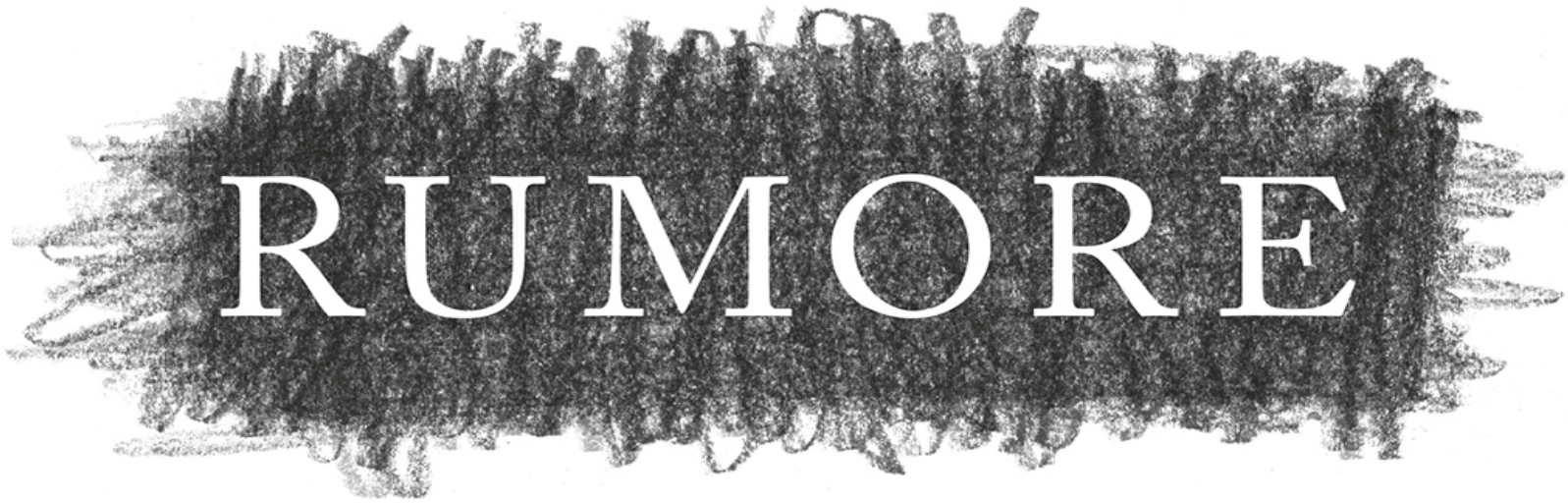
DANIEL
KAHNEMAN

PREMIO NOBEL PER L'ECONOMIA

OLIVIER SIBONY

CASS R. SUNSTEIN



RUMORE

Un difetto del ragionamento umano

UTET


UTET

Titolo originale: *Noise: A Flaw in Human Judgment*
Traduzione dall'inglese: Eleonora Gallitelli

Copyright © 2021 by Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein
All rights reserved

Per l'edizione italiana:
© 2021, DeA Planeta Libri S.r.L.
Redazione: Via Inverigo, 2 – 20151 Milano
www.deaplanetalibri.it/
ebook.deaplanetalibri.it

Prima edizione ebook: settembre 2021
ISBN: 978-88-511-9653-0

eBook realizzato da Punto Acuto
www.punto-acuto.it

Nessuna parte di questo volume può essere riprodotta, memorizzata o trasmessa in alcuna forma e con alcun mezzo, elettronico, meccanico o in fotocopia, in disco o in altro modo, compresi cinema, radio, televisione, senza autorizzazione scritta dall'Editore. Le fotocopie per uso personale del lettore possono essere effettuate nei limiti del 15% dietro pagamento alla SIAE del compenso previsto all'art. 68, commi 4 e 5, della legge 22 aprile 1941, n. 633.
Le riproduzioni per finalità di carattere professionale, economico o commerciale, o comunque per uso diverso da quello personale possono essere effettuate a seguito di specifica autorizzazione rilasciata da CLEARedi, corso di Porta Romana 108, 20122 Milano, e-mail info@clearedi.org e sito web www.clearedi.org

www.utetlibri.it

@utetlibri

@utetlibri

 @Utetlibri

Daniel Kahneman
Olivier Sibony
Cass R. Sunstein

RUMORE

Un difetto del ragionamento umano

Traduzione di Eleonora Gallitelli

UTET

Indice

Per Noga, Ori e Gili

D.K.

Per Fantin e Lélia

O.S.

Per Samantha

C.R.S.

Introduzione.

Due tipi di errore

Immaginate quattro squadre di amici che vanno a giocare al tiro a segno. A ogni squadra di cinque persone viene assegnato un fucile, e a ciascun partecipante spetta un tiro. La figura 1 mostra i risultati. L'ideale sarebbe che ogni tiro colpisse il centro del bersaglio.

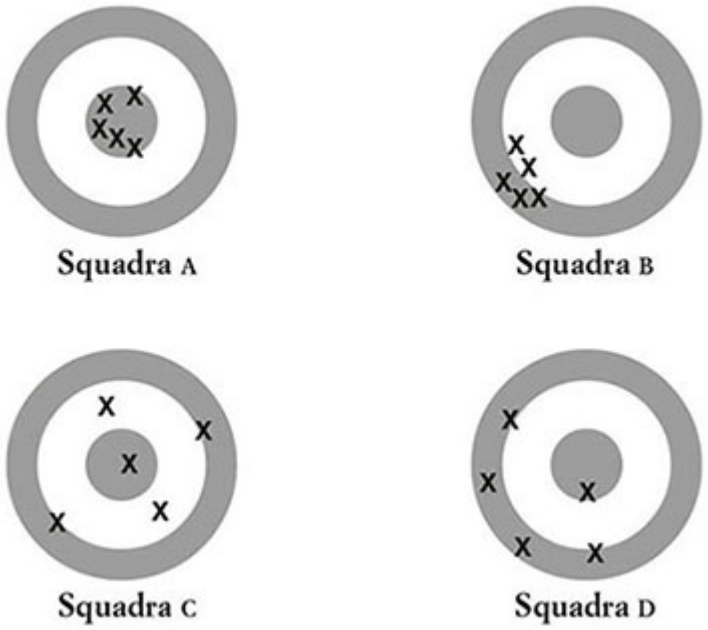


Figura 1. Quattro squadre

La squadra A ci va molto vicina. I tiri si concentrano nel tondino centrale, in una configurazione quasi perfetta.

Possiamo dire che la squadra B è affetta da *bias* – cioè da un errore sistematico, che ricorre in maniera prevedibile in particolari circostanze –,

perché i suoi tiri cadono sistematicamente fuori dal centro del bersaglio. Trattandosi di una deviazione costante, come illustra la figura, è possibile fare una previsione: se un membro di quella squadra dovesse fare un altro tiro, c'è da scommettere che cadrebbe nella stessa zona dei primi cinque. La costanza di questa deviazione ci induce poi a cercare una spiegazione causale: forse il mirino del fucile assegnato a questo gruppo era nella posizione sbagliata.

La squadra C, invece, potremmo definirla affetta da *rumore*, perché i suoi tiri sono sparpagliati qua e là. Non c'è un'ovvia deviazione, in quanto i colpi si concentrano, grosso modo, in un'area uniforme intorno al centro del bersaglio. Se un membro della squadra facesse un altro tiro, difficilmente sapremmo prevedere dove andrebbe a colpire. Inoltre non ci viene in mente nessuna ipotesi interessante per spiegare i risultati di questa squadra. Capiamo che i suoi membri sono scarsi come tiratori, ma non sapremmo spiegare il perché di tanto rumore.

La squadra D è affetta tanto da bias quanto da rumore. Come per la B, i suoi tiri mancano sistematicamente il centro, e come per la C, sono sparpagliati qua e là.

Ma questo non è un libro sul tiro a segno: a noi interessa l'errore umano. Bias e rumore – deviazione sistematica e dispersione casuale – sono due diverse componenti dell'errore. I bersagli ne illustrano la differenza.[1]

Il tiro a segno funge da metafora per gli sbagli che è possibile commettere nel giudicare, specialmente nelle decisioni di vario tipo che occorre prendere per conto di un'organizzazione. In tali situazioni, troveremo i due tipi di errore illustrati nella figura 1. Certi giudizi sono affetti da un bias: mancano sistematicamente il bersaglio, mentre altri sono intaccati dal rumore, come quando individui che dovrebbero convenire su

un punto colpiscono zone diversissime del bersaglio. Molte organizzazioni, purtroppo, sono funestate tanto dal bias quanto dal rumore.

La figura 2 illustra una differenza importante tra bias e rumore: indica cosa vedreste al tiro a segno se vi mostrassero solo il retro dei bersagli delle quattro squadre, senza alcuna indicazione del punto a cui stavano mirando.

Osservando soltanto il retro è impossibile dire se si sia avvicinata di più al centro la squadra A o la squadra B, ma si capisce subito che nelle squadre C e D c'è un certo rumore, mentre in A e B no. In effetti, riguardo alla dispersione la figura 2 ci dà le stesse informazioni della figura 1, né più né meno. Una delle proprietà generali del rumore è che è possibile riconoscerlo e misurarlo senza sapere nulla del bersaglio o del bias.

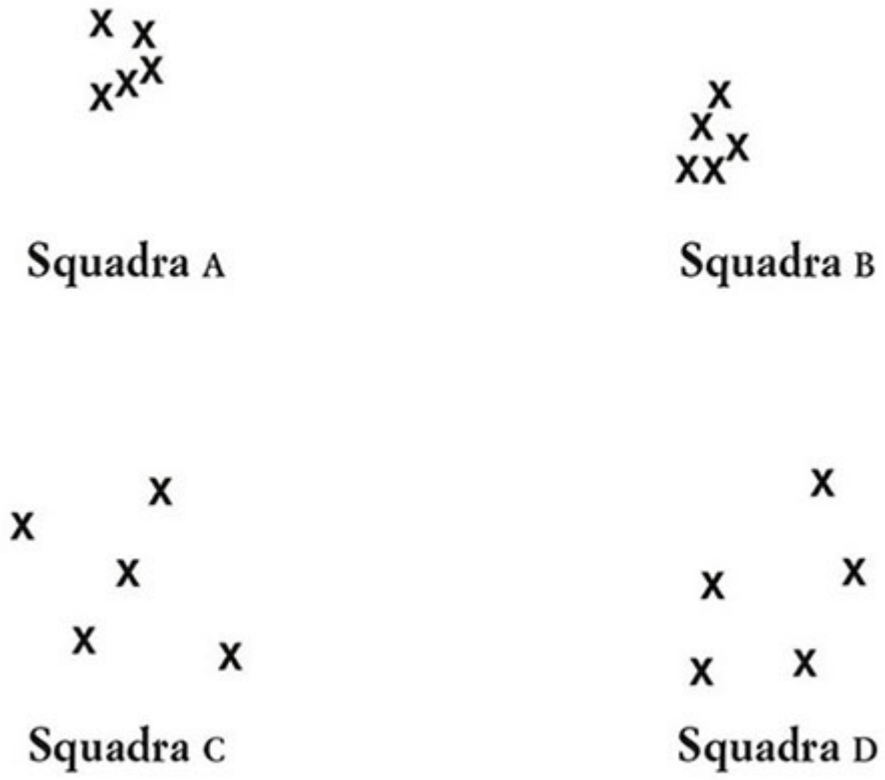


Figura 2. Il retro dei quattro bersagli

La suddetta proprietà è cruciale per i fini di questo libro, perché molte delle conclusioni che trarremo provengono da giudizi di cui non si conosce, né forse si può conoscere, la veridicità. Quando più medici formulano diagnosi diverse per lo stesso caso, è possibile analizzare questa difformità di giudizio senza sapere di cosa soffre il paziente. Quando dei produttori cinematografici ipotizzano il mercato potenziale di un film, possiamo studiare la variabilità delle loro proposte senza sapere se poi il film abbia avuto successo o sia mai stato prodotto. Non serve sapere chi ha ragione per valutare quanto varino i giudizi su uno stesso caso: per misurare il rumore non dobbiamo fare altro che guardare dietro il bersaglio.

Per comprendere un errore di giudizio occorre capirne sia il bias sia il rumore. Qualche volta, come vedremo, il problema principale è il rumore, ma nei discorsi sull'errore umano e nelle organizzazioni di tutto il mondo è raro che il rumore venga riconosciuto: è sempre il bias a occupare il centro della scena. Il rumore fa solo da comparsa o, spesso, resta addirittura fuori dalla scena. Il tema del bias è stato affrontato in migliaia di articoli scientifici e decine di testi divulgativi, ma sono in pochi a fare cenno al problema del rumore. In questo libro ci proponiamo di ristabilire un equilibrio.

Spesso, nelle decisioni che prendiamo giorno per giorno, il tasso di rumore è scandalosamente alto. Ecco alcuni esempi dell'enorme peso che ha il rumore in situazioni che richiederebbero grande accuratezza di giudizio:

- *La medicina è affetta da rumore.* Di fronte allo stesso paziente, medici diversi esprimono giudizi diversi sull'eventualità che il paziente abbia un tumore alla pelle o al seno, soffra di cardiopatia, tubercolosi, polmonite, depressione e una pletora di altre patologie. Il rumore è particolarmente elevato in ambito psichiatrico, dove, per forza di cose, il giudizio soggettivo è importante. Tuttavia, si ritrova un alto tasso di rumore anche in aree del tutto inaspettate, per esempio nella lettura delle radiografie.

- *Le decisioni sull'affidamento dei minori sono affette da rumore.*[2] Gli assistenti sociali devono valutare se i minori sono a rischio di maltrattamenti e, in tal caso, se darli in affido. L'elevata quantità di rumore all'interno del sistema si traduce in una maggiore propensione da parte di alcuni assistenti sociali rispetto ad altri a ricorrere alla procedura di affidamento. A distanza di anni, molti degli sfortunati che sono passati per le mani di questi coordinatori inclementi ne pagano le tristi conseguenze: tassi più alti di delinquenza e gravidanze precoci, e redditi più bassi.
- *Le previsioni sono affette da rumore.* Chi si occupa per lavoro di previsioni offre pronostici altamente variabili sulle probabilità di vendita di un nuovo prodotto, sulle probabilità di aumento del tasso di disoccupazione, sulle probabilità di fallimento delle aziende in difficoltà; insomma, praticamente su tutto. Sono in disaccordo non solo l'uno con l'altro, ma anche con se stessi. Per esempio, quando, in due giorni diversi, agli stessi sviluppatori di software[3] fu chiesta una stima del tempo di completamento di un certo compito, le loro previsioni in termini di ore di lavoro differirono in media del 71%.
- *Le decisioni sulle richieste di asilo sono affette da rumore.*[4] Per un richiedente asilo l'ammissione negli Stati Uniti è una specie di lotteria. Da uno studio di casi assegnati a giudici diversi in maniera casuale è emerso che un giudice aveva ammesso il 5% dei richiedenti, un altro l'88%. Il titolo dello studio è di per sé molto eloquente: *La roulette dei rifugiati.* (E qui di roulette ne vedremo tante.)
- *Le decisioni dell'ufficio del personale di un'azienda sono affette da rumore.* Gli addetti al reclutamento effettuano valutazioni diversissime dei medesimi candidati, e anche gli indici di performance degli stessi impiegati sono altamente variabili e dipendono più da chi effettua la valutazione che dalla performance stessa.
- *Le decisioni sulla libertà provvisoria sono affette da rumore.* Che a un accusato venga concessa la libertà provvisoria o che sia invece mandato in carcere in attesa del processo dipende in larga parte dall'identità del giudice. Alcuni sono più indulgenti di altri, anche quando si tratta di valutare quali imputati presentino il più elevato rischio di fuga o di recidiva.
- *La scienza forense è affetta da rumore.* Tendiamo a pensare che l'identificazione tramite impronte digitali sia infallibile, ma, a volte, tra gli esperti c'è chi individua una corrispondenza tra un'impronta trovata sulla scena del delitto e quella di un sospettato, e chi no. Non solo non vi è accordo tra gli esperti, ma capita che, analizzando la stessa impronta in circostanze diverse, lo stesso esperto formuli giudizi incoerenti. Una simile variabilità è documentata in diversi ambiti della scienza forense e perfino nelle analisi del DNA.
- *Le decisioni sull'assegnazione dei brevetti sono affette da rumore.*[5] Gli autori di un importante studio sulle domande di brevetti fanno notare quanto rumore si riscontri anche in quel campo: «La concessione di un brevetto dipende in larga misura dall'esaminatore che per puro caso viene assegnato alla domanda». Tale variabilità, naturalmente, costituisce un problema in termini di equità.

Tutte queste situazioni ad alto tasso di rumore sono solo la punta di un grosso iceberg. In qualsiasi tipo di giudizio umano ci sarà con ogni probabilità un certo grado rumore; cercare di debellarlo, così come eliminare il bias, è l'unico modo che abbiamo per migliorare la qualità dei nostri giudizi.

Questo libro consta di sei parti. Nella prima analizzeremo la differenza tra rumore e bias, e mostreremo come le organizzazioni, sia pubbliche sia private, possano essere "rumorose", a volte in una misura sconvolgente. Per comprendere il problema, partiremo dai giudizi formulati in due ambiti: quello delle sentenze penali (settore pubblico) e quello delle assicurazioni (settore privato). A prima vista queste due aree non potrebbero essere più diverse, ma, nella prospettiva del rumore, hanno molto in comune. Per dimostrarlo, introdurremo il concetto di "controllo del rumore", volto a misurare il livello di disaccordo tra professionisti che si occupano degli stessi casi all'interno di un'organizzazione.

Nella seconda parte indagheremo sulla natura del giudizio umano e cercheremo di capire come misurare l'accuratezza e l'errore. Come detto, i giudizi sono affetti da bias e rumore, e proveremo a descrivere come, sorprendentemente, questi due tipi di errore abbiano un ruolo analogo. Analizzeremo il rumore occasionale, che consiste nella variabilità dei giudizi formulati sullo stesso caso da parte della stessa persona o dello stesso gruppo in occasioni diverse, e vedremo come questo influisca nelle discussioni di gruppo attraverso fattori apparentemente irrilevanti, come chi prende la parola per primo.

La terza parte si concentrerà su un tipo di giudizio già ampiamente studiato: il giudizio predittivo. Ci soffermeremo sui principali vantaggi dell'affidarsi a regole, formule e algoritmi piuttosto che agli esseri umani quando si tratta di fare previsioni: contrariamente a quanto spesso si

crede, il motivo non è tanto la superiorità delle regole, quanto la loro assenza di rumore. Discuteremo del limite qualitativo fondamentale del giudizio predittivo, ovvero l'ignoranza oggettiva del futuro, che, insieme al rumore, contribuisce a limitare la qualità della predizione. Infine affronteremo una domanda che a questo punto quasi certamente vi sarete posti: se il rumore è davvero così onnipresente, come mai ve ne siete accorti soltanto adesso?

Nella quarta parte ci sposteremo nell'ambito della psicologia, analizzando le cause principali del rumore. Tra queste vi sono le differenze interpersonali dovute a vari fattori, come la personalità e lo stile cognitivo; le variazioni idiosincratiche del peso attribuito a diversi elementi; gli usi differenti che le persone fanno delle stesse scale di valutazione. Cercheremo poi di capire perché la gente tende a ignorare il rumore e spesso non si sorprende davanti a eventi e giudizi che non avrebbe mai potuto prevedere.

La quinta parte affronterà il problema pratico di come migliorare i propri giudizi ed evitare di sbagliare. (I lettori interessati soprattutto alle applicazioni pratiche della riduzione del rumore possono saltare la trattazione delle sfide della previsione e della psicologia del giudizio affrontata nella terza e quarta parte, e passare direttamente a questa sezione.) Considereremo i tentativi per contrastare il rumore condotti in ambito medico, aziendale, formativo, governativo e non solo. Presenteremo alcune tecniche di riduzione del rumore, che riuniremo sotto l'etichetta di "igiene decisionale", illustreremo cinque casi studio in ambiti noti per l'elevata presenza di rumore e in cui si è lavorato molto per ridurlo, provando a trarre conclusioni istruttive dai diversi gradi di successo ottenuti. I casi di studio riguardano diagnosi mediche inaffidabili, indicatori di performance, scienze forensi, decisioni sulle assunzioni e

previsioni in senso generale. Concluderemo presentando un sistema che abbiamo chiamato "protocollo di valutazione mediata": un approccio per la valutazione delle opzioni disponibili valido in tutti i campi, che comprende varie importanti prassi di igiene decisionale ed è teso a ridurre il rumore e pervenire a giudizi più affidabili.

Qual è il giusto livello di rumore? Questa domanda verrà affrontata nella sesta parte. Contro ogni aspettativa, il giusto livello non è zero. In certe aree è impossibile eliminare il rumore. In altre, è troppo costoso. In altre ancora, gli sforzi per ridurre il rumore andrebbero a discapito di importanti valori in contrasto tra loro: per esempio, potrebbero abbattere il morale e dare alle persone coinvolte l'impressione di essere trattate come rotelle di un ingranaggio. Quando si cerca la risposta negli algoritmi, sorgono le obiezioni più disparate, e qui ne affronteremo alcune, ma ciò non toglie che l'attuale livello di rumore sia inaccettabile. Esortiamo quindi le organizzazioni pubbliche e private a effettuare dei controlli del rumore e a sforzarsi di ridurlo con la massima energia e serietà. Così facendo, potrebbero abbattere le disparità più diffuse, nonché i costi, in molte aree.

Tenendo in mente questo obiettivo, concluderemo ogni capitolo con qualche breve citazione. Potrete prenderle alla lettera oppure adattarle alle questioni che più vi premono, che si tratti di salute, sicurezza, istruzione, denaro, lavoro, tempo libero o altro. Comprendere il problema del rumore, e cercare di risolverlo, è un processo in via di definizione che richiede uno sforzo collettivo, a cui tutti noi abbiamo l'opportunità di contribuire. Questo libro è stato scritto nella speranza di poter cogliere questa opportunità.

[1] Impiegando archi e frecce al posto dei fucili, il matematico svizzero Daniel Bernoulli propose la stessa analogia nel 1778 in un saggio sui problemi di valutazione. D. Bernoulli, *The Most Probable Choice Between Several Discrepant Observations and the Formation Therefrom of the Most Likely Induction*, in "Biometrika", 48(1961), n. 1-2, pp. 3-18, [doi.org/10.1093/biomet/48.1-2.3].

[2] J.J. Doyle Jr., *Child Protection and Child Outcomes: Measuring the Effects of Foster Care*, in "American Economic Review", 95(2007), n. 5, pp. 1583-1610.

[3] S. Grimstad, M. Jørgensen, *Inconsistency of Expert Judgment-Based Estimates of Software Development Effort*, in "Journal of Systems and Software", 80(2007), n. 11, pp. 1770-1777.

[4] A.I. Schoenholtz, J. Ramji-Nogales, P.G. Schrag, *Refugee Roulette: Disparities in Asylum Adjudication*, in "Stanford Law Review", 60(2007), n. 2.

[5] M.A. Lemley, B. Sampat, *Examiner Characteristics and Patent Office Outcomes*, in "Review of Economics and Statistics", 94(2012), n. 3, pp. 817-827. Vedi anche I. Cockburn, S. Kortum, S. Stern, *Are All Patent Examiners Equal? The Impact of Examiner Characteristics*, working paper 8980(2002), [www.nber.org/papers/w8980]; e M.D. Frakes, M.F. Wasserman, *Is the Time Allocated to Review Patent Applications Inducing Examiners to Grant Invalid Patents? Evidence from Microlevel Application Data*, in "Review of Economics and Statistics", 99(2017), n. 3, pp. 550-563.

PRIMA PARTE

Trovare il rumore

È inaccettabile che persone condannate per lo stesso reato ricevano sentenze drasticamente diverse, come per esempio cinque anni di reclusione per una e libertà vigilata per l'altra; eppure, in molte situazioni, è proprio quello che accade. Certo, la giustizia penale è affetta anche da bias, ma in questo primo capitolo ci concentreremo sul rumore, e in particolare su quel che accadde quando un famoso giudice richiamò l'attenzione su questo problema e, trovandolo vergognoso, lanciò una crociata che in un certo senso cambiò il mondo (anche se non abbastanza). Questa storia riguarda gli Stati Uniti, ma siamo certi che si possano ritrovare (e si ritroveranno) vicende simili in molti altri paesi, anzi, è probabile che in altre nazioni il problema del rumore sia ancora più grave. Attraverso questo esempio vogliamo dimostrare come il rumore possa portare a grandi ingiustizie.

Le condanne penali generano risultati particolarmente drammatici, ma ci occuperemo anche del settore privato, un altro campo in cui la posta in gioco può essere decisamente alta. Per argomentare questa affermazione, nel prossimo capitolo parleremo di una grande compagnia assicurativa, in cui i sottoscrittori hanno il compito di fissare i premi per i potenziali clienti, mentre i periti liquidatori devono giudicare l'entità delle richieste di risarcimento. Si tende a immaginare che si tratti di compiti semplici e meccanici, e che professionisti diversi perverranno, grosso modo, alle stesse cifre; tuttavia, i risultati di un esperimento molto accurato di controllo del rumore hanno sorpreso noi e sconcertato i dirigenti della

compagnia. L'alto tasso di rumore, come abbiamo avuto modo di appurare, è costato molto caro alla società: questa è una dimostrazione di come il rumore possa causare anche grosse perdite economiche.

Entrambi gli esempi si basano su studi condotti su un gran numero di persone, a cui viene chiesto di esprimere un gran numero di giudizi. Ma molti giudizi importanti sono *singoli*, non ripetuti: come gestire un'opportunità d'affari unica, come scegliere se lanciare o no un nuovo prodotto, come affrontare una pandemia, come decidere se assumere o no qualcuno che non risponde al profilo standard. È possibile trovare traccia di rumore anche in decisioni da prendere in situazioni eccezionali come queste? Tenderemmo a rispondere negativamente: dopotutto, il rumore è una variabilità indesiderata, e come si può avere variabilità in decisioni singole? Nel terzo capitolo cercheremo di rispondere a questa domanda, tenendo presente che il giudizio a cui si arriva, perfino in una situazione in apparenza straordinaria, è solo uno in un orizzonte di possibilità. Troveremo molto rumore anche lì.

Il tema che emerge da questi tre capitoli si può sintetizzare in una frase, che costituisce uno dei pilastri di questo libro: *dove c'è giudizio, c'è rumore, e più di quanto non si pensi.* Cerchiamo di comprenderne l'entità.

1

Delitto e rumoroso castigo

Poniamo che qualcuno venga condannato per aver commesso un reato – taccheggio, possesso di eroina, violenza privata o rapina a mano armata. Quale pena dovrà scontare?

La risposta a questa domanda non dovrebbe dipendere dal giudice a cui viene assegnato il caso, dalle condizioni meteorologiche di quel giorno o dal risultato di una partita di football giocata il giorno prima. Sarebbe vergognoso se tre persone condannate per lo stesso reato ricevessero una pena radicalmente diversa: libertà vigilata per il primo, due anni di reclusione per il secondo, dieci anni per il terzo. Eppure questo avviene in molte nazioni, e non solo nel passato, ma ancora oggi.

Per molto tempo, in ogni parte del mondo, i giudici hanno goduto di un ampio margine discrezionale quando si è trattato di decidere quale fosse la sentenza appropriata. In molti paesi, diversi esperti hanno tessuto le lodi di questa discrezionalità, considerandola tanto giusta quanto umana: le sentenze penali, secondo loro, dovevano tenere conto di una molteplicità di fattori legati non solo al reato, ma anche all'indole dell'imputato e alle circostanze. Insomma, la personalizzazione della sentenza era all'ordine del giorno. Se i giudizi fossero stati vincolati da determinate regole, i criminali avrebbero subìto un trattamento disumano, perché non sarebbero stati visti come individui unici con il diritto a portare l'attenzione sulla propria situazione particolare. L'idea stessa di processo

giusto sembrava, per molti, richiedere una discrezionalità giudiziale illimitata.

Negli anni settanta questo entusiasmo generale per la discrezionalità giudiziale iniziò a scemare per un semplice motivo: l'allarmante evidenza del rumore. Nel 1973 il celebre giudice Marvin Frankel portò il problema nel dibattito pubblico. Prima della sua nomina a giudice, Frankel, paladino della libertà di parola e appassionato sostenitore dei diritti umani, contribuì alla fondazione della Lawyers Committee for Human Rights, un'organizzazione oggi nota come Human Rights First.

Frankel, che sapeva essere molto severo, era indignato per il livello di rumore nel sistema penale, e motivò così il suo impegno:

> Un imputato condannato per una rapina a una banca federale poteva ricevere fino a un massimo di venticinque anni, vale a dire che la pena poteva variare da zero a venticinque anni. Capii subito che quel numero non dipendeva tanto dal caso o dal singolo imputato, quanto dal singolo giudice, ovvero dalle sue opinioni, preferenze e bias. Pertanto, lo stesso imputato poteva ottenere una sentenza molto diversa nello stesso processo a seconda del giudice a cui veniva affidato il suo caso.[1]

Frankel non forniva alcun tipo di analisi statistica a sostegno della sua argomentazione, ma snocciolava una serie di aneddoti molto incisivi su disparità ingiustificate nel trattamento di persone simili. Due uomini, entrambi incensurati, avevano subìto una condanna per avere incassato assegni falsi rispettivamente dell'ammontare di 58,40 e 35,20 dollari. Il primo fu condannato a quindici *anni* di reclusione, il secondo a trenta *giorni*. Per atti di appropriazione indebita analoghi tra loro, un uomo era stato condannato a centodiciassette *giorni* di carcere, un altro a venti *anni*. Citando numerosi casi di questo tipo, Frankel deplorò quelli che definì «i poteri pressoché incontrollati e indiscriminati»[2] dei giudici federali, che sfociavano in «crudeltà arbitrarie perpetrate quotidianamente»,[3] a suo parere inaccettabili in un «governo delle leggi, non degli uomini».[4]

Fece appello al Congresso per porre fine a quella «discriminazione», come definì quelle crudeltà arbitrarie. Con quel termine intendeva riferirsi soprattutto al rumore, presente sotto forma di inspiegabili variazioni nelle condanne, ma a preoccuparlo era anche il bias, che portava a disparità di giudizio su basi razziali e socioeconomiche. Per contrastare rumore e bias, insistette affinché non fossero permesse differenze di trattamento nei procedimenti penali, se non «giustificate da relativi test in grado di emettere formulazioni e istanze con un livello di obiettività sufficiente a garantire risultati migliori degli *ukase* idiosincratici di particolari funzionari, giudici o altri soggetti».[5] (Qui "*ukase* idiosincratici" è un'espressione un po' esoterica per riferirsi agli editti personali dei giudici.) Ma, soprattutto, Frankel insistette per ridurre il rumore attraverso «un profilo dettagliato o un elenco di fattori che includano, ove possibile, una forma di valutazione numerica o comunque oggettiva».[6]

Frankel espose le sue idee nei primi anni settanta, quindi non arrivò al punto da difendere quella che definì la «sostituzione delle persone con le macchine», ma ci andò sorprendentemente vicino. A suo dire, «il principio di legalità richiede un corpus di regole impersonali, applicabili in maniera generalizzata, vincolante per i giudici come per chiunque altro». Sostenne senza mezzi termini la necessità di avvalersi dei «computer come ausilio a un pensiero ordinato nella formulazione delle sentenze».[7] Raccomandò inoltre l'istituzione di una commissione che analizzasse le sentenze stesse.[8]

Il libro di Frankel divenne uno dei testi più influenti nella storia del diritto penale, negli Stati Uniti come nel resto del mondo. Il suo lavoro tuttavia risentiva di una certa informalità: era sconvolgente ma impressionistico e, per verificare la sua tesi, diversi studiosi avviarono indagini sul livello di rumore nelle condanne penali.

Un primo studio su vasta scala di questo tipo fu diretto dallo stesso giudice Frankel nel 1974. A cinquanta giudici di vari distretti venne chiesto di stabilire una condanna per casi ipotetici riassunti in identici resoconti presentenza. Il principale risultato fu che «la norma era l'assenza di uniformità di giudizio»[9] e che la variazione delle pene era «sconcertante».[10] Uno spacciatore di eroina poteva restare in carcere da uno a dieci anni a seconda del giudice.[11] La pena per una rapina in banca variava da cinque a diciotto anni di carcere.[12] Dallo studio emerse che in un caso di estorsione la condanna variava dalla pena stratosferica di vent'anni di reclusione e sessantacinquemila dollari di multa ad appena tre anni di carcere senza alcuna multa.[13] Ma l'aspetto più sorprendente era che in sedici casi su venti non vi era unanimità nemmeno sulla necessità stessa di una pena detentiva.

Questo studio fu seguito da una serie di altri, da cui emersero livelli di rumore altrettanto sbalorditivi. Nel 1977, per esempio, William Austin e Thomas Williams condussero un sondaggio che coinvolse quarantasette giudici,[14] a cui fu chiesto di pronunciarsi sugli stessi cinque casi di reati minori. Le descrizioni dei casi consistevano in una sintesi delle informazioni di cui si servono i giudici per emettere una sentenza, come l'accusa, la deposizione, i precedenti penali (se presenti), la condizione sociale e le evidenze relative all'indole dell'imputato. Lo studio rilevò «una discrepanza sostanziale»: in un caso di furto con scasso, per esempio, le condanne proposte andavano da cinque anni di carcere a soli trenta giorni (più una multa di cento dollari), mentre in un caso di possesso di marijuana certi giudici proposero la detenzione, altri la libertà vigilata.

Uno studio molto più ampio condotto nel 1981 coinvolse duecentootto giudici federali, ai quali vennero presentati gli stessi sedici casi ipotetici. I risultati furono scioccanti:

> Solo in tre casi su sedici vi era un accordo unanime sulla detenzione. Inoltre, anche quando i giudici concordavano sull'appropriatezza della carcerazione, vi era una variazione sostanziale nella lunghezza del periodo di detenzione raccomandato. In un caso di frode in cui il periodo di carcerazione medio era di 8,5 anni, il termine proposto più lungo fu l'ergastolo. In un altro, il periodo medio era di 1,1 anni, eppure il periodo più lungo fu di quindici anni.[15]

Per quanto rivelatori, è molto probabile che questi studi, basati su esperimenti strettamente controllati, sottovalutino la portata del rumore nella realtà della giustizia penale. Nella vita reale i giudici sono esposti a una mole di informazioni di gran lunga superiore rispetto ai dettagli inseriti ad arte nelle storielle presentate a chi si sottoponeva a questi esperimenti. Alcune di queste informazioni aggiuntive naturalmente sono rilevanti, ma vi è ampia evidenza che quelle irrilevanti, ovvero i piccoli fattori apparentemente casuali, possono produrre grandi differenze negli esiti. Per esempio, si è scoperto che i giudici sono più propensi a concedere la libertà condizionale all'inizio della giornata o dopo la pausa pranzo che non immediatamente prima di una pausa. Un giudice affamato è più severo.[16]

Uno studio condotto su migliaia di sentenze di tribunali minorili[17] ha riscontrato che quando la squadra di football locale perde una partita nel fine settimana, i giudici prendono decisioni più dure il lunedì (e, in misura minore, per il resto della settimana). Gli imputati neri sono decisamente i più penalizzati da questo incremento di severità. Un altro studio ha analizzato un milione e mezzo di decisioni giudiziarie emesse nell'arco di tre decenni, arrivando alla conclusione analoga che i giudici sono più severi nei giorni successivi a una sconfitta della squadra cittadina che nei giorni successivi a una vittoria.[18]

L'analisi di sei milioni di sentenze formulate da giudici francesi nell'arco di dodici anni ha dimostrato che questi sono più clementi con gli imputati

nel giorno del loro compleanno.[19] (Cioè del compleanno degli imputati; sospettiamo che possano essere più clementi anche nel giorno del proprio compleanno, ma, a quanto ci risulta, questa ipotesi non è stata verificata.) I giudici possono essere influenzati perfino da un fattore irrilevante come la temperatura esterna.[20] Da una rassegna di 207 000 decisioni prese dal tribunale dell'immigrazione in quattro anni si evince che le variazioni giornaliere della temperatura possono avere un effetto significativo: con il caldo, è meno probabile che venga concesso l'asilo. Insomma, se sei vittima di una persecuzione politica nel tuo paese e chiedi asilo all'estero, devi sperare e forse anche pregare che la tua udienza cada in una giornata fresca.

Ridurre il rumore nelle sentenze

Negli anni settanta le argomentazioni di Frankel e i dati empirici che le confortavano giunsero all'attenzione di Edward M. Kennedy, fratello del presidente assassinato John F. Kennedy, nonché uno dei membri più autorevoli del senato americano, che ne rimase sconcertato e inorridito, e già nel 1975 presentò un progetto di riforma del meccanismo sanzionatorio. Il tentativo non ebbe successo, ma Kennedy era implacabile. Dati alla mano, per anni continuò a fare pressione per arrivare all'approvazione di quella legge, finché nel 1984, di fronte all'evidenza di quella variabilità ingiustificata, il Congresso approvò il Sentencing Reform Act.

La nuova legge mirava a ridurre il rumore del sistema limitando «la discrezionalità incondizionata conferita ai giudici e alle autorità del *parole*, responsabili dell'imposizione e dell'adempimento delle sentenze». In particolare, i membri del Congresso facevano riferimento a una disparità di

trattamento «ingiustificatamente ampia»,[21] citando dati specifici relativi all'area di New York, dove le pene per casi identici potevano variare da tre a vent'anni di prigione. Come auspicato dal giudice Frankel, questa legge portò all'istituzione della US Sentencing Commission, con il chiaro intento di promulgare linee guida che vincolassero i giudici nel procedimento di determinazione della pena e che stabilissero dei limiti temporali ben definiti per le condanne penali.

L'anno successivo la commissione stabilì le linee guida, generalmente basate sulla media delle condanne per crimini simili su diecimila casi reali analizzati. Il giudice della Corte suprema Stephen Breyer, che ebbe un ruolo importante nel processo di riforma, difese questo ricorso a sentenze passate, notando un insanabile disaccordo all'interno della commissione. «Perché la commissione non si è messa a un tavolo e non ha cercato di ragionare sul problema senza basarsi sui precedenti storici? In sintesi: non ci siamo riusciti. Non ci siamo riusciti perché da ogni parte vengono avanzati ottimi argomenti che portano in direzioni opposte. [...] Provate a stilare un elenco di tutti i reati esistenti per grado di punibilità, [...] poi confrontate il vostro elenco con quello dei vostri conoscenti e vedete se corrispondono. Ebbene, vi anticipo che la risposta è no.»[22]

In ottemperanza alle linee guida, i giudici devono considerare due fattori per stabilire una pena: il crimine e i precedenti penali dell'imputato. A ogni crimine è associato uno dei quarantatré «livelli di reato» in base alla gravità, mentre i precedenti penali corrispondono principalmente al numero e alla gravità delle passate condanne dell'imputato. Una volta considerati il crimine e i precedenti penali, le linee guida presentano una gamma di pene relativamente limitata, autorizzando una differenza tra la pena più dura e quella più lieve pari a un massimo di sei mesi o al 25%. Ai giudici è concesso di discostarsi dalle linee guida qualora riconoscano una

circostanza aggravante o attenuante, ma tali scostamenti devono essere giustificati davanti a una corte d'appello.[23]

Per quanto vincolanti, queste linee guida sono molto meno rigide di quanto avesse desiderato il giudice Frankel, e concedono ai giudici un notevole spazio di manovra. Detto ciò, diversi studi condotti con varie metodologie su periodi storici differenti sono giunti alla medesima conclusione: le linee guida riducono il rumore. In linguaggio tecnico, «riducono la variazione netta della pena attribuibile all'identità casuale dell'autorità giudicante».[24]

Lo studio più elaborato, realizzato dalla commissione stessa,[25] confrontava le sentenze di processi per rapina in banca, spaccio di cocaina, spaccio di eroina e peculato emesse nel 1985 (prima che entrassero in vigore le linee guida) con sentenze pronunciate tra il 19 gennaio 1989 e il 30 settembre 1990. I trasgressori venivano associati in relazione ai fattori che in base alle linee guida erano rilevanti ai fini della determinazione della pena. Per ciascun tipo di reato le differenze di valutazione tra i giudici erano molto inferiori nel secondo periodo, dopo l'approvazione del Sentencing Reform Act.

Secondo un altro studio, la differenza attesa nella lunghezza delle pene comminate dai vari giudici tra il 1986 e il 1987 era del 17% o di 4,9 mesi, cifra che si riduceva all'11% o a 3,9 mesi tra il 1988 e il 1993.[26] Uno studio indipendente che considerava periodi diversi riscontrò un analogo successo nella riduzione delle disparità tra i giudici, definite come le differenze nella lunghezza media della pena comminata tra giudici con un numero simile di casi.[27]

A dispetto dei risultati, le linee guida subirono una valanga di critiche. Alcuni, compresi molti giudici, ritenevano che alcune pene fossero troppo severe – quindi un problema di bias, non di rumore. Un'obiezione per noi

molto più interessante mossa da numerosi giudici era che le linee guida fossero profondamente ingiuste, perché impedivano di prendere in adeguata considerazione i particolari di ciascun caso. Quella riduzione del rumore comportava un'inaccettabile meccanizzazione del processo decisionale. La docente di diritto di Yale Kate Stith e il giudice federale José Cabranes scrissero che «non occorre cecità, occorrono equità e discernimento», i quali «possono manifestarsi esclusivamente in un giudizio che tenga conto delle complessità del caso singolo».[28]

Questa obiezione condusse a vigorose contestazioni delle linee guida, alcune per ragioni legali, altre per motivazioni politiche. Contestazioni che non portarono a nulla, finché, per motivi tecnici che non avevano niente a che fare con il dibattito qui riportato, la Corte suprema revocò le linee guida nel 2005.[29] Dopo la deliberazione della Corte, esse assunsero un valore meramente consultivo. Si noti che gran parte dei giudici federali fu molto soddisfatta di quella decisione: il 75% preferiva il nuovo regime, mentre solo il 3% riteneva che fosse preferibile l'obbligatorietà.[30]

Quale effetto ebbe questo passaggio da indicazioni vincolanti a consultive? La docente di diritto di Harvard Crystal Yang indagò sulla questione, non mediante un esperimento o un sondaggio, ma con una massiccia raccolta di dati relativi a sentenze penali reali, che coinvolgevano circa quattrocentomila imputati. Il principale risultato della ricerca fu che, considerando molti indicatori, le disparità tra i giudici erano sensibilmente aumentate dopo il 2005. Quando le linee guida erano vincolanti, gli imputati che erano stati condannati da un giudice relativamente severo avevano ricevuto una pena più lunga di 2,8 mesi rispetto a quella che avrebbero ottenuto da un ipotetico giudice medio. Quando le indicazioni divennero puramente consultive, la disparità raddoppiò. Con parole che ricordano quelle impiegate dal giudice Frankel quarant'anni prima, Yang scrisse che i

suoi «risultati sollevano grandi problemi di equità, poiché l'identità del giudice assegnato al caso incide in maniera significativa sul trattamento discriminatorio di imputati simili condannati per reati simili».[31]

Dopo che le linee guida divennero consultive, i giudici furono più inclini a basare le sentenze sui propri valori personali. Le indicazioni vincolanti, al contrario, avevano ridotto tanto il bias quanto il rumore. Dopo la decisione della Corte suprema, vi fu un aumento significativo nella disparità tra le condanne degli imputati afroamericani e quelle dei bianchi colpevoli degli stessi reati. Al contempo i giudici donne divennero più inclini dei loro colleghi uomini a esercitare quella maggiore discrezionalità per formulare sentenze più indulgenti, e lo stesso accadde per i giudici nominati da presidenti democratici.

Tre anni dopo la morte di Frankel, avvenuta nel 2002, la revoca delle linee guida obbligatorie riportò in vita quello che era stato il suo incubo: una legge senza ordine.

La storia della lotta del giudice Frankel a favore di linee guida per la commisurazione della pena mette in luce alcuni dei punti chiave che tratteremo in questo libro. Primo, giudicare è difficile perché il mondo è pieno di complicazioni e di incertezze; tale complessità, evidente nel sistema giudiziario, si riscontra in quasi tutte le situazioni che richiedono un giudizio professionale: in senso lato, vi rientrano i giudizi di medici, infermieri, avvocati, ingegneri, insegnanti, architetti, produttori cinematografici, addetti alla selezione del personale, editori, dirigenti aziendali di qualsiasi tipo e direttori sportivi. Ogni volta che c'è di mezzo un giudizio, è inevitabile che vi sia disaccordo.

Secondo, il livello di tale disaccordo è di gran lunga superiore al previsto. Se sono in pochi a obiettare al principio della discrezionalità giudiziale, quasi tutti disapprovano le enormi disparità che ne derivano. Il *rumore*

sistemico, ovvero una variabilità indesiderata in giudizi che in teoria dovrebbero essere identici, può generare ingiustizie incontrollate, alti costi economici ed errori di vario tipo.

Terzo, ridurre il rumore è possibile: l'approccio auspicato da Frankel e attuato dalla Sentencing Commission, ovvero l'utilizzo di regole e linee guida, è uno dei tanti in grado di farlo. Diversi tipi di giudizi richiederanno diversi approcci. Inoltre, alcuni metodi impiegati per ridurre il rumore possono avere allo stesso tempo il medesimo effetto anche sul bias.

Quarto, spesso gli sforzi per ridurre il rumore sollevano obiezioni e incorrono in serie difficoltà. Anche di questo occorre tenere conto, o la lotta al rumore sarà destinata a fallire.

A proposito del rumore nelle sentenze

«Gli esperimenti mostrano grandi disparità tra i giudici nelle pene proposte per casi identici. Tale variabilità è evidentemente ingiusta. La pena di un imputato non dovrebbe dipendere dal giudice a cui, in maniera del tutto casuale, viene assegnato il suo caso.»

«Le condanne penali non dovrebbero dipendere dall'umore del giudice nel corso dell'udienza o dalle condizioni meteorologiche.»

«Le linee guida costituiscono uno dei modi possibili per affrontare il problema, ma molti le disapprovano perché limitano la discrezionalità dei giudici, che può essere necessaria per garantire imparzialità e accuratezza di giudizio. Dopotutto ogni caso è un caso a sé, o no?»

[1] M. Frankel, *Criminal Sentences: Law Without Order*, in "25 Inst. for Sci. Info. Current Contents / Soc. & Behavioral Scis.: This Week's Citation Classic", 14(23 giugno 1986), n. 2A-6, disponibile al link [www.garfield.library.upenn.edu/classics1986/A1986C697400001.pdf].

[2] M. Frankel, *Criminal Sentences: Law Without Order*, Hill and Wang, New York 1973, p. 5.

[3] Ivi, p. 103.

[4] Ivi, p. 5.

[5] Ivi, p. 11.

[6] Ivi, p. 114.

[7] Ivi, p. 115.

[8] Ivi, p. 119.

[9] A. Partridge, W.B. Eldridge, *The Second Circuit Sentence Study: A Report to the Judges of the Second Circuit August 1974*, Federal Judicial Center, Washington, DC 1974, p. 9.

[10] US Senate, *Comprehensive Crime Control Act of 1983: Report of the Committee on the Judiciary, United States Senate, on S. 1762, Together with Additional and Minority Views*, US Government Printing Office, Washington, DC 1983, report n. 98-225.

[11] A. Partridge, W.B. Eldridge, *Second Circuit Sentence Study*, cit., A-11.

[12] Ivi, A-9.

[13] Ivi, A-5, A-7.

[14] W. Austin, T.A. Williams III, *A Survey of Judges' Responses to Simulated Legal Cases: Research Note on Sentencing Disparity*, in "Journal of Criminal Law & Criminology", 68(1977), p. 306.

[15] J. Bartolomeo *et al.*, *Sentence Decisionmaking: The Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity*, in "Journal of Criminal Law and Criminology", 72(1981), n. 2. (Vedi capitolo 6 per un'ampia trattazione.) Vedi anche US Senate, *Comprehensive Crime Control Act of 1983*, cit., report n. 44.

[16] S. Danziger, J. Levav, L. Avnaim-Pesso, *Extraneous Factors in Judicial Decisions*, in "Proceedings of the National Academy of Sciences of the United States of America", 108(2011), n. 17, pp. 6889-6892.

[17] O. Eren, N. Mocan, *Emotional Judges and Unlucky Juveniles*, in "American Economic Journal: Applied Economics", 10(2018), n. 3, pp. 171-205.

[18] D.L. Chen, M. Loecher, *Mood and the Malleability of Moral Reasoning: The Impact of Irrelevant Factors on Judicial Decisions*, in "SSRN Electronic Journal", 21 settembre 2019, pp. 1-70, [users.nber.org/dlchen/papers/Mood_and_the_Malleability_of_Moral_Reasoning.pdf].

[19] D.L. Chen, A. Philippe, *Clash of Norms: Judicial Leniency on Defendant Birthdays*, 2020, disponibile al link [ssrn.com/abstract=3203624].

[20] A. Heyes, S. Saberian, *Temperature and Decisions: Evidence from 207 000 Court Cases*, in "American Economic Journal: Applied Economics", 11(2019), n. 2, pp. 238-265.

[21] US Senate, *Comprehensive Crime Control Act of 1983*, cit., report n. 38.

[22] Il giudice Breyer è citato in J. Rosen, *Breyer Restraint*, in "New Republic", 11 luglio 1994, pp. 19, 25.

[23] US Sentencing Commission, *Guidelines Manual (2018)*, www.ussc.gov/sites/default/files/pdf/guidelines-manual/2018/GLMFull.pdf.

[24] J.M. Anderson, J.R. Kling, K. Stith, *Measuring Interjudge Sentencing Disparity: Before and After the Federal Sentencing Guidelines*, in "Journal of Law and Economics", 42(1999), n. S1, pp. 271-308.

[25] US Sentencing Commission, *The Federal Sentencing Guidelines: A Report on the Operation of the Guidelines System and Short-Term Impacts on Disparity in Sentencing, Use of Incarceration, and Prosecutorial Discretion and Plea Bargaining*, voll. 1-2, US Sentencing Commission, Washington, DC 1991.

[26] J.M. Anderson, J.R. Kling, K. Stith, *Measuring Interjudge Sentencing Disparity*, cit.

[27] P.J. Hofer, K.R. Blackwell, R. Barry Ruback, *The Effect of the Federal Sentencing Guidelines on Inter-Judge Sentencing Disparity*, in "Journal of Criminal Law and Criminology", 90(1999), pp. 239, 241.

[28] K. Stith, J. Cabranes, *Fear of Judging: Sentencing Guidelines in the Federal Courts*, University of Chicago Press, Chicago 1998, p. 79.

[29] 543 U.S. 220(2005).

[30] US Sentencing Commission, *Results of Survey of United States District Judges, January 2010 through March 2010*, giugno 2010, question 19, table 19, [https://bit.ly/3B8s62Q].

[31] C. Yang, *Have Interjudge Sentencing Disparities Increased in an Advisory Guidelines Regime? Evidence from Booker*, in "New York University Law Review", 89(2014), pp. 1268-1342; in part. 1278, 1334.

2

Un sistema rumoroso

La nostra prima esperienza in fatto di rumore, all'origine del nostro interesse per il tema, non aveva niente a che vedere con il sistema penale. Si trattava piuttosto di un piccolo incidente per risolvere il quale una compagnia assicurativa si era rivolta alla società di consulenza a cui due di noi erano affiliati.

Naturalmente non tutti sono avvezzi al mondo delle assicurazioni, ma i risultati a cui siamo pervenuti mostrano la portata del problema del rumore nelle organizzazioni a scopo di lucro, alle quali le decisioni rumorose possono costare care. La nostra collaborazione con la compagnia assicurativa aiuta a spiegare perché spesso questo problema passi inosservato e come si possa intervenire per risolverlo.

I dirigenti della compagnia stavano valutando l'opportunità di intraprendere azioni mirate per aumentare la coerenza interna – cioè per ridurre il rumore – tra i giudizi delle persone che prendevano importanti decisioni finanziarie per conto della società. Tutti concordavano sul valore della coerenza, ma anche sul fatto che tali giudizi non potessero mai essere del tutto coerenti, in quanto informali e in parte soggettivi. Un certo livello di rumore è inevitabile; ciò su cui non vi era accordo, però, era quanto quel livello fosse alto. I dirigenti dubitavano che il rumore potesse essere un serio problema per la loro società, ma saggiamente accettarono di risolvere la diatriba attraverso un semplice esperimento che chiameremo "controllo

del rumore". I risultati li sorpresero, e inoltre l'esperimento si rivelò una perfetta illustrazione del problema del rumore.

Una lotteria che crea rumore

In tutte le grandi società molti professionisti sono autorizzati a esprimere giudizi vincolanti per la società stessa. Questa compagnia assicurativa, per esempio, si avvale di numerosi sottoscrittori che stabiliscono l'importo dei premi per i vari rischi finanziari, come assicurare una banca contro perdite dovute a frodi o a operazioni disoneste. Impiega inoltre molti periti liquidatori, che prevedono il costo delle future richieste di risarcimento e conducono le trattative per le richieste effettive.

Ciascuna delle grandi filiali della società ha diversi sottoscrittori qualificati, e le richieste di preventivo vengono assegnate a quelli disponibili sul momento. A tutti gli effetti, si può dire che il sottoscrittore responsabile di un determinato preventivo venga estratto a sorte, come in una lotteria.

Il valore esatto del preventivo ha importanti conseguenze per la compagnia: un premio alto è vantaggioso se il preventivo viene accettato, ma espone al rischio di perdere il cliente a favore di un concorrente; per contro, un premio basso ha maggiori probabilità di essere accettato, ma è meno vantaggioso per la compagnia. Per ogni rischio c'è un prezzo ottimale, quello giusto, né troppo alto né troppo basso, e vi sono buone probabilità che il giudizio medio di un ampio gruppo di professionisti non si discosti troppo da questo optimum. I prezzi più alti o più bassi di questo valore comportano dei costi, ed è così che la variabilità dei giudizi rumorosi incide in negativo sul bilancio.

Anche il lavoro dei periti assicurativi ha un forte impatto sulle finanze della compagnia. Poniamo, per esempio, che venga presentata una richiesta di risarcimento per conto di un operaio (il richiedente) che ha subìto una perdita permanente dell'uso della mano destra in un incidente sul lavoro. Anche in questo caso, come abbiamo visto per i sottoscrittori, a una richiesta viene assegnato un determinato perito sulla base della disponibilità. Il perito raccoglie informazioni e fornisce una stima dei costi finali alla compagnia, dopodiché si occupa di avviare una trattativa con il rappresentante legale del richiedente per garantire che quest'ultimo riceverà l'indennizzo promesso nella polizza, al contempo tutelando la compagnia da richieste eccessive.

La stima iniziale è importante, perché stabilisce l'obiettivo implicito del perito nelle successive trattative con il richiedente. La compagnia assicurativa, inoltre, ha l'obbligo legale di accantonare il costo previsto di ogni richiesta di risarcimento (cioè di avere sempre abbastanza denaro disponibile per poterla soddisfare). Anche in questo caso, vi è un valore ottimale dal punto di vista della compagnia. Non è detto che si arrivi a un accordo, perché l'avvocato del richiedente potrebbe decidere di adire le vie legali se reputa l'offerta troppo misera; d'altro canto, un accantonamento eccessivamente prodigo potrebbe dare al perito troppa libertà di accogliere richieste poco serie. Il giudizio del perito, dunque, è dirimente per la compagnia, e lo è ancora di più per il richiedente.

Per sottolineare il ruolo della sorte nella selezione del sottoscrittore e del perito, utilizzeremo il termine "lotteria": nella gestione ordinaria della compagnia, a un caso viene assegnato un unico professionista, e a nessuno è dato sapere cosa sarebbe successo se al suo posto fosse capitato un altro collega.

Le lotterie hanno una loro funzione, e non è detto che siano inique. Quelle impiegate per assegnare "onori" come l'accesso a un corso universitario o "oneri" come la leva militare sono accettabili, in quanto hanno una loro finalità. Ma quelle coinvolte nei giudizi di cui parliamo non assegnano proprio niente: non fanno altro che produrre incertezza. È come se una compagnia assicurativa impiegasse sottoscrittori non affetti da rumore per il calcolo del premio ottimale, ma poi intervenisse un dispositivo di assegnazione casuale a modificare il preventivo che effettivamente arriva al cliente. È chiaro che una simile lotteria non avrebbe alcuna giustificazione. Come non è giustificabile un sistema in cui il risultato dipende dall'identità di una persona selezionata in maniera del tutto casuale per esprimere un giudizio professionale.

Il controllo del rumore rivela il rumore sistemico

La lotteria che estrae un particolare giudice per pronunciare una sentenza penale o un unico tiratore per rappresentare una squadra crea variabilità, ma tale variabilità passa inosservata. Un controllo del rumore come quello condotto sui giudici federali in rapporto alla determinazione della pena permette di far luce su quel rumore. Durante il controllo lo stesso caso viene valutato da più individui, di cui viene messa in evidenza la variabilità delle risposte.

I giudizi dei sottoscrittori e dei periti assicurativi si prestano particolarmente bene a questo esercizio, perché le loro decisioni si basano su informazioni scritte. In preparazione del controllo del rumore, i dirigenti della compagnia assicurativa elaborarono una descrizione dettagliata di cinque casi paradigmatici per ciascun gruppo (sottoscrittori e periti).[1] Ai dipendenti venne chiesto di valutare due o tre casi ciascuno, in

maniera indipendente, e non gli venne detto che lo studio mirava a esaminare la variabilità dei loro giudizi.

Prima di proseguire, provate a pensare a come rispondereste voi alle seguenti domande: in una compagnia assicurativa efficiente, selezionando in maniera casuale due sottoscrittori o periti qualificati, quanto vi aspettate che differiscano le loro stime per uno stesso caso? Andando più nello specifico, quale sarebbe la differenza percentuale tra le due stime rispetto alla loro media?

Queste stesse domande furono poste a numerosi dirigenti della compagnia e, negli anni, raccogliemmo le stime effettuate da un'ampia varietà di persone con diversi ruoli professionali. A sorpresa, una risposta ricorreva con maggiore frequenza di tutte le altre: la gran parte dei dirigenti della compagnia assicurativa ipotizzava uno scostamento uguale o inferiore al 10%. Quando chiedemmo a 828 amministratori delegati e alti dirigenti di un'ampia gamma di settori quale variazione si aspettassero di trovare in analoghi giudizi di esperti, anche lì la risposta mediana e la più frequente fu il 10% (al secondo posto c'era il 15%). Una differenza del 10% vorrebbe dire, per esempio, che uno dei due sottoscrittori stabilisce un premio di 9500 e l'altro di 10 500 dollari. Una differenza non trascurabile, ma che ci aspettiamo possa essere tollerata da un'organizzazione.

Il nostro controllo del rumore fece emergere differenze di gran lunga superiori. Secondo la nostra misurazione, la differenza mediana tra i sottoscrittori era del 55%, circa cinque volte superiore a quella ipotizzata dalla maggior parte delle persone, compresi gli stessi dirigenti della compagnia assicurativa. Questo risultato implica, per esempio, che quando un sottoscrittore stabilisce un premio di 9500 dollari, l'altro non lo fissa a 10 500 ma a 16 700 dollari. Per i periti assicurativi il rapporto mediano era del

43%. Notate che qui ci riferiamo ai valori mediani: nella metà dei casi la differenza tra i due giudizi era ancora più alta.

I dirigenti a cui riferimmo i risultati del nostro controllo capirono subito che l'entità del rumore riscontrato costituiva un problema in termini di costi. Uno di loro stimò che per la compagnia il costo annuale del rumore nelle sottoscrizioni – contando sia la perdita di affari per i preventivi eccessivi sia quella per i contratti sottocosto – era nell'ordine delle centinaia di milioni di dollari.

Nessuno era in grado di dire con precisione quale fosse il livello di errore (o di bias), perché nessuno poteva stabilire con sicurezza il valore ottimale per ciascun caso. I dati mostravano che il prezzo richiesto a un cliente dipende in misura sconvolgente dalla lotteria che assegna una determinata operazione a un determinato dipendente. I clienti non sarebbero molto contenti di sapersi coinvolti in una simile lotteria senza il proprio consenso: in generale, chi si affida a un'organizzazione si aspetta di trovare un sistema che formuli giudizi coerenti e affidabili; non si aspetta un rumore sistemico.

Variabilità indesiderata versus diversità voluta

Una caratteristica distintiva del rumore sistemico è il suo essere *indesiderato*, ma è bene sottolineare fin da ora che la variabilità nei giudizi spesso è tutt'altro che sgradita.

Prendiamo le questioni di gusto o di preferenza. Se dieci critici cinematografici guardano lo stesso film, se dieci assaggiatori valutano lo stesso vino, se dieci persone leggono lo stesso romanzo, non ci aspettiamo che abbiano tutti la stessa opinione. La diversità di gusto è benaccetta e del tutto prevista: nessuno (o quasi) vorrebbe vivere in un mondo in cui tutti

hanno esattamente le stesse preferenze. Ma questa stessa diversità può aiutare a spiegare gli errori che nascono quando il gusto personale viene confuso con il giudizio professionale. Se un produttore cinematografico decide di lanciarsi in un progetto insolito (poniamo sull'ascesa e il declino del telefono a disco) perché gli piace la sceneggiatura, potrebbe aver preso un grosso abbaglio se quel progetto non piace a nessun altro.

La variabilità di giudizio è prevista e gradita anche in situazioni competitive in cui i giudizi migliori verranno premiati: quando diverse società (o diversi gruppi di lavoro all'interno della stessa organizzazione) concorrono per arrivare a soluzioni innovative per uno stesso problema riscontrato dai clienti, non è auspicabile che adottino lo stesso approccio; quando diversi gruppi di ricercatori affrontano un problema scientifico come lo sviluppo di un vaccino, tutti noi ci auguriamo che lo guardino da prospettive diverse. Perfino gli esperti di previsioni di tanto in tanto sembrano porsi in concorrenza gli uni con gli altri. L'analista che annuncia correttamente una recessione che nessuno aveva previsto assurgerà di sicuro alla fama, mentre chi non si discosta mai dall'opinione dominante è condannato a restare nel buio. In questi contesti, di nuovo, la variabilità delle idee e dei giudizi è benvista, perché la variazione è solo il primo passo. In un secondo momento i risultati di questi giudizi verranno messi a confronto, e vincerà il migliore. Nel mercato, come in natura, non si dà selezione senza variazione.

Le questioni di gusto e gli ambienti competitivi pongono interessanti problemi di giudizio. Noi, però, abbiamo deciso di concentrarci sui giudizi in cui la variabilità *non è* desiderabile. Il rumore sistemico è un problema che riguarda appunto i sistemi, ovvero le organizzazioni, non i mercati. Quando gli operatori finanziari effettuano diverse stime del valore di un titolo azionario, alcuni di loro ci guadagneranno, altri no: il disaccordo è

alla base del mercato. Ma se uno di questi operatori viene scelto a caso per effettuare una valutazione per conto della sua società, e scopriamo che i suoi colleghi, nella medesima società, effettuerebbero valutazioni molto diverse, allora ci troviamo davanti a un rumore sistemico, il che è un problema.

Abbiamo ricevuto un'elegante illustrazione di questo fenomeno quando abbiamo presentato i nostri risultati agli alti dirigenti di una società di gestione patrimoniale, sollecitandoli a condurre loro stessi un controllo esplorativo del rumore. Loro hanno raccolto l'invito e chiesto a quarantadue investitori esperti della società di stimare il *fair value* di un titolo (il prezzo al quale per gli investitori sarebbe indifferente comprare o vendere). Gli investitori hanno basato la loro analisi su una breve descrizione dell'impresa quotata, che riportava il conto economico semplificato, lo stato patrimoniale e il rendiconto finanziario degli ultimi tre anni, e una proiezione dei due successivi. Il rumore mediano, ricavato in maniera analoga a quello della compagnia assicurativa, è risultato essere del 41%. Differenze di tale portata tra investitori di una stessa società che si avvalgono degli stessi metodi di valutazione non sono certo rassicuranti.

Ogni volta che chi esprime un giudizio viene selezionato a caso in un insieme di individui parimenti qualificati, come nel caso di questa società di gestione patrimoniale, del sistema penale e della compagnia assicurativa di cui si è parlato in precedenza, il rumore è un problema. Il rumore sistemico affligge molte organizzazioni: una procedura di assegnazione di fatto casuale spesso determina quale dottore vi assisterà in ospedale, a quale giudice verrà affidata la vostra causa in tribunale, quale esaminatore analizzerà la vostra domanda di brevetto, quale operatore del servizio clienti risponderà al vostro reclamo e così via. La variabilità indesiderata, in

questi giudizi, può causare seri problemi, da una semplice perdita di profitti fino a un'iniquità sociale dilagante.

Spesso si parte dal presupposto errato che la variabilità di giudizio indesiderata non sia importante, perché si presume che gli errori casuali si eliminino a vicenda. È vero, gli errori di giudizio positivi e negativi riguardanti uno stesso caso tenderanno a cancellarsi a vicenda, e spiegheremo nel dettaglio come sia possibile sfruttare questo aspetto per ridurre il rumore. I sistemi rumorosi, però, non esprimono molteplici giudizi sullo stesso caso, ma su casi diversi. Se una polizza assicurativa è troppo costosa e un'altra è troppo economica, in media il prezzo potrà sembrare giusto, ma la compagnia assicurativa avrà fatto due errori che le costeranno cari. Se due criminali che dovrebbero essere entrambi condannati a cinque anni di reclusione ricevono, rispettivamente, una condanna a tre anni e una a sette anni, non possiamo dire che, in media, sia stata fatta giustizia. Nei sistemi rumorosi, gli errori non si compensano. Si sommano.

L'illusione di accordo

Da vari decenni si va accumulando un'ampia letteratura sul rumore nel giudizio professionale. Poiché eravamo a conoscenza di questa letteratura, i risultati ottenuti dal controllo del rumore nella compagnia assicurativa non ci hanno sconvolto; ciò che ci ha sorpresi, invece, è stata la reazione dei dirigenti a cui abbiamo riferito tali risultati: nessuno di loro si aspettava un simile tasso di rumore. Nessuno ha messo in discussione la validità del controllo e nessuno ha preteso che il tasso di rumore riscontrato fosse accettabile, eppure il problema del rumore – e il suo costo – sembrava del tutto nuovo per l'organizzazione. Era come un'infiltrazione nello

scantinato: veniva tollerato non perché fosse ritenuto accettabile, ma perché nessuno se n'era accorto.

Com'è possibile? Come possono dei professionisti con lo stesso ruolo e la stessa posizione esprimere giudizi tanto diversi senza rendersene conto? Come mai i dirigenti non avevano osservato questo fenomeno, che pure riconoscevano come una grossa minaccia per le prestazioni e la reputazione della loro società? Ci siamo resi conto che il problema del rumore sistemico è spesso ignorato nelle organizzazioni, e che questa comune disattenzione nei confronti del rumore è tanto diffusa quanto interessante. I controlli del rumore suggerivano che degli stimati professionisti, e le organizzazioni che li assumevano, mantenevano un'"illusione di accordo", pur discordando, di fatto, nei giudizi professionali che formulavano quotidianamente.

Per provare a comprendere come nasce questa illusione di accordo, mettetevi nei panni di un sottoscrittore in una giornata lavorativa qualsiasi. Avete più di cinque anni di esperienza e sapete di essere stimati dai vostri colleghi, che voi stessi rispettate e apprezzate. Siete inoltre consapevoli di essere bravi nel vostro lavoro. Dopo avere analizzato con attenzione il complesso dei rischi a cui è esposta una società finanziaria, concludete che sia appropriato assegnarle un premio di 200 000 dollari. Il problema, per quanto complicato, non è molto diverso da quelli che vi trovate a risolvere tutti i giorni.

Immaginate ora che vi venga detto che i vostri colleghi di ufficio hanno ricevuto le stesse informazioni e valutato lo stesso rischio. Credereste che almeno la metà ha fissato un premio superiore a 255 000 o inferiore a 145 000 dollari? Non è facile da accettare. Anzi, sospettiamo che i sottoscrittori venuti a conoscenza del controllo del rumore, pur accettandone la validità,

non abbiano mai creduto che le nostre conclusioni li riguardassero personalmente.

Molti di noi vivono nella convinzione che il mondo sia esattamente come appare; da qui a credere che gli altri vedano il mondo come noi il passo è breve. Convinzioni come questa, che rientrano nel cosiddetto "realismo ingenuo",[2] sono fondamentali perché ci danno l'impressione che esista una realtà condivisa. Raramente le mettiamo in discussione. In ogni momento abbiamo una certa interpretazione del mondo che ci circonda, e di norma non ci sforziamo più di tanto per trovare delle alternative plausibili. Ci basta una sola interpretazione, che percepiamo come vera; non affrontiamo la vita immaginando modi alternativi di vedere la realtà.

Nel caso dei giudizi professionali, la convinzione che altri vedano il mondo proprio come noi viene rafforzata quotidianamente in vari modi. Innanzitutto condividiamo con i nostri colleghi un linguaggio comune e una serie di regole sulle considerazioni di cui tenere conto nelle decisioni, e a ciò si aggiunge l'esperienza rassicurante di essere d'accordo con gli altri sull'assurdità dei giudizi che violano queste regole. Interpretiamo i dissensi occasionali con i colleghi come loro errori di giudizio, e raramente abbiamo l'opportunità di riscontrare che queste regole condivise sono vaghe, sufficienti per scartare alcune possibilità ma non abbastanza specifiche da farci arrivare a una risposta comune nei vari casi particolari. Possiamo convivere serenamente con i nostri colleghi senza renderci mai conto che in realtà loro non vedono il mondo come noi.

Durante un colloquio, una sottoscrittrice ci parlò della sua lunga esperienza nel dipartimento a cui faceva capo: «Nei primi tempi mi confrontavo con il mio supervisore in tre casi su quattro. [...] Dopo qualche anno, non era più necessario. Ora sono considerata un'esperta. [...] Nel tempo sono diventata sempre più sicura dei miei giudizi». Come molti di

noi, questa persona ha maturato una piena fiducia nella propria capacità di giudizio sostanzialmente esercitandola.

La spiegazione psicologica alla base di questo processo è ben nota: la fiducia si acquista attraverso l'esperienza soggettiva di giudizi effettuati in maniera sempre più facile e disinvolta, anche perché simili a quelli già espressi in passato in casi analoghi. Nel tempo, imparando a entrare in sintonia con le sue stesse decisioni passate, questa sottoscrittrice è diventata più sicura dei propri giudizi. Non ha mai detto che, dopo la fase di apprendistato iniziale, ha imparato ad accordarsi con gli altri, ha verificato fino a che punto è d'accordo con loro o ha cercato di evitare che le sue decisioni si allontanassero da quelle dei suoi colleghi.

Per la compagnia assicurativa, soltanto il controllo del rumore fu in grado di spazzare via quell'illusione di accordo. Com'era possibile che i dirigenti della società fossero rimasti all'oscuro del problema del rumore? Le risposte possibili sono tante, ma quella che sembra preponderante, in molti contesti, è l'imbarazzo del disaccordo. La maggior parte delle organizzazioni preferisce il consenso e l'armonia al dissenso e al conflitto. Spesso le procedure vigenti sembrano progettate di proposito per ridurre al minimo le occasioni di reale disaccordo e, quando queste si verificano, accantonarle con una spiegazione rassicurante.

Nathan Kuncel, professore di psicologia alla University of Minnesota ed eminente studioso delle previsioni sulle prestazioni, ci ha raccontato una storia che illustra bene il problema. Kuncel stava aiutando l'ufficio ammissioni di una scuola a rivedere il proprio processo decisionale, che funzionava in questo modo: un membro del personale leggeva una domanda di ammissione, la valutava, e poi, dopo avere allegato la sua valutazione, la passava a un collega, che a sua volta esprimeva il suo parere. Kuncel suggerì, per motivi che capiremo più avanti, che sarebbe

stato preferibile occultare la valutazione del primo lettore, in modo da non influenzare il secondo. La scuola rispose: «Prima facevamo così, ma il disaccordo era tale che abbiamo deciso di passare al sistema attuale». Per questa scuola, come per altre organizzazioni, evitare il conflitto è importante almeno quanto prendere la giusta decisione.

Soffermiamoci su un altro meccanismo a cui fanno ricorso molte società: le analisi retrospettive dei giudizi infelici. Come forma di apprendimento, queste analisi sono utili. Ma se davvero è stato commesso un errore, nel senso che un giudizio ha trasgredito le norme della professione, parlarne non sarà un problema: gli esperti concluderanno agilmente che quel giudizio era lontanissimo dall'opinione generale. (Oppure lo liquideranno come una rara eccezione.) I giudizi errati sono molto più facili da identificare rispetto a quelli corretti. Riconoscere un errore marchiano ed emarginare i colleghi inetti non aiuterà i professionisti a riconoscere quanto sono in disaccordo su giudizi ritenuti ampiamente accettabili. Al contrario, il facile consenso sui giudizi sbagliati potrebbe perfino rafforzare l'illusione di accordo. In questo modo non si arriverà mai a comprendere la pervasività del rumore sistemico.

Ci auguriamo che conveniate con noi che il rumore sistemico sia un problema serio. La sua esistenza non ci sorprende: il rumore è una conseguenza della natura informale del giudizio; tuttavia, come vedremo nei prossimi capitoli, il livello di rumore che si osserva nelle organizzazioni dopo un'indagine accurata è quasi sempre sconcertante. Questo ci porta a una semplice conclusione: dove c'è giudizio, c'è rumore, e più di quanto non si pensi.

A proposito del rumore sistemico nelle compagnie assicurative

«Facciamo affidamento sulla qualità dei giudizi professionali di sottoscrittori, periti assicurativi e altri. Assegniamo ogni caso a un esperto, ma partiamo dal presupposto errato che un altro esperto arriverebbe a un giudizio simile.»

«Il rumore sistemico è cinque volte maggiore di quanto pensavamo, o di quanto siamo disposti a tollerare. Senza un controllo del rumore, non ce ne saremmo mai accorti. Il controllo del rumore ha spazzato via ogni illusione di accordo.»

«Il rumore sistemico è un problema serio: costa centinaia di milioni di dollari.»

«Dove c'è giudizio, c'è rumore, e più di quanto non si pensi.»

[1] I dirigenti della compagnia assicurativa elaborarono descrizioni dettagliate di casi rappresentativi, con rischi e richieste di risarcimento simili a quelli gestiti dai dipendenti nella vita reale: sei casi riguardavano il lavoro dei periti liquidatori del ramo danni e infortuni, e quattro quello dei sottoscrittori specializzati nel rischio finanziario. Ai dipendenti fu data mezza giornata di pausa dal normale carico di lavoro per valutare due o tre casi ciascuno, e fu chiesto loro di lavorare in maniera indipendente, senza rivelare che lo studio era finalizzato all'esame della variabilità dei loro giudizi. In tutto ottenemmo ottantasei giudizi da quarantotto sottoscrittori e centotredici giudizi da sessantotto periti liquidatori.

[2] D.W. Griffin, L. Ross, *Subjective Construal, Social Inference, and Human Misunderstanding*, in "Advances in Experimental Social Psychology", 24(1991), pp. 319-359; R.J. Robinson *et al.*, *Actual Versus Assumed Differences in Construal: 'Naive Realism' in Intergroup Perception and Conflict*, in "Journal of Personality and Social Psychology", 68(1995), n. 3, p. 404; L. Ross, A. Ward, *Naive Realism in Everyday Life: Implications for Social Conflict and Misunderstanding*, in E.S. Reed, E. Turiel, T. Brown (a cura di), *Values and Knowledge*, Lawrence Erlbaum Associates, Mahwah, NJ 1996.

3

Decisioni singole

I casi che abbiamo discusso finora chiamano in causa giudizi espressi ripetutamente. Qual è la giusta pena per chi viene condannato per furto? Qual è il giusto premio assicurativo a fronte di un certo rischio? Mentre ogni caso è in un certo senso un caso a sé, questo tipo di giudizi comporta *decisioni ricorrenti*: i medici fanno diagnosi sui pazienti, i giudici esaminano le richieste di libertà condizionale, i responsabili delle ammissioni considerano le candidature pervenute, i commercialisti predispongono le dichiarazioni dei redditi: tutti esempi di decisioni ricorrenti.

In questo tipo di decisioni è possibile dimostrare la presenza di rumore eseguendo un controllo simile a quello presentato nel capitolo precedente: la variabilità indesiderata è facile da definire e da misurare quando professionisti intercambiabili prendono decisioni su casi simili. Tuttavia sembra molto più difficile, o forse perfino impossibile, applicare l'idea di rumore a una classe di giudizi che chiamiamo *decisioni singole*.

Consideriamo, per esempio, la crisi sanitaria del 2014, quando in Africa occidentale molta gente stava morendo per il virus Ebola. Poiché il mondo è interconnesso, alcune proiezioni indicavano che le infezioni si sarebbero rapidamente diffuse in tutto il mondo e avrebbero colpito l'Europa e il Nordamerica con particolare intensità. Negli Stati Uniti c'era chi chiedeva con insistenza di sospendere i voli provenienti dalle regioni colpite e intervenire con decisione per chiudere i confini. Vi erano intense pressioni

politiche per compiere azioni di questo tipo, promosse da figure di spicco competenti in materia.

Il presidente Barack Obama si trovò di fronte a una delle decisioni più difficili della sua presidenza, una scelta che non si era mai trovato a dover prendere prima, né che avrebbe mai più dovuto prendere in seguito. Decise di non chiudere i confini, ma di inviare in Africa occidentale un contingente di tremila persone, tra operatori sanitari e soldati, e guidò una coalizione internazionale di varie nazioni non sempre affiatate, che impiegarono le loro risorse e competenze per contrastare il problema alla fonte.

Singolo versus ricorrente

Le decisioni che vengono prese una volta soltanto, come la risposta al virus Ebola del presidente Obama, sono singole perché non vengono prese regolarmente dallo stesso individuo o gruppo, non prevedono una risposta preconfezionata e si distinguono per caratteristiche del tutto eccezionali. Per gestire l'Ebola il presidente e la sua squadra non avevano precedenti su cui basarsi. Spesso le decisioni politiche più importanti costituiscono esempi di decisioni singole, al pari delle scelte fatidiche dei comandanti militari.

Nella sfera privata, le decisioni che prendiamo quando accettiamo un lavoro, acquistiamo una casa o facciamo una proposta di matrimonio presentano queste stesse caratteristiche. Anche se non si tratta di un primo lavoro, una prima casa o un primo matrimonio, si ha comunque il senso di essere di fronte a una decisione irripetibile. Nel mondo degli affari, spesso gli amministratori delle società sono chiamati a prendere quelle che a loro sembrano decisioni irripetibili: se lanciare o no un'innovazione potenzialmente dirompente, quali sedi chiudere durante una pandemia, se

aprire uno studio in un paese straniero o cedere a un governo che cerca di regolamentare la loro attività.

Verosimilmente vi è una continuità, non un salto di categoria, tra decisioni singole e ricorrenti: da una parte anche i sottoscrittori potranno incorrere in casi che reputeranno irripetibili o straordinari, dall'altra, se è la quarta volta che acquisti una casa, probabilmente un simile acquisto comincerà a sembrarti una decisione ricorrente. Ma questi sono esempi estremi, che indicano chiaramente che vi è una differenza notevole tra un tipo di decisione e l'altra. Andare in guerra è una cosa, procedere alla revisione del bilancio annuale un'altra.

Il rumore nelle decisioni singole

Di norma le decisioni singole vengono considerate tutt'altra cosa rispetto ai giudizi ricorrenti espressi da dipendenti intercambiabili di grandi organizzazioni. Se i giudizi ricorrenti sono di pertinenza delle scienze sociali, le decisioni singole ad alto rischio sono il terreno d'elezione della trattazione storica o dei guru del management. Gli approcci a questi due tipi di decisioni sono di solito molto diversi tra loro. Spesso le analisi delle decisioni ricorrenti prendono una piega statistica, con esperti di scienze sociali che valutano una serie di decisioni simili per individuare i modelli soggiacenti, identificare le regolarità e misurarne l'accuratezza. Al contrario, tipicamente il dibattito sulle decisioni singole si ferma a un'interpretazione superficiale: viene condotto a posteriori e si concentra sulle cause di ciò che è accaduto. Le analisi storiche, come lo studio dei successi e dei fallimenti aziendali, mirano a comprendere come sia stato espresso un giudizio sostanzialmente eccezionale.

La natura delle decisioni singole solleva un interrogativo importante per uno studio del rumore. Abbiamo definito il rumore come la variabilità indesiderabile nei giudizi relativi a uno stesso problema; poiché i problemi particolari non si ripropongono mai in maniera del tutto identica, nel loro caso questa definizione non è applicabile. Dopotutto, la storia non si ripete. Non potremo mai paragonare la decisione di Obama di inviare operatori sanitari e soldati in Africa occidentale nel 2014 con le decisioni prese da altri presidenti americani per gestire un altro problema particolare in un altro momento particolare (anche se si possono avanzare delle ipotesi). Potrei anche essere d'accordo sulla possibilità di paragonare la mia decisione di sposare qualcuno in particolare con quella presa da un'altra persona simile a me, ma questo paragone non sarà utile quanto quello tra i preventivi dei sottoscrittori per uno stesso caso, di cui abbiamo parlato nel capitolo precedente. Io e la mia metà siamo unici. Non è possibile osservare in forma diretta la presenza del rumore nelle decisioni singole.

Eppure questo tipo di decisioni non è esente dai fattori che producono rumore in quelle ricorrenti. Al tiro a segno i tiratori della squadra C (quella affetta da rumore) potrebbero piegare il mirino del fucile in direzioni diverse o magari non avere una presa ferma: se osservassimo solo il primo tiratore della squadra, non avremmo idea del livello di rumore complessivo della squadra, ma le fonti di rumore sarebbero comunque presenti. Analogamente, quando prendiamo una decisione singola, dobbiamo immaginare che, al posto nostro, un'altra persona con le nostre stesse competenze e i nostri medesimi obiettivi e valori non perverrebbe alla nostra stessa conclusione a partire dagli stessi dati. E, nel decidere, dovremmo riconoscere che avremmo potuto prendere una decisione diversa se qualche aspetto irrilevante della situazione o del processo decisionale fosse stato diverso.

Detto in altri termini, non è possibile misurare il rumore insito in una decisione singola, ma se pensiamo in maniera controfattuale, sappiamo per certo che il rumore c'è. Come la presa instabile del tiratore implica che un singolo colpo *sarebbe potuto* finire altrove, così il rumore presente nei decisori e nel processo decisionale implica che una decisione singola *avrebbe potuto* essere diversa.

Consideriamo tutti i fattori che influiscono su una decisione singola. Se gli esperti chiamati ad analizzare la minaccia posta dal virus Ebola e a predisporre i piani di intervento fossero stati persone diverse, con una formazione ed esperienze di vita diverse, la proposta da loro presentata al presidente Obama sarebbe stata la stessa? Se i medesimi fatti fossero stati presentati in maniera leggermente differente, il confronto si sarebbe svolto nello stesso modo? Se le figure chiave fossero state di umore diverso o si fossero incontrate durante una tempesta di neve, la decisione finale sarebbe stata un'altra? Vista in questa luce, la decisione singola non sembra così inesorabile: sulla base di vari fattori di cui non siamo neanche a conoscenza, avrebbe potuto benissimo differire del tutto da quella presa nella realtà.

Per fare un altro esempio di pensiero controfattuale, consideriamo la risposta alla pandemia da Covid-19 di paesi e regioni diversi: perfino quando il virus li ha colpiti più o meno allo stesso tempo e in maniera simile, ci sono state grosse differenze nelle reazioni. Questa variazione costituisce una prova evidente del rumore nel processo decisionale di paesi diversi. Ma cosa sarebbe accaduto se l'epidemia avesse colpito un solo paese? In quel caso non avremmo osservato alcuna variabilità, ma la nostra incapacità di osservarla non avrebbe reso la decisione meno rumorosa.

Controllare il rumore nelle decisioni singole

Questa discussione teorica è importante. Se le decisioni singole non sono meno rumorose di quelle ricorrenti, allora le strategie per ridurre il rumore nelle seconde dovrebbero migliorare anche la qualità delle prime.

Questa regola è più controintuitiva di quanto non sembri: quando dobbiamo prendere una decisione unica nel suo genere, d'istinto probabilmente la tratteremo come tale. Alcuni sostengono addirittura che le leggi del pensiero probabilistico siano del tutto irrilevanti nelle decisioni singole prese in condizioni di incertezza, e che tali decisioni richiedano un approccio radicalmente diverso.

Le osservazioni che abbiamo compiuto suggeriscono l'opposto. Nell'ottica della riduzione del rumore, *una decisione singola è una decisione ricorrente che avviene una volta sola*. Che prendiate una decisione una volta oppure cento, dovreste comunque puntare a ridurre sia il bias sia il rumore. E le azioni che riducono l'errore dovrebbero essere efficaci tanto nelle decisioni uniche quanto in quelle ripetute.

A proposito delle decisioni singole

«Il modo in cui ti poni nei confronti di un'opportunità insolita ti espone al rumore.»

«Ricorda: una decisione singola è una decisione ricorrente presa una volta sola.»

«Le esperienze personali che ti hanno reso chi sei non sono molto rilevanti per questa decisione.»

SECONDA PARTE

La mente come strumento di misura

Nella vita quotidiana come nella scienza, la misurazione consiste nell'impiego di uno strumento per assegnare un valore su una scala a un oggetto o a un evento. Misuriamo la lunghezza di un tappeto in centimetri usando un metro a nastro e la temperatura in gradi Celsius o Fahrenheit mediante un termometro.

La formulazione di un giudizio presuppone un'operazione simile: quando i giudici determinano il periodo di detenzione appropriato per un reato, assegnano un valore su una scala. Lo stesso dicasi per i sottoscrittori che stabiliscono un valore monetario per assicurare un rischio o i medici che fanno una diagnosi. (Non è detto che la scala sia numerica: anche «colpevole oltre ogni ragionevole dubbio», «melanoma avanzato» e «si consiglia intervento chirurgico» sono dei giudizi.)

Possiamo dunque descrivere il giudizio come una *misurazione il cui strumento è la mente umana.* Nell'idea di misurazione è implicito l'obiettivo dell'accuratezza: avvicinarsi alla verità riducendo al minimo l'errore. Lo scopo del giudizio, dunque, non è fare colpo, prendere posizione o convincere. È importante sottolineare che il concetto di giudizio a cui qui si fa riferimento è mutuato dalla letteratura psicologica, ed è molto più ristretto rispetto all'accezione comune del termine. Come *giudizio* non è un sinonimo di *pensiero*, così *dare giudizi accurati* non significa *avere giudizio.*

Nella nostra definizione, un giudizio è una conclusione riassumibile in una parola o una frase. Se un analista dell'intelligence scrive un lungo rapporto concludendo che un certo regime è instabile, solo la conclusione

può dirsi un giudizio. Il *giudizio*, come la *misurazione*, comprende sia l'attività mentale di formulare un giudizio sia il prodotto di tale attività. Di tanto in tanto useremo la parola "giudice" come termine tecnico per riferirci a persone che danno giudizi, anche quando non hanno niente a che fare con il sistema giudiziario.

Malgrado l'obiettivo sia l'accuratezza, la perfezione non si raggiunge né nella misurazione scientifica né tantomeno nel giudizio. Si incorre sempre in qualche errore, che si tratti di bias o di rumore.

Per sperimentare come il rumore e il bias contribuiscano all'errore, vi invitiamo a fare un gioco che vi ruberà meno di un minuto. Se avete uno smartphone con un cronometro, probabilmente avrete una funzione "lap timer" che permette di misurare intervalli consecutivi senza fermare il tempo o guardare lo schermo. Il vostro obiettivo è di registrare cinque giri consecutivi di esattamente dieci secondi senza guardare il telefono. Prima di cominciare potete osservare sullo schermo quanto durano un paio di intervalli di dieci secondi. Bene, ora partite.

Ora guardate la durata dei giri registrati sul telefono (che pure non è privo di rumore, ma ne risente pochissimo). Vedrete che i giri non sono tutti di dieci secondi esatti, ma, al contrario, variano sensibilmente. Avete cercato di riprodurre esattamente lo stesso tempo, ma non ci siete riusciti: la variabilità che non siete riusciti a controllare è un esempio di rumore.

Questo risultato non deve sorprendervi, perché il rumore è un dato universale, tanto sul piano della fisiologia che su quello della psicologia. La variabilità tra gli individui è un dato biologico; nello stesso baccello non troverete due piselli identici. Anche all'interno della stessa persona c'è una certa variabilità: il nostro battito cardiaco non è perfettamente regolare, non possiamo ripetere lo stesso gesto due volte con la massima precisione, e quando l'audiologo ci controlla l'udito, ci saranno dei suoni debolissimi

che non sentiremo mai e altri fortissimi che sentiremo sempre, ma anche suoni che certe volte sentiremo e altre no.

Ora guardate le cinque misurazioni sul vostro telefono. Notate uno schema ricorrente, per esempio tutti e cinque i giri sono inferiori a dieci secondi, cosa che sembrerebbe indicare che il vostro orologio interno va troppo veloce? In questo piccolo test il bias è dato dalla differenza, positiva o negativa, tra la media dei vostri giri e i dieci secondi. Il rumore, invece, è costituito dalla variabilità dei risultati, come nella dispersione dei colpi sul bersaglio vista in precedenza. In statistica la misura più comune della variabilità è la *deviazione standard*,[1] che impiegheremo per misurare il rumore nei giudizi.

La maggior parte dei giudizi, specialmente quelli *predittivi*, sono assimilabili alla misurazione che avete appena svolto. Quando facciamo una previsione, cerchiamo di avvicinarci a un valore reale: chi fa previsioni economiche, per esempio, punta ad avvicinarsi il più possibile, poniamo, al valore reale dell'aumento del prodotto interno lordo dell'anno successivo, mentre un medico punta a fare la diagnosi corretta. (Notiamo che il termine *previsione*, nell'accezione tecnica utilizzata in questo libro, non si riferisce necessariamente al futuro: per i nostri scopi, la diagnosi di una patologia esistente è una previsione.)

Adotteremo spesso questa analogia tra giudizio e misurazione, perché essa ci permette di spiegare il ruolo del rumore nell'errore. Chi formula un giudizio predittivo è come un tiratore che mira al centro del bersaglio, o un fisico che cerca di misurare il peso reale di una particella: il rumore nei suoi giudizi implica un errore. In sostanza, se l'obiettivo è arrivare a un valore reale, due misure diverse non potranno essere entrambe corrette. Proprio come accade per gli strumenti di misurazione, in generale certe persone cadono in errore più di altre in determinati compiti, magari a causa di

lacune nella loro preparazione o nelle loro competenze. Detto ciò, al pari di qualsiasi strumento di misurazione, un individuo che esprime un giudizio non è mai perfetto, e ciò che a noi preme è comprendere e misurare i suoi errori.

Naturalmente i giudizi professionali sono quasi sempre più complessi della misurazione di un intervallo di tempo. Nel prossimo capitolo definiremo diversi tipi di giudizi professionali e indagheremo sui loro obiettivi. Nel capitolo 5 parleremo di come misurare l'errore e quantificare l'entità del rumore sistemico. Il capitolo 6 approfondirà il tema del rumore sistemico identificandone le componenti, ovvero i diversi tipi di rumore, mentre nel capitolo 7 ci concentreremo su una di queste componenti: il rumore occasionale. Infine, nel capitolo 8 mostreremo come spesso i gruppi amplifichino il rumore nei giudizi.

Da questi capitoli emerge una semplice conclusione: come ogni strumento di misurazione, anche la mente umana è imperfetta in quanto affetta da bias e rumore. Perché e in quale misura? Scopriamolo insieme.

[1] La deviazione standard di un insieme di numeri deriva da un'altra misura statistica, chiamata *varianza*. Per calcolarla occorre prima ottenere la distribuzione delle deviazioni dalla media e poi calcolare il quadrato di ciascuna deviazione. La varianza è la media di queste deviazioni quadratiche, mentre la deviazione standard è la radice quadrata della varianza.

4

Questioni di giudizio

Questo libro prende in considerazione i giudizi professionali in senso lato, e parte dal presupposto che chiunque li formuli sia competente e voglia fare la cosa giusta. Tuttavia, il concetto stesso di giudizio costringe ad ammettere, a malincuore, che non si può mai essere certi della sua correttezza.

Spesso si parla di "una questione di giudizio" o di "un giudizio opinabile". Affermare che domani sorgerà il sole o che la formula chimica del cloruro di sodio è NaCl non è considerato una questione di giudizio, perché tra persone ragionevoli ci aspettiamo di trovarci perfettamente d'accordo su questo. Una questione di giudizio presuppone un'incertezza sulla risposta, nonché un possibile e plausibile disaccordo tra persone ragionevoli e competenti.

Ma c'è un limite al livello di disaccordo ammissibile. Anzi, la parola *giudizio* viene impiegata principalmente proprio quando la gente crede che debba esserci un accordo. In ciò le questioni di giudizio differiscono dalle questioni di opinione o di gusto, in cui le differenze inconciliabili sono del tutto accettate. I dirigenti assicurativi sconvolti dai risultati del controllo del rumore non avrebbero problemi se i periti liquidatori si trovassero in disaccordo sui rispettivi meriti dei Beatles e dei Rolling Stones, o sulla bontà del tonno e del salmone.

Le questioni di giudizio, anche in ambito professionale, si situano a metà tra le questioni di fatto o di calcolo da una parte, e le questioni di gusto o di

opinione dall'altra, e sono contraddistinte dall'*aspettativa di un disaccordo limitato*.

Quale esatto livello di disaccordo sia accettabile rispetto a un giudizio è esso stesso un giudizio opinabile e dipende dalla difficoltà del problema. L'accordo è particolarmente semplice quando un giudizio è assurdo. Anche dei giudici che emettono sentenze molto diverse per un comune caso di frode converranno sul fatto che una multa di un dollaro e l'ergastolo sono entrambe pene irrazionali. Nei concorsi enologici i giudici sono molto divisi sull'esperto da premiare, ma spesso sono del tutto d'accordo su chi eliminare.[1]

L'esperienza del giudizio: un esempio

Prima di proseguire nella disamina dell'esperienza del giudizio, vi chiediamo di esprimerne uno voi stessi. Comprenderete meglio il resto del capitolo se svolgerete questo esercizio fino alla fine.

> Immaginate di far parte di una commissione che ha il compito di valutare i candidati per il ruolo di amministratore delegato di una società finanziaria con un discreto successo su base regionale, ma soggetta a una concorrenza sempre più incalzante. Vi viene chiesto di valutare la probabilità che il seguente candidato avrà successo a due anni dall'assunzione, dove per "successo" si intende semplicemente la capacità di mantenere l'incarico per i suddetti due anni. Indicate tale probabilità su una scala da 0 (impossibile) a 100 (certo).
>
> Michael Gambardi ha trentasette anni. Nei dodici anni successivi alla laurea presso la Harvard Business School, ha ricoperto diversi incarichi aziendali. Ha esordito come fondatore e investitore di due startup, poi fallite senza attrarre grandi finanziamenti. In seguito è entrato in una grande compagnia assicurativa e presto ha raggiunto la posizione di direttore generale regionale per l'Europa. In quel ruolo ha attuato un'azione importante per la risoluzione tempestiva dei sinistri. I suoi colleghi e sottoposti lo descrivono come efficiente ma anche dispotico e irritante, e nel corso del suo mandato vi è stato un notevole ricambio di dirigenti. Allo stesso tempo, ne attestano l'onestà e l'accettazione delle proprie responsabilità a fronte di insuccessi aziendali. Negli ultimi

due anni ha ricoperto il ruolo di amministratore delegato di una società finanziaria di medie dimensioni sull'orlo del fallimento ed è riuscito a consolidare la compagnia, dove è ritenuto un uomo di successo con cui è difficile lavorare. Si è detto interessato a spostarsi. Gli specialisti in risorse umane con cui ha avuto un colloquio qualche anno fa hanno espresso una valutazione eccellente sulla sua creatività ed energia, ma lo hanno anche descritto come arrogante e talvolta perfino tirannico.

Ricordiamo che Michael si è candidato alla posizione di amministratore delegato di una società finanziaria regionale di discreto successo che deve far fronte a una concorrenza sempre più spietata. Qual è la probabilità che, se Michael venisse assunto, resterebbe in carica per i due anni successivi? Indicate un numero preciso tra 0 e 100, prima di proseguire. Se necessario, rileggete la descrizione.

Se avete preso sul serio questo compito, probabilmente lo troverete difficile. Abbiamo una grande quantità di informazioni, molte delle quali apparentemente discordanti, e bisogna faticare per formarsi un'idea coerente da cui partire per formulare un giudizio. Per arrivare a questa idea, vi siete concentrati su alcuni dettagli che vi sembravano importanti e molto probabilmente ne avete trascurati altri. Se doveste motivare la vostra indicazione numerica, chiamereste in causa alcuni elementi salienti, ma non sufficienti a dare conto in tutto e per tutto del vostro giudizio.

Il processo di pensiero che avete seguito illustra diverse caratteristiche delle operazioni mentali che chiamiamo giudizi:

- Tra tutti gli spunti forniti dalla descrizione (che sono solo una parte di quanto vi occorrerebbe sapere), ne avete colti alcuni più di altri, senza essere pienamente coscienti delle vostre scelte. Avete notato che Gambardi è un cognome italiano? Vi ricordate quale università ha frequentato? Questo esercizio era pensato proprio per sovraccaricarvi di informazioni in modo che non poteste recuperare facilmente tutti i dettagli del caso, ed è ben probabile che ciò che ricordate di questa presentazione sia diverso da ciò che ricordano altri lettori. L'attenzione selettiva e la memoria selettiva introducono una variabilità tra le persone.
- Poi, in maniera informale, avete integrato tutti questi spunti in un'impressione generale delle prospettive di Gambardi. Qui la parola chiave è *informale*. Non avete elaborato un modello per rispondere alla domanda: senza che ne aveste piena consapevolezza, la vostra mente ha lavorato per costruire un'impressione coerente dei punti di forza e di debolezza di Michael, e delle sfide che gli si presentano. Questa informalità vi ha permesso di agire rapidamente.

Inoltre ha prodotto una variabilità: un processo formale come sommare dei numeri disposti in colonna garantisce risultati identici, ma nelle operazioni informali il rumore è inevitabile.

- Infine, avete convertito questa impressione generale in un numero da collocare su una scala delle probabilità di successo. Associare un numero da 0 a 100 a un'impressione è un processo decisivo, su cui torneremo nel capitolo 14. Anche qui, non sapete esattamente perché avete dato una tale risposta. Perché avete scelto, poniamo, 65 e non 61 o 69? Probabilmente, a un certo punto vi è venuto in mente un numero, e vi siete chiesti se quel numero vi sembrasse giusto; in caso negativo, ve ne è venuto in mente un altro. Anche questa parte del processo introduce una variabilità tra le persone.

Poiché ognuna di queste tre fasi della formazione di un giudizio complesso comporta una certa variabilità, non dovremmo sorprenderci nel trovare molto rumore nelle risposte su Michael Gambardi. Se sottoponeste il caso a qualche amico, probabilmente avreste stime diversissime della sua probabilità di successo. Quando lo abbiamo presentato a centoquindici studenti di Business Administration, le loro stime della probabilità di successo di Gambardi variavano da un minimo di 10 a un massimo di 95. Insomma, una quantità enorme di rumore.

Per inciso, forse vi sarete accorti che l'esercizio con il cronometro e il caso Gambardi illustrano due tipi di rumore. La variabilità di giudizio nei diversi tentativi con il cronometro è un rumore interno al singolo giudicante (voi stessi), mentre quella relativa al caso Gambardi è un rumore tra giudici diversi. In termini di misurazione, il primo problema illustra l'affidabilità *intrapersonale*, il secondo quella *interpersonale*.

La finalità del giudizio: il segnale interno

La vostra risposta alla domanda su Gambardi costituisce un giudizio predittivo, nell'accezione sopra specificata. Detto ciò, differisce notevolmente da altri giudizi predittivi come la temperatura massima di domani a Bangkok, il risultato della partita di stasera o l'esito delle

prossime elezioni. Se non siete d'accordo con un vostro amico su queste questioni, a un certo punto scoprirete chi aveva ragione; ma se dissentite su Gambardi, il tempo *non* vi dirà chi aveva ragione, per un semplice motivo: Gambardi non esiste.

Anche se il quesito si riferisse a una persona reale e ne conoscessimo l'esito, un singolo giudizio di probabilità (diverso da 0 o 100%) non può essere confermato o smentito, perché l'esito non rivela qual era la probabilità *ex ante*. Se un evento a cui era stata assegnata una probabilità del 90% non accade, quel giudizio di probabilità non è necessariamente sbagliato. Dopotutto, i risultati che hanno il 10% di probabilità di verificarsi, il 10% delle volte si verificano. L'esercizio su Gambardi è un esempio di giudizio predittivo *non verificabile*, per due diversi motivi: Gambardi è un personaggio immaginario, e la risposta è probabilistica.

Molti giudizi professionali rientrano in questa categoria. Se si escludono gli errori madornali, i sottoscrittori non sapranno mai, per esempio, se una particolare polizza aveva un prezzo troppo alto o troppo basso. Altre previsioni possono essere non verificabili perché condizionali: «Se andremo in guerra, ci annienteranno» è una previsione importante, ma è probabile che non venga provata (speriamo). Vi sono poi previsioni così a lungo termine che chi le elabora non potrà mai darne conto, come, per esempio, la stima delle temperature medie alla fine del ventunesimo secolo.

La non verificabilità dell'esercizio su Gambardi ha cambiato il vostro modo di affrontarlo? Vi siete chiesti, per esempio, se Gambardi fosse reale o fittizio? O se l'esito vi sarebbe stato rivelato nelle pagine successive? Avete riflettuto sul fatto che, anche in quel caso, quella rivelazione non avrebbe risposto alla domanda iniziale? Probabilmente no, perché queste considerazioni non vi sembravano pertinenti quando avete dato la vostra risposta.

La verificabilità non cambia l'esperienza del giudizio. Per certi versi, forse rifletterete di più su un problema di cui presto saprete la risposta, perché il timore di essere sbugiardati vi farà concentrare. D'altro canto, potreste anche rifiutarvi di arrovellarvi su un problema talmente ipotetico da essere assurdo («Se Gambardi avesse tre gambe e sapesse volare, sarebbe un amministratore migliore?»). Ma, in genere, ci si accosta a un problema ipotetico plausibile come se fosse vero. Questa somiglianza è importante per la ricerca psicologica, che impiega spesso problemi immaginari.

Poiché non c'è alcun esito, e probabilmente non vi siete neanche chiesti se ci sarebbe stato, non avete cercato di minimizzare l'errore rispetto a quell'esito, ma vi siete preoccupati soltanto di dare un giudizio corretto, di arrivare a un numero che vi desse abbastanza fiducia da poter sceglierlo come risposta. Naturalmente, non era la stessa fiducia che avreste avuto rispondendo a chi vi avesse chiesto quanto fa sei per quattro. Eravate consapevoli di una qualche incertezza (e, come vedremo, probabilmente eravate più incerti di quanto non sapeste), ma a un certo punto avete deciso che non potevate fare di meglio e vi siete accontentati di una data risposta.

Cosa vi ha fatto credere che foste arrivati al giudizio giusto, o almeno abbastanza giusto da sceglierlo come risposta? Noi riteniamo che questa sensazione sia un *segnale interno di completamento del giudizio*, indipendente da qualsiasi informazione esterna. La vostra risposta vi è sembrata giusta se avete avuto l'impressione che fosse abbastanza compatibile con i dati fattuali. Una risposta pari a 0 o a 100 non vi avrebbe dato questa impressione di compatibilità, in quanto avrebbe implicato una fiducia non coerente con i dati confusi, ambigui e discordanti forniti. Ma il numero che avete stabilito, qualunque sia, vi ha dato l'impressione di coerenza di cui

avevate bisogno. La finalità del giudizio, nella vostra esperienza, era il raggiungimento di una soluzione coerente.

La caratteristica essenziale di questo segnale interno è che l'impressione di coerenza fa parte dell'esperienza del giudizio. Non è subordinata a un esito reale. Di conseguenza, il segnale interno si attiva tanto per i giudizi non verificabili quanto per quelli reali e verificabili, e ciò spiega perché dare un giudizio su un personaggio fittizio come Gambardi non sembri diverso dal dare un giudizio sul mondo reale.

Come viene valutato il giudizio: l'esito e il processo

La verificabilità non cambia l'esperienza del giudizio sul momento, ma modifica la sua valutazione a posteriori.

I giudizi verificabili possono essere valutati da un osservatore oggettivo misurandone l'errore, ovvero la differenza tra il giudizio e il risultato. Se, stando alle previsioni del tempo, la massima di oggi sarebbe stata di 21 gradi e invece è di 18, l'errore è di +3 gradi. Evidentemente, questo approccio non funziona con i giudizi non verificabili come il problema di Gambardi, che non ha un vero esito. Come fare, allora, a stabilire quando un giudizio può dirsi buono?

La risposta è che c'è un secondo modo di valutare i giudizi, che vale sia per quelli verificabili sia per quelli non verificabili, e consiste nella valutazione del *processo* di giudizio. Quando parliamo di giudizi buoni o cattivi, possiamo riferirci o all'esito, per esempio il numero a cui siete arrivati nel caso Gambardi, o al processo, cioè cosa avete fatto per arrivare a quel numero.

Un possibile approccio alla valutazione del processo di giudizio sta nell'osservare se quel processo funziona quando applicato a un gran

numero di casi. Per esempio, consideriamo un esperto di politica che ha effettuato delle previsioni sui risultati alle elezioni locali di un'ampia rosa di candidati, assegnando a cento di loro una probabilità di vittoria del 70%: se, alla fine, settanta di loro verranno eletti, avremo una buona indicazione del talento dell'esperto nell'impiego della scala di probabilità. I giudizi di probabilità sono verificabili nell'insieme, anche se non è possibile dichiararne giusto o sbagliato uno in particolare. Analogamente, si può riscontrare un bias a favore o contro un particolare gruppo esaminando i risultati statistici di un numero consistente di casi.

Un'altra questione sollevata dal processo di giudizio è se sia conforme o no ai princìpi della logica o alla teoria della probabilità. Numerose ricerche sui bias cognitivi di giudizio si interrogano su questo aspetto.

Concentrarsi sul processo di giudizio invece che sull'esito permette di valutare la qualità dei giudizi non verificabili, come quelli su problemi fittizi o le previsioni a lungo termine: forse non saremo in grado di rapportarli a un risultato noto, ma possiamo comunque dire se sono stati formulati in maniera scorretta. Anche quando affronteremo il problema di *migliorare* i giudizi piuttosto che valutarli, ci concentreremo sul processo. Tutte le procedure per ridurre il bias e il rumore che suggeriamo in questo libro puntano all'adozione di un processo di giudizio che minimizzi l'errore in un corpus di casi simili.

Abbiamo posto a confronto due modi di valutare un giudizio: confrontandolo con un *esito* e valutando la qualità del *processo* con cui vi siamo arrivati. Si noti che, quando il giudizio è verificabile, le due modalità di valutazione potrebbero portare a conclusioni differenti per lo stesso caso. Un esperto abile e accorto che si avvale dei migliori strumenti e tecniche possibili spesso nelle sue previsioni trimestrali sull'inflazione non arriverà al valore corretto, ma, su un singolo trimestre, uno scimpanzé che

scegliesse tra varie opzioni lanciando delle freccette potrebbe anche azzeccare.

Per risolvere questa tensione tra i due diversi tipi di valutazione, gli studiosi dei processi decisionali danno una chiara indicazione: concentrarsi sul processo, non sul singolo risultato. Ammettiamo, tuttavia, che nella vita reale non sempre si segue questa linea: normalmente i professionisti vengono valutati sulla base di quanto i loro giudizi si avvicinino agli esiti verificabili, e se si chiede loro a cosa puntano, la risposta sarà proprio questa: avvicinarsi il più possibile al risultato misurabile.

In sostanza, di solito nei giudizi verificabili le persone sostengono di tendere a una previsione che coincida con l'esito. Ma ciò che in realtà cercano di raggiungere, a prescindere dalla verificabilità, è il segnale interno di completamento attivato dalla coerenza tra i dati fattuali del caso e il giudizio. Di conseguenza, ciò a cui dovrebbero mirare, come regola generale, è il processo che porti al miglior giudizio su un insieme di casi simili.

Giudizi valutativi

Fin qui in questo capitolo ci siamo concentrati su esercizi relativi al giudizio predittivo, e molti dei giudizi che prenderemo in esame appartengono a questa tipologia. Nel capitolo 1, in cui si parlava del giudice Frankel e del rumore nel processo di determinazione della pena dei giudici federali, abbiamo invece esaminato un altro tipo di giudizio. Emettere una condanna penale non è una previsione, ma un *giudizio valutativo* che cerca di assegnare una pena in linea con la gravità del reato. I giudici di un concorso di enologia o di una gara di tuffi esprimono giudizi valutativi, e lo stesso vale per i professori che attribuiscono un voto a un esame, i giudici delle

gare di pattinaggio sul ghiaccio e le commissioni che assegnano finanziamenti ai progetti di ricerca.

Un diverso tipo di giudizio valutativo è rappresentato dalle decisioni basate su opzioni multiple, ognuna con i suoi pro e i suoi contro. Pensiamo agli amministratori che devono selezionare un candidato da assumere, ai gruppi dirigenti che devono decidere tra più opzioni strategiche o anche a un presidente che deve capire come agire davanti a un'epidemia in Africa. Certo, tutte queste decisioni si basano sull'input di giudizi predittivi – per esempio, quale sarà la performance del candidato nel corso del primo anno, come reagirà il mercato azionario a una certa mossa strategica o con quale velocità si diffonderà l'epidemia se non si interviene –, ma le decisioni finali implicano un confronto tra i pro e i contro delle varie opzioni, confronto che si risolve in un giudizio valutativo.[2]

Come nei giudizi predittivi, anche in quelli valutativi ci si aspetta un grado di disaccordo limitato. Nessun giudice federale con un po' di autostima dirà: «Questa è la pena che ho stabilito, e non mi interessa se i miei colleghi la pensano diversamente». E i dirigenti che scelgono tra diverse opzioni strategiche possibili si aspettano che altri colleghi e osservatori in possesso delle stesse informazioni e con gli stessi obiettivi siano d'accordo con loro, o almeno che non dissentano troppo. I giudizi valutativi dipendono in una certa misura dai valori e dalle preferenze di chi li formula, ma non sono semplicemente una questione di gusto personale.

Per questi motivi il confine tra giudizi predittivi e valutativi è labile, e chi si trova a esprimere un giudizio spesso non ne è consapevole. I giudici che stabiliscono una condanna o i professori che valutano un esame riflettono molto sul proprio compito e cercano di fare la scelta "giusta", sviluppando una certa fiducia nei propri giudizi e nelle giustificazioni che li sostengono. Anche i professionisti pensano, agiscono e parlano in questi termini per

giustificare i propri giudizi predittivi («Quanto venderà questo nuovo prodotto?») e valutativi («Quali risultati ha raggiunto il mio assistente quest'anno?»).

Il problema del rumore

La presenza del rumore nei giudizi predittivi indica sempre che c'è qualcosa che non va. Se due medici sono in disaccordo su una diagnosi o due esperti fanno previsioni diverse sulle vendite del prossimo trimestre, almeno uno di loro sarà in errore; ciò può avvenire perché uno di loro è meno competente, quindi più incline a sbagliare, o a causa di qualche altra fonte di rumore. L'errore può comportare gravi perdite, che si tratti di salute o di denaro. A prescindere dalla motivazione, un giudizio non corretto può avere serie conseguenze su chi dovrà basarsi sulle diagnosi e le previsioni di queste persone.

Il rumore nei giudizi valutativi è problematico per un altro motivo. In ogni sistema in cui si presume che gli esperti siano intercambiabili e assegnati in maniera pressoché casuale, un ampio disaccordo sullo stesso caso tradisce le aspettative di coerenza e imparzialità. Se vi sono grandi differenze nelle condanne inferte a uno stesso imputato, entriamo nel campo delle «crudeltà arbitrarie» denunciate dal giudice Frankel. Perfino due giudici convinti del valore della pena personalizzata, nel caso si trovassero in disaccordo sulla condanna inflitta a un rapinatore, concorderanno sul fatto che se il livello di disaccordo è tale da trasformare un giudizio in una lotteria, c'è un problema. Lo stesso vale (con conseguenze meno drammatiche) nel caso in cui alla stessa prova d'esame vengano assegnati voti decisamente diversi, allo stesso ristorante valutazioni diverse sul rispetto delle norme di sicurezza, alla stessa

pattinatrice punteggi diversi, o quando una persona che soffre di depressione riceve una pensione di invalidità civile e un'altra con la stessa patologia no.

Perfino quando l'imparzialità ha conseguenze meno gravi, il rumore sistemico pone alcuni problemi. Chi riceve un giudizio valutativo si aspetta che i valori riflessi da tale giudizio siano quelli del sistema, non del singolo valutatore. C'è qualcosa che non va se un cliente che sporge un reclamo per un computer difettoso riceve un rimborso integrale e un altro solo una lettera di scuse, o se un dipendente che ha lavorato per cinque anni in un'azienda ottiene una promozione, mentre a un altro con un'identica performance la stessa promozione viene rifiutata. Il rumore sistemico sta proprio in questa incoerenza, che danneggia la credibilità del sistema.

Indesiderabile ma misurabile

Per misurare il rumore occorre semplicemente disporre di più giudizi relativi allo stesso problema. Non serve conoscere un valore reale. Come illustra l'esempio del tiro a segno dell'introduzione, quando osserviamo il retro del bersaglio, il centro è invisibile, ma la dispersione dei tiri è evidente. Una volta capito che tutti i tiratori puntavano allo stesso centro, sarà possibile misurare il rumore. È così che agisce il controllo del rumore: se chiediamo a un gruppo di esperti di elaborare una stima delle vendite del prossimo trimestre, il rumore sarà dato dal livello di dispersione delle loro previsioni.

Questa differenza tra bias e rumore è essenziale se ci si pone la finalità pratica di migliorare i giudizi. Potrà sembrare paradossale sostenere che sia possibile migliorare i giudizi quando non si è in grado di verificare se siano giusti o sbagliati, eppure si può fare, se si comincia con il misurare il

rumore. Che l'obiettivo del giudizio sia la semplice accuratezza o un più complesso bilanciamento tra valori diversi, il rumore è sempre indesiderabile e spesso misurabile. E una volta misurato, come vedremo nella parte 5, in genere può essere ridotto.

A proposito del giudizio professionale

«È una questione di giudizio. Non puoi aspettarti che tutti siano perfettamente d'accordo.»

«Sì, questa è una questione di giudizio, ma alcuni giudizi sono talmente fuorvianti da essere sbagliati.»

«La tua scelta del candidato era solo una questione di gusto, non un giudizio serio.»

«Una decisione richiede sia giudizi predittivi sia giudizi valutativi.»

[1] R.T. Hodgson, *An Examination of Judge Reliability at a Major U.S. Wine Competition*, in "Journal of Wine Economics", 3(2008), n. 2, pp. 105-113.

[2] Alcuni studiosi dei processi decisionali definiscono le decisioni come scelte tra più opzioni, e ritengono i giudizi quantitativi uno speciale caso di decisione in cui vi è un continuum di scelte possibili. In quest'ottica, i giudizi sono tipi particolari di decisioni. Qui adottiamo un approccio diverso: riteniamo che le decisioni basate su una scelta tra più opzioni discendano da un giudizio valutativo sottostante riguardo a ciascuna opzione. Consideriamo, cioè, le decisioni come un particolare tipo di giudizio.

5

Misurare l'errore

Va da sé che un bias persistente può causare grossi errori: se una bilancia aggiunge sempre un tot al vostro peso, o un manager pieno di entusiasmo puntualmente prevede che un progetto richiederà metà del tempo effettivamente richiesto o, anno dopo anno, un dirigente cauto è troppo pessimista sulle vendite future, si incorrerà in errori seri.

Ma arrivati a questo punto, abbiamo visto come anche il rumore possa causare grossi problemi: se un manager prevede quasi sempre che un progetto richiederà metà del tempo effettivo e alcune volte che richiederà il doppio del tempo, non ci aiuta sapere che il suddetto manager ha "mediamente" ragione. Gli errori si accumulano, non si cancellano a vicenda.

È importante, quindi, chiedersi in che modo e in quale misura il bias e il rumore contribuiscano all'errore. Questo capitolo si propone di trovare una risposta alla domanda. Il concetto alla base è molto semplice: in ogni tipo di giudizio professionale che punti all'accuratezza, *il bias e il rumore contribuiscono al calcolo dell'errore complessivo esattamente allo stesso modo.* In certi casi inciderà di più il bias, in altri (più spesso di quanto non si creda) il rumore, ma, in tutti i casi, la riduzione del rumore avrà lo stesso impatto sull'errore complessivo di un'analoga riduzione del bias. Per questo misurare e ridurre il rumore dovrebbe essere non meno prioritario di misurare e ridurre il bias.

Questa conclusione si fonda su un particolare metodo di misurazione dell'errore, un metodo di lunga data generalmente accettato in ambito scientifico e statistico. In questo capitolo ripercorreremo la sua storia, con qualche accenno ai calcoli matematici su cui è basato.

GoodSell dovrebbe ridurre il rumore?

Per cominciare, immaginiamo una grande società di distribuzione di nome GoodSell, che si avvale di molti esperti per le previsioni di vendita, con il compito di prevedere quale sarà la quota di mercato dell'azienda in varie regioni. Forse dopo aver letto un libro su questo tema, Amy Simkin, direttrice del dipartimento previsioni di GoodSell, ha condotto un controllo del rumore, chiedendo a tutti gli esperti di fornire una stima individuale della quota di mercato che verrà raggiunta in una data regione.

La figura 3 mostra i risultati (inverosimilmente omogenei) del controllo del rumore. Amy nota che le previsioni assumono la classica distribuzione a campana, anche nota come distribuzione normale o gaussiana.

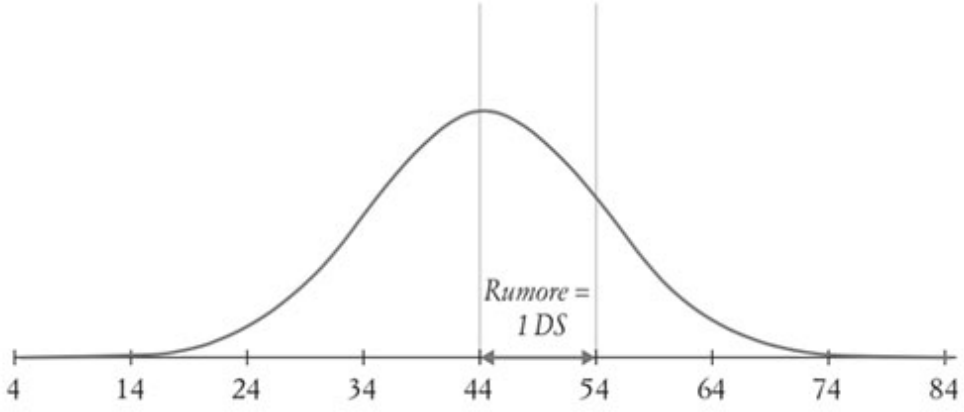


Figura 3. Distribuzione delle previsioni sulla quota di mercato di GoodSell in una data regione

La previsione più frequente, rappresentata dal picco della curva a campana, si attesta su una quota di mercato del 44%. Amy osserva inoltre che il

sistema di previsioni della società è piuttosto rumoroso: le previsioni, che se fossero tutte accurate sarebbero identiche, variano in misura rilevante.

Possiamo associare un numero al livello di rumore del sistema di previsioni di GoodSell: proprio come abbiamo fatto per gli intervalli di tempo nell'esperimento con il cronometro, possiamo calcolare la *deviazione standard* (DS) delle previsioni. Quest'ultima, come si evince dal nome, indica lo scostamento tipico dalla media, che in questo esempio è pari a dieci punti percentuali. Come in ogni distribuzione normale, circa due terzi delle previsioni si distanziano dalla media di una deviazione standard su entrambi i lati – in questo esempio, concentrandosi tra quote di mercato del 34% e del 54%. Ora Amy ha una stima del livello di rumore sistemico nelle previsioni della quota di mercato. (Un controllo del rumore più raffinato terrebbe conto di vari problemi previsionali per arrivare a una stima più solida, ma per i nostri fini ne basta uno.)

Come i dirigenti della compagnia assicurativa del capitolo 2, anche Amy è sconvolta dai risultati e intende prendere provvedimenti: questo inaccettabile livello di rumore indica che gli esperti non applicano con metodo le procedure che sono tenuti a seguire. Amy chiede quindi ai suoi superiori di rivolgersi a un consulente per raggiungere un'uniformità e un rigore maggiori nell'attività dei suoi esperti, ma purtroppo la sua richiesta non viene approvata. La risposta del suo capo sembra sensata: come possiamo ridurre l'errore, le chiede, senza sapere se le nostre previsioni sono giuste o sbagliate? Se nelle previsioni c'è un errore medio elevato (cioè un bias elevato), è ovvio che dovremmo innanzitutto risolvere quello, le dice. Prima di intraprendere qualsiasi azione per migliorare le nostre previsioni, conclude, dovremmo aspettare per capire se sono corrette.

A distanza di un anno dal primo controllo del rumore si apprende il risultato reale su cui gli esperti avevano fatto previsioni: la quota di mercato nella regione considerata è stata del 34%. Ora è noto anche l'errore di

ciascun esperto, ovvero la differenza tra la previsione e il risultato. L'errore è 0 per una previsione del 34%, è pari al 10% per la previsione media del 44%, ed è pari al –10% per una previsione a ribasso del 24%.

La figura 4 mostra la distribuzione degli errori. È la stessa di quella riportata nella figura 3, ma da ogni previsione è stato sottratto il valore reale (34%). La forma del grafico non è cambiata, e la deviazione standard (la nostra misura del rumore) è ancora pari al 10%.

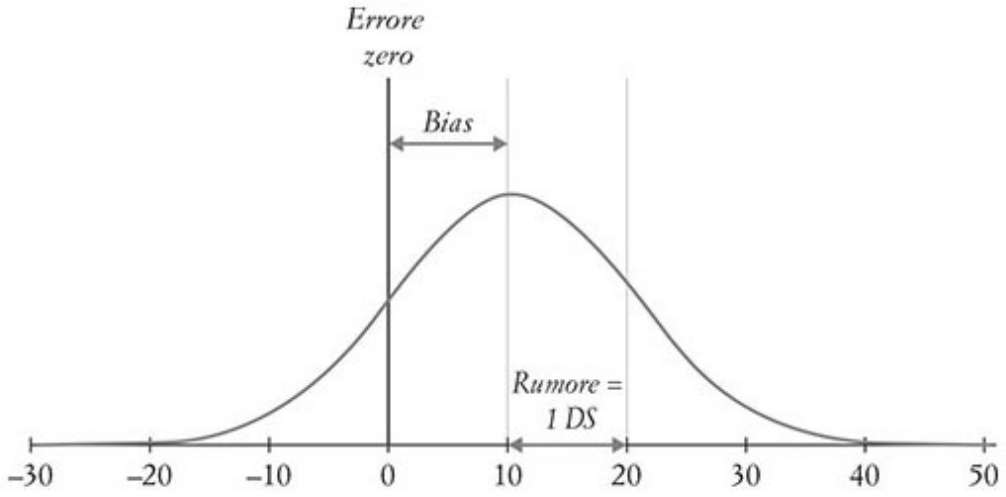


Figura 4. Distribuzione degli errori nelle previsioni di GoodSell per una data regione

La differenza tra le figure 3 e 4 è analoga a quella tra la disposizione dei tiri al bersaglio vista da dietro o davanti nelle figure 1 e 2 (vedi Introduzione). Così come non era necessario conoscere la posizione del bersaglio per osservare il rumore nei colpi, conoscere l'esito reale non aggiunge nulla a quanto già sapevamo sul rumore presente nelle previsioni degli esperti di GoodSell.

Amy Simkin e il suo capo ora hanno un elemento in più: il livello di bias nelle previsioni. Il bias non è altro che la media degli errori, in questo esempio pari anch'essa al 10%. Per caso, quindi, in questa serie di dati bias e rumore sono numericamente identici. (Sia chiaro che questa equivalenza non è assolutamente la norma, ma il fatto che qui abbiano uno stesso valore

ci permette di capire meglio il loro ruolo.) Notiamo che la maggior parte degli esperti ha commesso un errore ottimistico, cioè ha sovrastimato la quota di mercato che sarebbe stata raggiunta: gli errori si situano per lo più a destra della linea verticale dell'errore zero. (Applicando le proprietà della distribuzione normale, scopriamo infatti che questo è vero nell'84% dei casi.)

Con malcelata soddisfazione il capo di Amy le fa notare che aveva ragione lui: nelle previsioni vi era una grande distorsione. Ora è ancora più evidente, quindi, che il bias andrebbe ridotto. Ma Amy continua a chiedersi se non sarebbe stata una buona idea, un anno fa come ora, ridurre anche il rumore. Che valore avrebbe un simile miglioramento rispetto a una riduzione del bias?

Minimi quadrati

Per rispondere alla domanda di Amy occorre un "criterio di valutazione" degli errori, un metodo per ponderare e aggregare i singoli errori in un'unica misurazione dell'errore complessivo. Fortunatamente disponiamo di un simile strumento: si tratta del *metodo dei minimi quadrati,* elaborato nel 1795[1] da Carl Friedrich Gauss, il noto prodigio della matematica nato nel 1777 che avviò le sue grandi scoperte negli anni dell'adolescenza.

Gauss propose un criterio per valutare quanto incide ogni singolo errore sull'errore complessivo. La sua misura dell'errore complessivo, chiamato *errore quadratico medio* (in inglese *mean squared error* o MSE), corrisponde alla media dei quadrati dei singoli errori di misurazione.

La dettagliata argomentazione di Gauss a sostegno di questo approccio alla misurazione dell'errore complessivo esula dalle finalità di questo libro e la sua soluzione non è di immediata evidenza. Perché allora impiegare i quadrati degli errori? Sembra un'idea bizzarra e del tutto campata in aria.

Eppure, come vedremo, poggia su un'intuizione su cui non potrete che convenire anche voi.

Per motivare questa scelta, spostiamoci su un problema apparentemente molto diverso ma in realtà identico. Immaginate che vi si dia un righello e vi si chieda di misurare la lunghezza di una linea con precisione millimetrica. Potete svolgere cinque misurazioni, rappresentate dai triangoli rovesciati disposti sopra la linea nella figura 5.

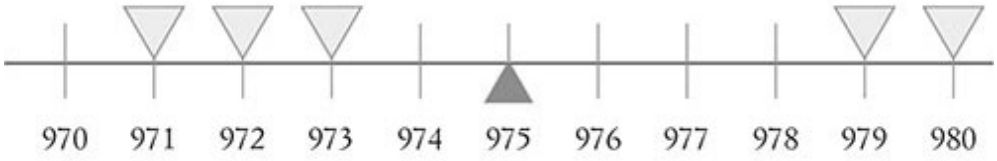


Figura 5. Cinque misurazioni della stessa lunghezza

Come vedete, le cinque misurazioni si situano tutte tra 971 e 980 millimetri. Qual è la vostra stima della vera lunghezza della linea? Le risposte più ovvie sono due. Una possibilità è il valore della mediana, la misurazione che sta tra le due inferiori e le due superiori, pari a 973 millimetri. L'altra possibilità è la media aritmetica, che in questo esempio è pari a 975 millimetri, qui indicata con un triangolo con la punta verso l'alto. È probabile che intuitivamente optereste per la media, e non sbaglierete, perché la media contiene più informazioni: tiene conto dell'entità dei numeri, mentre la mediana considera solo il loro ordine.

Vi è uno stretto legame tra questo problema di valutazione, sul quale avete una chiara intuizione, e quello della misurazione dell'errore complessivo che qui ci interessa. In effetti, sono le due facce della stessa medaglia. Questo perché la stima migliore è quella che riduce al minimo l'errore complessivo delle misurazioni disponibili. Pertanto, se la vostra intuizione che la media offra la stima migliore è corretta, per misurare

l'errore complessivo dovreste scegliere una formula che vi restituisca proprio la media aritmetica in quanto valore che riduce al minimo l'errore.

L'errore quadratico medio è l'unica definizione dell'errore complessivo ad avere questa proprietà. Nella figura 6 abbiamo calcolato il valore dell'errore quadratico medio delle cinque misurazioni per dieci valori interi possibili della lunghezza reale della linea. Se, per esempio, il valore reale fosse 971, gli errori nelle cinque misurazioni sarebbero 0, 1, 2, 8 e 9. I quadrati di questi errori, sommati, sono pari a 150, e la loro media è 30. È un numero elevato, a riprova del fatto che alcune misure sono lontane dal valore reale. Come vedete, l'errore quadratico medio si riduce man mano che ci avviciniamo a 975 (la media) e torna ad aumentare oltre quel punto. La media è la stima migliore perché è il valore che riduce al minimo l'errore complessivo.

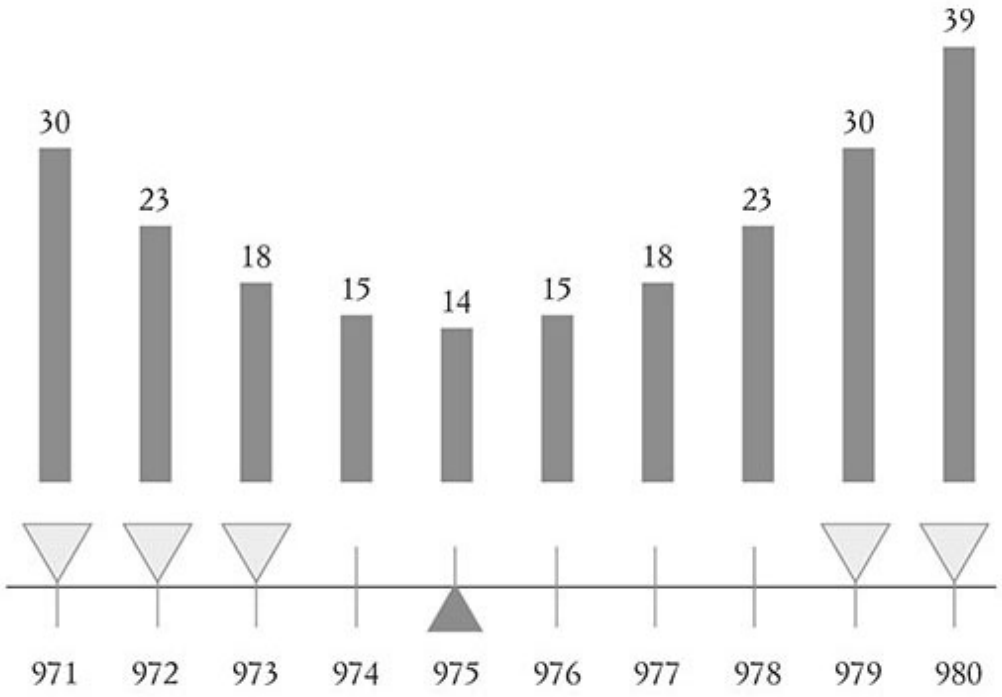


Figura 6. Errore quadratico medio (MSE) per dieci possibili valori della lunghezza reale

Vedete, inoltre, che l'errore complessivo aumenta rapidamente quando la vostra stima si scosta dalla media. Quando aumenta di soli 3 millimetri, per

esempio da 976 a 979, l'errore quadratico medio raddoppia. Questa è una sua caratteristica fondamentale: l'elevazione al quadrato dà agli errori più grandi un peso di gran lunga superiore rispetto a quello degli errori più piccoli.

Ora capiamo perché la formula di Gauss per misurare l'errore complessivo è chiamata errore quadratico medio e perché questo approccio è chiamato metodo dei minimi quadrati. L'elevazione al quadrato degli errori è un'idea cruciale, e qualsiasi altra formula sarebbe incompatibile con la vostra intuizione che la media rappresenti la stima migliore.

I vantaggi dell'approccio di Gauss furono presto riconosciuti da altri matematici. Tra i suoi tanti colpi di genio, Gauss impiegò l'errore quadratico medio (e altre innovazioni matematiche) per sciogliere un enigma di fronte al quale si erano arresi i più importanti astronomi d'Europa: la riscoperta di Cerere, un asteroide che era stato localizzato per breve tempo per poi scomparire dopo essere entrato in congiunzione con il sole nel 1801. Gli astronomi cercarono di ipotizzare la sua traiettoria, ma il metodo con cui calcolavano l'errore di misurazione dei loro telescopi era sbagliato, e il pianeta non riapparve nel luogo indicato dai loro risultati. Gauss ricalcolò l'errore impiegando il metodo dei minimi quadrati. Quando gli astronomi puntarono i telescopi verso il punto da lui indicato, trovarono Cerere.

Presto il metodo dei minimi quadrati fu adottato da scienziati di varie discipline, e a distanza di oltre due secoli, è ancora la modalità standard per valutare gli errori quando si punta alla massima accuratezza. La ponderazione degli errori a partire dai loro quadrati è centrale nelle analisi statistiche, e l'errore quadratico medio è alla base della maggior parte delle applicazioni di tutte le discipline scientifiche. Come vedremo, le implicazioni di questo approccio sono sorprendenti.

Le equazioni di errore

Il ruolo del bias e del rumore nell'errore complessivo è riassumibile in due espressioni che chiameremo *equazioni di errore*. La prima scompone l'errore presente in una singola misurazione nelle due componenti che ormai vi saranno note: il bias (l'errore medio) e un "errore da rumore" residuale. L'errore da rumore è positivo quando è superiore al bias, negativo quando è inferiore. La media degli errori da rumore è zero. La prima equazione di errore, quindi, non ci dice niente di nuovo.

Errore in una singola misurazione = Bias + Errore da rumore

La seconda equazione di errore è una scomposizione del metodo dei minimi quadrati, ovvero la misura dell'errore complessivo appena illustrata. Con un semplice calcolo algebrico è possibile dimostrare che l'errore quadratico medio è pari alla somma dei quadrati del bias e del rumore.[2] (Ricordiamo che il rumore è la deviazione standard delle misurazioni, che è identica alla deviazione standard degli errori da rumore.) Pertanto:

Errore complessivo (MSE, o errore quadratico medio) = Bias^2 + Rumore^2

La struttura di questa equazione (una somma di due quadrati) forse vi farà tornare in mente il classico teorema di Pitagora: in un triangolo rettangolo la somma dei quadrati costruiti sui cateti è equivalente al quadrato costruito sull'ipotenusa. Questa analogia vi aiuterà a visualizzare l'equazione di errore, in cui errore quadratico medio, quadrato del bias e quadrato del rumore sono le aree dei quadrati costruiti sui tre lati di un triangolo rettangolo. La figura 7 mostra come l'errore quadratico medio (l'area del

quadrato più scuro) sia pari alla somma delle aree degli altri due quadrati. Nell'immagine a sinistra il rumore è superiore al bias; in quella a destra è superiore il bias, ma l'errore quadratico medio è lo stesso, e l'equazione di errore regge in entrambi i casi.

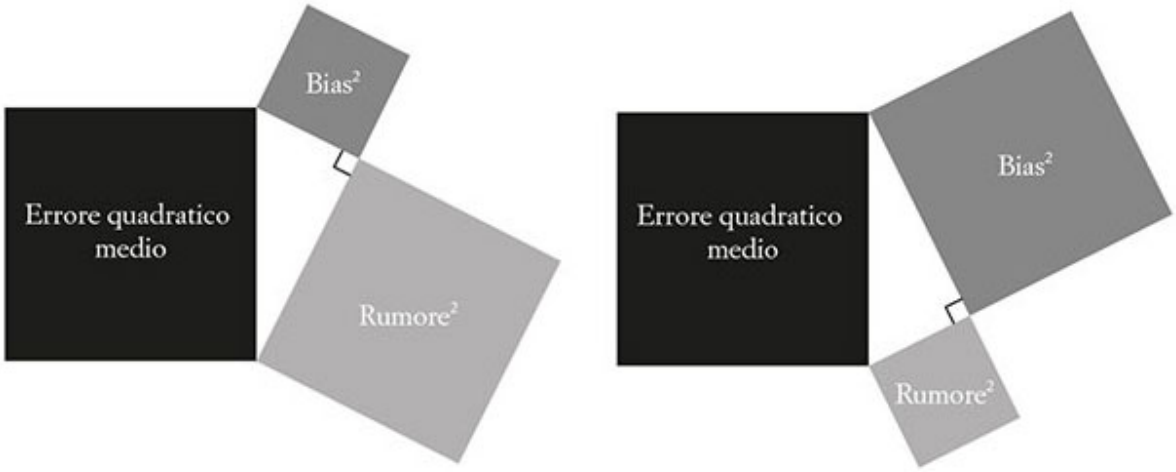


Figura 7. Due scomposizioni dell'errore quadratico medio

Come suggeriscono l'espressione matematica e la sua rappresentazione grafica, il bias e il rumore occupano un ruolo identico nell'equazione di errore. Sono indipendenti l'uno dall'altro e hanno lo stesso peso nella determinazione dell'errore complessivo. (Si noti che nei capitoli successivi, quando andremo ad analizzare le componenti del rumore, impiegheremo una simile scomposizione in una somma di quadrati.)

L'equazione di errore fornisce una risposta alla domanda pratica sollevata da Amy: come incide sull'errore complessivo un'analoga riduzione del rumore o del bias? La risposta è evidente: bias e rumore sono intercambiabili nell'equazione di rumore, quindi, che si limiti l'uno o l'altro, la riduzione dell'errore complessivo sarà la medesima. Nella figura 4, a fronte di un bias e un rumore identici (entrambi pari al 10%), anche il rispettivo apporto all'errore complessivo è identico.

L'equazione di errore va indubbiamente a sostegno dell'intuizione di Amy Simkin che occorresse ridurre il rumore. Ogni volta che ci accorgiamo del rumore dovremmo sforzarci di ridurlo! L'equazione mostra inoltre che il capo di Amy si sbagliava a dire che GoodSell dovesse aspettare di misurare il bias nelle previsioni per prendere provvedimenti. Rispetto all'errore complessivo, il rumore e il bias sono indipendenti: il beneficio di una riduzione del rumore non dipende dal livello di bias.

Questa idea è profondamente controintuitiva, ma cruciale. Per illustrarla, nella figura 8 mostriamo l'effetto della riduzione di uno stesso livello di bias e di rumore. Per facilitare la comprensione dei risultati espressi nei due grafici, la distribuzione degli errori originaria (quella riportata nella figura 4) è illustrata con una linea tratteggiata.

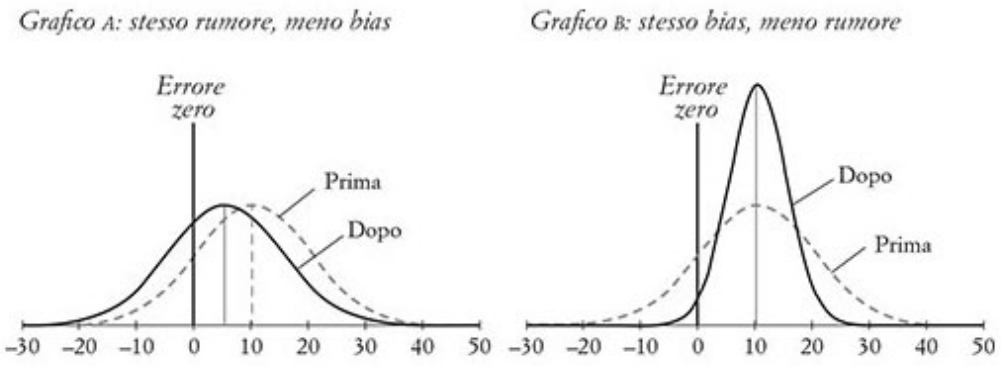


Figura 8. Distribuzione degli errori con un dimezzamento del bias e con un dimezzamento del rumore

Nel grafico A presumiamo che il capo di Amy abbia deciso di procedere in questo modo: una volta scoperta l'origine del bias, è riuscito a dimezzarlo (forse dando un riscontro agli esperti troppo ottimisti), ma nulla è cambiato riguardo al rumore. Il miglioramento è evidente: l'intera distribuzione delle previsioni si è avvicinata al valore reale.

Nel grafico B mostriamo cosa sarebbe successo se la richiesta di Amy fosse stata accolta: il bias sarebbe rimasto invariato, mentre il rumore si sarebbe

dimezzato. Qui il paradosso è che la riduzione del rumore sembra aver peggiorato la situazione: ora le previsioni sono più concentrate (meno rumorose) ma non più accurate (non meno affette da bias). Se prima l'84% delle previsioni era prossimo al valore reale, ora quasi tutte (il 98%) commettono un errore di sopravvalutazione. La riduzione del rumore sembra aver reso le previsioni ancora più errate. Non certo il miglioramento a cui aspirava Amy.

Malgrado le apparenze, tuttavia, l'errore complessivo si è ridotto di uguale misura nei due grafici. L'illusione che nel grafico B vi sia stato un peggioramento sorge da un'intuizione errata sul bias: la misura del bias che qui ci interessa non è il rapporto tra errori positivi e negativi, ma l'errore medio, che corrisponde alla distanza tra il picco della curva a campana e il valore reale. Nel grafico B questo errore medio non è cambiato rispetto alla situazione iniziale: è sempre del 10%, ancora alto ma non peggiore. Certo, ora la presenza del bias colpisce di più, perché copre una porzione superiore dell'errore complessivo (l'80% invece del 50%), ma questo accade perché il rumore è stato ridotto. Al contrario, nel grafico A è stato ridotto il bias ma non il rumore. Pertanto, l'errore quadratico medio è lo stesso in entrambi i grafici: ridurre il rumore o il bias del medesimo livello ha lo stesso effetto sull'errore quadratico medio.

Come illustra questo esempio, l'errore quadratico medio si pone in contrasto con le intuizioni più comuni sulla valutazione dei giudizi predittivi. Per ridurlo al minimo bisogna puntare a evitare gli errori più grandi. Nella misurazione della lunghezza, per esempio, l'effetto di ridurre un errore da 11 a 10 centimetri è 21 volte superiore all'effetto di passare da un errore di 1 centimetro a una misurazione perfetta. Purtroppo in questo caso l'intuito ci porta nella direzione opposta a quella auspicabile: si tende a voler arrivare a tutti i costi a una misurazione perfetta e quindi a essere molto sensibili ai piccoli errori, ma non si bada affatto alla differenza tra due

grossi errori.[3] Anche se mirate davvero a formulare un giudizio accurato, la vostra reazione emotiva ai risultati potrebbe essere incompatibile con l'obiettivo scientifico dell'accuratezza.

Naturalmente qui la soluzione migliore sarebbe ridurre tanto il rumore quanto il bias: poiché si tratta di due fenomeni indipendenti, non c'è motivo di scegliere tra la posizione di Amy Simkin e quella del suo capo. Quindi se GoodSell decide di ridurre il rumore, il fatto che tale riduzione renda più visibile – anzi, lampante – il bias potrebbe rivelarsi un grosso vantaggio. Dopo aver ridotto il rumore, di sicuro la società si porrà come prossimo obiettivo la riduzione del bias.

È anche vero che, qualora il bias fosse molto più elevato del rumore, ridurre quest'ultimo sarebbe una priorità meno urgente. Ma l'esempio di GoodSell ci permette di sottolineare un altro punto importante. In questo modello semplificato siamo partiti dal presupposto di un livello identico di rumore e bias. Vista la struttura dell'equazione di errore, è identico anche il loro contributo all'errore totale: incidono entrambi del 50%. Eppure, come abbiamo osservato, l'84% degli esperti ha commesso un errore che va in una certa direzione, ovvero quella di una stima eccessivamente ottimistica. Occorre un livello così alto di bias (sei persone su sette che commettono un errore nella stessa direzione!) per arrivare allo stesso effetto del rumore. Non dovremmo sorprenderci, quindi, di fronte a situazioni in cui il rumore sia superiore al bias.

Abbiamo illustrato l'applicazione dell'equazione di errore a un caso singolo, una particolare regione in cui opera GoodSell, ma naturalmente è sempre auspicabile eseguire un controllo del rumore in contemporanea su più casi. Il procedimento è lo stesso: l'equazione di errore viene applicata ai singoli casi, e poi si ottiene un'equazione generale dai valori medi di MSE, quadrato del bias e quadrato del rumore nei vari casi. Sarebbe stato meglio se Amy Simkin avesse avuto a disposizione previsioni multiple per più

regioni, dagli stessi esperti o da esperti diversi; la media dei risultati le avrebbe fornito un quadro più accurato del bias e del rumore nel sistema di previsioni di GoodSell.

Il costo del rumore

L'equazione di errore costituisce il fondamento concettuale di questo libro. È la *ratio* alla base dell'obiettivo della riduzione del rumore sistemico nei giudizi predittivi, un obiettivo che, in linea di principio, non è meno importante della riduzione del bias statistico. (Si noti che il termine "bias statistico" non è un sinonimo di discriminazione sociale, ma indica semplicemente l'errore medio in una serie di giudizi.)

L'equazione di errore e le conclusioni che ne abbiamo tratto si basano sull'impiego dell'errore quadratico medio come misura dell'errore generale. Tale metodo è senz'altro appropriato per i giudizi puramente predittivi, comprese le previsioni e le stime, che puntano ad avvicinarsi a un valore reale con la massima accuratezza (cioè il minor bias) e la massima precisione (cioè il minor rumore) possibili.

L'equazione di errore, tuttavia, non è applicabile ai giudizi valutativi, perché in quel caso è più difficile impiegare il concetto di errore, che dipende dall'esistenza di un valore reale. Peraltro, anche se si riuscissero a identificare gli errori, il loro costo sarebbe raramente simmetrico e difficilmente proporzionale al loro quadrato.

Per una ditta che produce ascensori, per esempio, le conseguenze di eventuali errori nella stima del carico massimo sarebbero ovviamente asimmetriche: sottostimarlo avrebbe un costo, ma sovrastimarlo potrebbe rivelarsi catastrofico. L'errore quadratico è altrettanto irrilevante quando si tratta di decidere a che ora uscire di casa per prendere un treno: in questa decisione le conseguenze di arrivare alla stazione in ritardo di uno o cinque

minuti sono le stesse. E quando la compagnia assicurativa citata nel capitolo 2 stabilisce il prezzo di una polizza o stima l'ammontare di un risarcimento, gli errori per eccesso e per difetto hanno entrambi un costo, ma non è detto che sia lo stesso.

Questi esempi mettono in luce la necessità di specificare qual è il ruolo dei giudizi predittivi e valutativi nelle decisioni. Secondo una massima ampiamente accettata, quando si prende una decisione non si dovrebbero mischiare fatti e valori. Bisognerebbe decidere sulla base di giudizi predittivi obiettivi e accurati, non influenzati da speranze e paure o da preferenze e valori. Per la ditta produttrice di ascensori il primo passo da compiere sarebbe un calcolo oggettivo del carico massimo tecnicamente ammissibile in diverse soluzioni ingegneristiche. La sicurezza diventerà una considerazione dirimente solo nella seconda fase, quando un giudizio valutativo determinerà la scelta di un margine di sicurezza accettabile per stabilire la capacità massima. (A dire il vero tale scelta dipenderà anche molto da giudizi fattuali quali, per esempio, quelli riguardo a costi e benefici di detto margine di sicurezza.) Analogamente, il primo passo per decidere quando avviarsi verso la stazione dovrebbe essere una determinazione obiettiva delle probabilità di tempi di spostamento diversi. I costi relativi di perdere il treno e arrivare troppo presto in stazione entrano in gioco solo nella scelta del rischio che siete disposti a correre.

La stessa logica è applicabile a molte altre decisioni importanti. Un comandante dell'esercito deve considerare diversi fattori quando si tratta di decidere se lanciare un'offensiva o no, ma gran parte dei dati di intelligence di cui dispone rientrano nella sfera dei giudizi predittivi. Un governo che reagisce a una crisi sanitaria come una pandemia deve valutare i pro e i contro di varie azioni possibili, ma non può effettuare alcuna valutazione se non dispone di previsioni accurate sulle probabili conseguenze di ognuna di queste scelte (compresa quella di non intervenire).

In tutti questi esempi le decisioni finali richiedono dei giudizi valutativi: chi decide deve considerare varie opzioni e applicare i propri valori per compiere la scelta ottimale. Ma le decisioni dipendono dalle relative previsioni, che dovrebbero essere slegate da qualsiasi valore. Se il loro obiettivo è l'accuratezza, cioè avvicinarsi il più possibile al centro del bersaglio, l'errore quadratico medio è la misura più appropriata dell'errore. Le procedure per ridurre il rumore potranno migliorare i giudizi predittivi, a patto che non aumentino ulteriormente il bias.

A proposito dell'equazione di errore

«Stranamente, ridurre il bias e il rumore della stessa misura ha il medesimo effetto in termini di accuratezza.»

«Ridurre il rumore nei giudizi predittivi è sempre utile, che si conosca o meno il bias.»

«Quando la distribuzione dei giudizi è tale che l'84% si colloca al di sopra e il 16% al di sotto del valore reale, il bias è molto accentuato. È in casi come questo che bias e rumore si equivalgono.»

«Ogni decisione comporta giudizi predittivi che dovrebbero mirare unicamente all'accuratezza. I fatti vanno sempre tenuti separati dai valori.»

[1] Il primo a scrivere sul metodo dei minimi quadrati fu Adrien-Marie Legendre nel 1805. Gauss sostenne di averlo utilizzato già dieci anni prima, e in seguito lo collegò allo sviluppo di una teoria dell'errore e alla curva normale che porta il suo nome. Questa disputa sulla primogenitura ha generato un grande dibattito tra gli storici, che propendono per la versione di Gauss. Vedi S.M. Stigler, *Gauss and the Invention of Least Squares*, in "Annals of Statistics", 9(1981), pp. 465-474; Id., *The History of Statistics: The Measurement of Uncertainty Before 1900*, Belknap Press of Harvard University Press, Cambridge, MA 1986.

[2] Abbiamo definito il rumore come la deviazione standard degli errori; pertanto, il rumore al quadrato è la varianza degli errori. La definizione di *varianza* è "la media dei quadrati meno il quadrato della media". Poiché l'errore medio è il bias, il "quadrato della media" è il quadrato del bias, da cui: $\text{Rumore}^2 = \text{Errore quadratico medio} - \text{Bias}^2$.

[3] B.J. Dietvorst, S. Bharti, *People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error*, in "Psychological Science", 31(2020), n. 10, pp. 1302-1314.

6

L'analisi del rumore

Nel capitolo precedente si è parlato della variabilità nella misurazione o nel giudizio relativi a un singolo caso. Quando ci concentriamo su un singolo caso, la variabilità di giudizio costituisce un errore, scomponibile a sua volta in bias e rumore. Naturalmente i sistemi di giudizio che stiamo esaminando, compresi quelli che coinvolgono tribunali e compagnie assicurative, sono concepiti per gestire casistiche diverse tra cui operare un discrimine: i giudici federali e i periti liquidatori non sarebbero molto utili se esprimessero lo stesso giudizio per tutti i casi che si trovassero ad affrontare. Molte volte la variabilità di giudizi su casi diversi è intenzionale.

Tuttavia, la variabilità su uno stesso caso non è auspicabile e costituisce un rumore sistemico; come vedremo, un controllo del rumore in cui si chieda alle stesse persone di esprimere giudizi su vari casi permette di pervenire a un'analisi più dettagliata di tale rumore.

Un controllo del rumore sulle sentenze

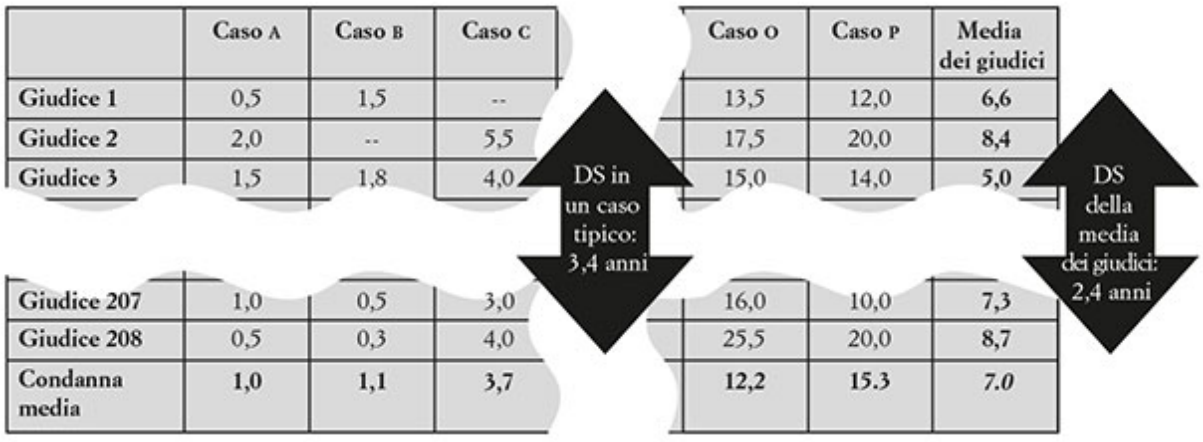
Per illustrare l'analisi del rumore in casi multipli, consideriamo un dettagliatissimo controllo del rumore condotto sulle sentenze dei giudici federali, i cui risultati furono pubblicati nel 1981 a sostegno del movimento per la riforma del sistema di determinazione della pena descritto nel capitolo 1.[1] Lo studio si concentrava sulle decisioni dei giudici, ma se ne possono trarre insegnamenti più generali, estendibili ad altri giudizi

professionali. Il controllo del rumore mirava ad andare oltre le evidenze del rumore – lampanti ma aneddotiche – raccolte dal giudice Frankel e da altri, allo scopo di «determinare il livello di disparità nelle sentenze» in maniera più sistematica.

Gli autori elaborarono sedici casi ipotetici basati su imputati giudicati colpevoli in attesa di condanna. Le illustrazioni ritraevano rapine o casi di frode che si differenziavano su sei dimensioni: per esempio, l'imputato poteva essere l'esecutore materiale o un complice del reato, avere o non avere precedenti penali, avere usato un'arma (nel caso della rapina) o no, e così via.

I ricercatori organizzarono dei colloqui ben strutturati con un campione nazionale di 208 giudici federali in attività. Nel corso di novanta minuti a ogni giudice venivano presentati tutti e sedici i casi e gli veniva chiesto di emettere una sentenza.[2]

Per comprendere appieno i risvolti di questo studio, riteniamo possa essere utile provare a visualizzare i risultati. Immaginate una grande tabella con sedici colonne per i reati, a cui attribuiremo una lettera dalla A alla P, e 208 righe per i giudici, numerate da 1 a 208. Ogni cella, da A1 a P208, mostra la pena detentiva comminata da un certo giudice in una certa causa penale. La figura 9 illustra una possibile configurazione di queste 3328 celle. Per studiare il rumore ci concentreremo sulle sedici colonne, ciascuna delle quali equivale a un singolo controllo del rumore.



	Caso A	Caso B	Caso C	Caso O	Caso P	Media dei giudici
Giudice 1	0,5	1,5	--	13,5	12,0	6,6
Giudice 2	2,0	--	5,5	17,5	20,0	8,4
Giudice 3	1,5	1,8	4,0	15,0	14,0	5,0
Giudice 207	1,0	0,5	3,0	16,0	10,0	7,3
Giudice 208	0,5	0,3	4,0	25,5	20,0	8,7
Condanna media	**1,0**	**1,1**	**3,7**	**12,2**	**15.3**	***7.0***

Figura 9. Una rappresentazione dello studio sulle sentenze

Condanna media

Non esiste un metodo oggettivo per definire quale sia il "valore reale" di una condanna in un caso particolare. In questo paragrafo considereremo come condanna "giusta" per ciascun caso la media delle 208 condanne (ovvero la condanna media). Come si diceva nel capitolo 1, la Sentencing Commission americana partì da questo stesso presupposto quando adottò la prassi media dei casi passati come punto di partenza per l'elaborazione di linee guida per la determinazione della pena. Questo assunto postula un valore di bias pari a zero nel giudizio medio su ogni caso.

Siamo pienamente consapevoli che, in realtà, questo sia un presupposto sbagliato: è ben probabile che il giudizio medio in alcuni casi sia distorto rispetto al giudizio medio su altri casi molto simili, per esempio sulla base di una discriminazione razziale. La varianza delle distorsioni da caso a caso – a volte positiva, altre negativa – è una notevole fonte di errore e parzialità. Ad aumentare la confusione si aggiunge il fatto che questa varianza viene spesso chiamata proprio "bias".[3] In questo capitolo, e in questo libro in generale, la nostra analisi si concentra sul rumore, che costituisce una diversa fonte di errore. Il giudice Frankel pose grande enfasi sull'ingiustizia

insita nel rumore, pur sottolineando al contempo la presenza del bias (inclusa la discriminazione razziale). Analogamente, il nostro focus sul rumore non intende sminuire l'importanza della misurazione e della lotta ai bias condivisi.

Per comodità, la condanna media di ciascun caso è indicata nell'ultima riga della tabella. I casi sono posti in ordine di gravità crescente: la condanna media nel caso A è di un anno, nel caso P è di 15,3 anni. La pena detentiva media di tutti e sedici i casi è di 7 anni.[4]

Per cominciare ipotizziamo un mondo perfetto in cui tutti i giudici sono impeccabili strumenti di misurazione della giustizia e la determinazione della pena non è affetta da alcun rumore. In simili circostanze come si presenterebbe la figura 9? Evidentemente tutte le celle della colonna del caso A sarebbero identiche, perché tutti i giudici attribuirebbero all'imputato del caso A la stessa condanna di un anno esatto. Lo stesso varrebbe per tutte le altre colonne: i numeri di ciascuna riga, naturalmente, varierebbero, perché si tratta di casi diversi, ma ogni riga sarebbe identica alla precedente e alla successiva. La differenza tra i casi sarebbe l'unica fonte di variabilità della tabella.

Purtroppo il mondo della giustizia non è perfetto. I giudici non sono identici e la variabilità all'interno di ogni colonna è elevata, indicando la presenza di rumore nei giudizi espressi per ciascun caso. Nelle condanne c'è maggiore variabilità del dovuto, e in questo studio cercheremo di analizzarla.

La lotteria delle sentenze

Partiamo dall'ipotesi di un mondo perfetto appena formulata, in cui tutti i casi ricevono la stessa pena da parte di ciascun giudice. Qui ogni colonna

conterrebbe 208 numeri identici. Ora aggiungiamo il rumore percorrendo dall'alto in basso la colonna e modificando qualche numero qua e là, quindi riducendo o allungando il periodo di detenzione rispetto alla condanna media. Non essendo tutte uguali, queste modifiche creeranno una variabilità all'interno della colonna. Questa variabilità è il rumore.

Il risultato fondamentale di questo studio sta nella grande quantità di rumore osservabile *nei giudizi sullo stesso caso*. La misura del rumore all'interno di ogni caso è la deviazione standard dalla pena detentiva a esso assegnata. La condanna media nei sedici casi analizzati era di 7 anni, mentre la deviazione standard da quella media era pari a 3,4 anni.[5]

Anche se il concetto di *deviazione standard* vi fosse noto, potrà comunque essere utile darne una descrizione concreta. Immaginate di prendere due giudici qualsiasi e calcolare la differenza tra i loro giudizi su un caso. Ora ripetete questa operazione per tutte le coppie di giudici e tutti i casi, e calcolate la media dei risultati. Questa misura, la *differenza assoluta media*, vi darà un'idea della lotteria in cui è coinvolto un imputato di un tribunale federale. Presumendo che i giudizi abbiano una distribuzione normale, sarà pari a 1,128 volte la deviazione standard, pertanto la differenza media tra due condanne qualsiasi per lo stesso caso sarà di 3,8 anni. Nel capitolo 2 abbiamo parlato della lotteria in cui è coinvolto il cliente che debba stipulare una particolare polizza assicurativa; la lotteria a cui è soggetto l'imputato di un processo penale ha conseguenze assai più importanti.

Una differenza assoluta media tra i giudici di 3,8 anni quando la condanna media è di 7 anni è un risultato inquietante e, a nostro parere, inaccettabile. Eppure vi è motivo di sospettare che nell'effettiva conduzione della giustizia il rumore sia ancora più alto. Innanzitutto i partecipanti al controllo del rumore avevano di fronte dei casi fittizi, insolitamente facili da confrontare e presentati in immediata successione. Nella vita reale non è

altrettanto semplice mantenere la massima coerenza. In secondo luogo, in un'aula di tribunale i giudici hanno molte più informazioni, e ogni nuova informazione, a meno che non sia decisiva, genera nuove occasioni per ciascun giudice di differire da ogni altro. Per queste ragioni, sospettiamo che il rumore a cui sono esposti gli imputati in tribunale sia perfino maggiore di quello sin qui osservato.

Alcuni giudici sono severi: rumore di livello

Nella fase successiva dell'analisi gli autori hanno suddiviso il rumore in componenti distinte. La prima riflessione che forse vi verrà in mente – e che venne in mente al giudice Frankel – è che il rumore sia dovuto a una variazione nella propensione dei giudici a emettere condanne severe. Come confermerà qualsiasi avvocato difensore, ogni giudice ha una sua reputazione: alcuni sono noti per essere "forcaioli", più severi del giudice medio, altri per essere "compassionevoli", più clementi del giudice medio. Chiameremo queste deviazioni *errori di livello*. (Ricordiamo che qui per errore intendiamo una deviazione dalla media; in realtà, un errore potrebbe anche sanare un'ingiustizia, se il giudice medio ha torto.)

La variabilità degli errori di livello è presente in ogni atto di giudizio. Pensiamo a quando nella valutazione di una prestazione alcuni supervisori sono più generosi di altri, nelle previsioni sulla quota di mercato alcuni esperti sono più ottimistici di altri, o nel prescrivere un'operazione alla schiena certi ortopedici sono più interventisti di altri.

Ogni riga della figura 9 mostra le condanne emesse da un unico giudice. La condanna media emessa da ogni giudice, indicata nella colonna all'estrema destra della tabella, dà una misura del suo livello di severità. È evidente che i giudici differiscono ampiamente su questo piano: la

deviazione standard dei valori della colonna all'estrema destra è di 2,4 anni. Questa variabilità non ha niente a che fare con la giustizia: al contrario, capite che le differenze nella condanna media riflettono la variazione tra i giudici rispetto ad altre caratteristiche, come l'estrazione sociale, le esperienze personali, le opinioni politiche, i pregiudizi e così via. I ricercatori hanno esaminato l'atteggiamento generale dei giudici verso le condanne, per esempio se pensano che l'obiettivo principale della condanna sia l'interdizione (l'allontanamento del criminale dalla società), la riabilitazione o la deterrenza, e hanno scoperto che quelli che si pongono come obiettivo principale la riabilitazione tendono ad assegnare pene detentive più brevi e periodi di affidamento ai servizi sociali più lunghi rispetto a quelli che puntano alla deterrenza o all'interdizione. Inoltre, i giudici attivi nel Sud degli Stati Uniti hanno emesso condanne significativamente più lunghe rispetto ai giudici di altre parti del paese. Si è riscontrato insomma, come era prevedibile, un rapporto tra ideologia conservatrice e severità delle condanne.

In generale si può concludere che il livello medio delle condanne sia paragonabile a un tratto di personalità. A partire da questo studio si potrebbero porre i giudici su una scala che va da molto duri a molto clementi, come un test di personalità potrebbe misurare il loro grado di estroversione o amicalità. Alla stregua di altri tratti, ci aspetteremmo che la severità delle condanne fosse correlata a fattori genetici, alle esperienze personali e ad altri aspetti della personalità: nessuno di questi tratti ha a che fare con il caso o con l'imputato. Adottiamo il termine *rumore di livello* per indicare la variabilità dei giudizi medi dei giudici, che coincide con la variabilità degli errori di livello.

I giudici sono diversi: il rumore strutturale

Come mostrano le frecce nere della figura 9, il rumore di livello è di 2,4 anni, quello sistemico di 3,4. Questa discrepanza indica che il rumore sistemico va oltre le differenze attestabili nella severità media dei singoli giudici. Chiamiamo questa ulteriore componente *rumore strutturale*.

Per comprendere il rumore strutturale torniamo alla figura 9 e concentriamoci su una cella a caso, per esempio la C3. La condanna media nel caso C è indicata in fondo alla colonna: come vedete, è di 3,7 anni. Guardiamo ora la colonna all'estrema destra per individuare la condanna media emessa dal giudice 3 nei vari casi: è di 5,0 anni, esattamente 2,0 anni in meno della media totale. Se la variazione nella severità dei giudici fosse l'unica fonte di rumore della colonna 3, potremmo ipotizzare che la condanna della cella C3 sia di 3,7 – 2,0 = 1,7 anni. Ma il dato effettivo della cella C3 è 4 anni, da cui si evince che il giudice 3 è stato particolarmente duro nella condanna emessa per quel caso.

Seguendo questa logica additiva dovremmo riuscire a prevedere ogni condanna di ogni colonna della tabella, ma in realtà nella maggior parte delle celle troveremmo delle deviazioni da questo semplice modello.[6] Osservando un'intera riga, scopriamo che i giudici non sono ugualmente severi in tutti i casi: in alcuni sono più duri della loro media personale, in altri più clementi. Chiamiamo queste deviazioni residue *errori strutturali.* Trascrivendo gli errori strutturali di ogni cella della tabella, scopriremmo che in totale ammontano a zero per ogni giudice (riga) e per ogni caso (colonna). Tuttavia, gli errori strutturali non si elidono a vicenda in termini di rumore, perché nel calcolo di quest'ultimo i valori di ogni cella vengono elevati al quadrato.

C'è un modo più semplice per dimostrare che il modello additivo delle condanne non regge. Nella tabella vedete che le condanne medie riportate in fondo a ogni colonna aumentano progressivamente da sinistra a destra,

mentre ciò non avviene all'interno delle righe. Il giudice 208, per esempio, emette una condanna molto più elevata per l'imputato del caso O che per quello del caso P. Se i singoli giudici classificassero i casi in base al periodo di detenzione che ritengono appropriato, le loro classificazioni sarebbero diverse.

Con il termine *rumore strutturale* indichiamo la variabilità appena identificata, perché essa riflette la struttura complessa dell'atteggiamento dei giudici nei casi particolari. Un giudice, per esempio, potrà essere più duro della media in generale, ma relativamente più clemente verso criminali dei ceti più elevati; un altro potrà essere incline a condanne leggere, ma più severo quando il delinquente è recidivo; un terzo potrà essere vicino alla severità media, ma sarà bendisposto quando il trasgressore è solo un complice e più duro quando la vittima è una persona anziana. (Adottiamo il termine *rumore strutturale* per una questione di praticità. Il termine statistico corretto per il rumore strutturale è *interazione giudice × caso*. Ci scusiamo con chi ha una formazione statistica per lo sforzo di traduzione richiesto.)

Nel contesto della giustizia penale, alcune reazioni idiosincratiche ai casi potrebbero essere dettate dalla personale "filosofia della pena" del giudice. Altre potrebbero essere dovute ad associazioni di cui il giudice non è neppure consapevole, come nel caso di un imputato che gli ricordi un criminale particolarmente efferato o magari di un'imputata che assomigli a sua figlia. A prescindere dalla motivazione, questi atteggiamenti strutturali non sono frutto del caso: ci aspettiamo che ricorrano ogni volta che al giudice si ripresenti un caso analogo. Ma poiché il rumore strutturale, nella pratica, è difficile da prevedere, aggiunge ulteriore incertezza alla già imprevedibile lotteria della determinazione della pena. Come osservato dagli autori dello studio, «le differenze strutturali tra i giudici

relativamente all'influenza delle caratteristiche del crimine o del criminale» costituiscono «un'ulteriore forma di disparità di condanna».[7]

Forse vi sarete accorti che la scomposizione del rumore sistemico in rumore di livello e rumore strutturale segue la stessa logica dell'equazione di errore illustrata nel capitolo precedente, in cui l'errore veniva scomposto in bias e rumore. Questa volta potremmo avere un'equazione del tipo:

$$\textbf{Rumore sistemico}^2 = \textbf{Rumore di livello}^2 + \textbf{Rumore strutturale}^2$$

Questa espressione può essere rappresentata graficamente proprio come l'equazione di errore originaria (figura 10). Qui l'abbiamo illustrata attraverso un triangolo con due cateti uguali: questo perché, nello studio delle condanne, il rumore strutturale e il rumore di livello contribuiscono all'incirca nella stessa misura al rumore sistemico.[8]

Il rumore strutturale è dilagante. Poniamo che dei medici debbano decidere se ricoverare o no qualcuno, che delle società debbano decidere chi assumere, che degli avvocati debbano decidere quali casi seguire o che dei produttori hollywoodiani debbano decidere quali programmi televisivi finanziare: in tutti questi casi ci sarà un rumore strutturale, in cui diversi decisori arriveranno a classificazioni diverse dei casi.

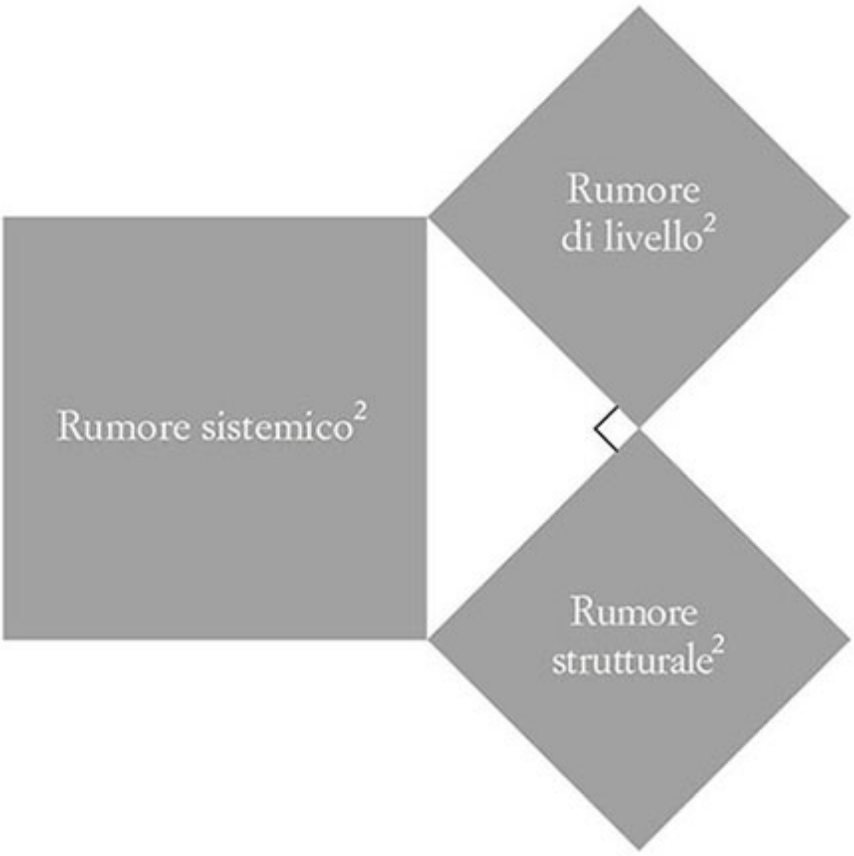


Figura 10. Scomporre il rumore sistemico

Le componenti del rumore

La nostra trattazione del rumore strutturale ha glissato su un'importante complicazione: il possibile contributo dell'errore casuale.

Ricordate l'esercizio con il cronometro? Quando avete cercato di misurare più volte un intervallo di dieci secondi, avete riscontrato che i vostri risultati variavano da un giro all'altro, in quella che si definisce una variabilità intrapersonale. Allo stesso modo, i giudici non avrebbero emesso le stesse esatte condanne per questi sedici casi se fosse stato chiesto loro di giudicarli *ex novo* in un'altra occasione. Anzi, come vedremo, non avrebbero emesso le stesse condanne neanche se lo studio originario fosse stato condotto in un giorno diverso della stessa settimana. Se un giudice è di

buonumore perché sua figlia ha ricevuto una buona notizia, perché la sua squadra del cuore ha vinto la sera prima o magari perché è una bella giornata, il suo giudizio potrebbe essere più indulgente. Questa variabilità intrapersonale è concettualmente diversa dalle più stabili differenze interpersonali appena discusse, ma è difficile distinguere tra queste due fonti di variabilità. Chiameremo la variabilità dovuta a effetti transitori *rumore occasionale*.

In effetti, in questo studio non abbiamo considerato il rumore occasionale: nel controllo del rumore abbiamo deciso di interpretare gli specifici modelli di condanna dei giudici come il risultato di atteggiamenti stabili. Questo è senz'altro un assunto ottimistico, ma vi sono motivi oggettivi per credere che il rumore occasionale non abbia rivestito un ruolo centrale in questo studio: sicuramente i giudici che vi hanno partecipato, nella loro lunga esperienza, saranno pervenuti ad alcune idee fisse sul significato di varie caratteristiche di crimini e imputati. Nel prossimo capitolo esploreremo nel dettaglio il rumore occasionale e mostreremo come questo possa essere separato dalla componente stabile del rumore strutturale.

Riassumendo, abbiamo analizzato diversi tipi di rumore. Il *rumore sistemico* è la variabilità indesiderata che interviene nei giudizi espressi sullo stesso caso da più individui. Abbiamo identificato le sue due componenti principali, che possono essere disgiunte quando gli stessi individui valutano più casi:

- Il *rumore di livello* è la variabilità nel livello medio dei giudizi espressi da giudici diversi.
- Il *rumore strutturale* è la variabilità nelle reazioni dei giudici a casi particolari.

Nel presente studio il peso del rumore di livello e del rumore strutturale è pressoché uguale. Tuttavia, la componente che abbiamo identificato come

rumore strutturale contiene indubbiamente un certo grado di *rumore occasionale*, che può essere considerato alla stregua di un errore casuale.

Qui abbiamo presentato a titolo illustrativo i risultati di un controllo del rumore nel sistema giudiziario, ma la stessa analisi potrebbe essere applicata a qualsiasi controllo del rumore, in campo aziendale, medico, governativo e quant'altro. Il rumore di livello e quello strutturale (che comprende a sua volta quello occasionale) contribuiscono entrambi al rumore sistemico. Li incontreremo più volte nel corso della nostra trattazione.

A proposito dell'analisi del rumore

«Il rumore di livello si ha quando i giudici mostrano diversi livelli di severità, mentre il rumore strutturale si verifica quando sono in disaccordo tra loro su quali imputati meritino un trattamento più o meno severo. Una parte del rumore strutturale è costituita dal rumore occasionale, che si manifesta quando i giudici sono in disaccordo con se stessi.»

«In un mondo perfetto, gli imputati verrebbero affidati alla giustizia; nel mondo reale, vengono affidati a un sistema rumoroso.»

[1] J. Bartolomeo *et al.*, *Sentence Decisionmaking: The Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity*, in "Journal of Criminal Law and Criminology", 72(1981), n. 2, pp. 524-554; INSLAW, Inc. *et al.*, *Federal Sentencing: Towards a More Explicit Policy of Criminal Sanctions III-4*, 1981.

[2] La condanna poteva prevedere qualsiasi combinazione di pene detentive, servizi sociali e sanzioni. Per semplicità, qui ci concentriamo soprattutto sulla componente principale delle condanne, la pena detentiva, tralasciando le altre due.

[3] In una configurazione simile, con una molteplicità di casi e di giudici, la versione estesa dell'equazione di errore presentata nel capitolo 5 comprende un termine che riflette questa varianza. Nello specifico, se definiamo un *bias generale* come l'errore medio di tutti i casi, e se questo errore non è identico da caso a caso, avremo una varianza dei bias dei casi. L'equazione diventa la seguente: Errore quadratico medio = Bias generale^2 + Varianza dei bias dei casi + Rumore sistemico^2.

[4] I numeri citati in questo capitolo derivano dallo studio originario come segue.

Innanzitutto gli autori riferiscono che l'effetto principale di *crimine e criminale* ammonti al 45% della varianza totale (vedi J. Bartolomeo *et al.*, *Sentence Decisionmaking*, cit. 2, tabella 6). Tuttavia, a noi qui interessa l'effetto di ciascun caso in generale, comprese tutte le caratteristiche presentate ai giudici, come gli eventuali precedenti penali dell'imputato o l'eventuale arma del delitto. Secondo la nostra definizione tutte queste caratteristiche rientrano nella *varianza dei casi reali*, non nel rumore. Pertanto abbiamo reintegrato nella varianza dei casi le interazioni tra i vari aspetti di ciascun caso (che coprono l'11% della varianza totale; vedi J. Bartolomeo *et al.*, cit., tabella 10). Di conseguenza, abbiamo riassegnato alla varianza dei casi un peso del 56%, all'effetto principale del giudice (rumore di livello) un peso del 21% e alle interazioni nella varianza totale un peso del 23%. Il rumore sistemico è quindi pari al 44% della varianza totale.

La varianza delle sentenze giuste può essere calcolata a partire da Bartolomeo *et al.*, cit., p. 89, seguendo la tabella che elenca le condanne medie per ciascun caso, e ammonta a 15. Se questo numero è pari al 56% della varianza totale, allora quest'ultima è pari a 26,79 e la varianza del rumore sistemico è pari a 11,79. La radice quadrata di quella varianza è la deviazione standard di un caso rappresentativo, pari a 3,4 anni.

L'effetto principale del giudice, o rumore di livello, è pari al 21% della varianza totale. La radice quadrata di tale varianza è la deviazione standard attribuibile al rumore di livello del giudice, pari a 2,4 anni.

[5] Questo valore è la radice quadrata della media delle varianze delle condanne sui sedici casi. L'abbiamo calcolata come spiegato nella nota precedente.

[6] L'ipotesi dell'additività parte essenzialmente dal presupposto che la severità di un giudice incrementi di una quantità costante il tempo di detenzione, ma difficilmente tale ipotesi sarà corretta: è più probabile che la severità del giudice aumenti di una quantità proporzionale alla condanna media. Questa questione è stata trascurata nell'analisi originaria, cosa che non consente in alcun modo di valutarne l'importanza.

[7] J. Bartolomeo *et al.*, *Sentence Decisionmaking*, cit., p. 23.

[8] La seguente equazione regge: $(\text{Rumore sistemico})^2 = (\text{Rumore di livello})^2 + (\text{Rumore strutturale})^2$. La tabella mostra che il rumore sistemico è di 3,4 anni, quello di livello di 2,4. Ne consegue che anche il rumore strutturale sarà di circa 2,4 anni. Il calcolo ha un intento illustrativo: i valori reali sono leggermente diversi a causa degli errori di arrotondamento.

7

Rumore occasionale

Un giocatore di basket professionista si prepara per un tiro libero. Si posiziona sulla linea, si concentra e tira. Ha compiuto questa sequenza di movimenti infinite volte. Farà canestro?

Non lo sappiamo, come non lo sa neanche lui. Nella National Basketball Association, anche nota come NBA, in media i giocatori realizzano un tiro libero tre volte su quattro. Alcuni giocatori, naturalmente, sono più bravi di altri, ma nessuno segna nel 100% dei casi. I migliori di tutti i tempi (che in data attuale sono Stephen "Steph" Curry, Steve Nash e Mark Price) superano di poco il 90% di canestri su tiro libero.[1] I peggiori di tutti i tempi arrivano circa al 50% (il grande Shaquille O'Neal, per esempio, mise a segno soltanto il 53% dei suoi tiri liberi).[2] Benché l'anello si trovi sempre alla stessa altezza e alla stessa distanza, e la palla abbia sempre lo stesso peso, non è facile acquisire la capacità di ripetere la precisa sequenza di gesti richiesta per segnare. La variabilità è la norma, non solo tra i giocatori ma anche per lo stesso giocatore: il tiro libero è un'altra forma di lotteria, con una possibilità di successo molto più elevata se a tirare è Curry piuttosto che O'Neal, ma pur sempre una lotteria.

Da cosa deriva questa variabilità? Sappiamo che infiniti fattori possono influenzare i giocatori sulla linea dei tiri liberi: la stanchezza di una lunga partita, la pressione mentale di una gara tirata, il calore della tifoseria di casa o i fischi dei tifosi della squadra avversaria. Se uno come Curry o come Nash sbaglia, ci appigliamo a una di queste giustificazioni, ma in realtà è

difficile arrivare a capire in che modo incidano esattamente tutti questi fattori. La variabilità della prestazione di un tiratore è una forma di rumore.

La seconda lotteria

La variabilità nei tiri liberi o in altri processi fisici non è certo una sorpresa. Siamo tutti abituati alla variabilità del nostro corpo: il battito cardiaco, la pressione sanguigna, i riflessi, il tono di voce e il tremito delle mani cambiano a seconda delle circostanze, e, anche se ci sforziamo di firmare sempre allo stesso modo, la nostra firma sarà ogni volta leggermente diversa.

Meno facile da osservare è la variabilità delle nostre menti. Certo, capita a tutti di cambiare idea, anche senza disporre di nuove informazioni. Il film che ci ha fatto sbellicare dalle risate ieri sera adesso ci sembra banale e irrilevante; la persona su cui ieri abbiamo espresso un giudizio severo ora sembra meritare la nostra indulgenza; un ragionamento che non ci piaceva o non capivamo, una volta recepito ci appare essenziale. Ma, come indicano questi esempi, di solito associamo tali cambiamenti a questioni relativamente minori e per lo più soggettive.

Di fatto, le nostre opinioni cambiano senza un motivo evidente, e lo stesso succede con questioni soggette al giudizio attento e ponderato di professionisti esperti. Capita spesso, per esempio, di ottenere diagnosi significativamente diverse dallo stesso medico se gli si presenta due volte lo stesso caso (vedi capitolo 22). Quando, in un grande concorso enologico americano, gli esperti hanno assaggiato due volte gli stessi vini, è capitato che abbiano dato un identico punteggio solo al 18% di questi (di solito ai peggiori).[3] Un criminologo può giungere a diverse conclusioni quando si

trova a esaminare le stesse impronte digitali per la seconda volta a distanza di poche settimane (vedi capitolo 20). Un esperto di informatica può fornire stime notevolmente diverse del tempo di completamento dello stesso intervento in due occasioni diverse.[4] Insomma, come un giocatore di basket non farà mai due tiri esattamente uguali, così non sempre noi formuleremo giudizi identici sugli stessi fatti in due occasioni diverse.

Abbiamo paragonato la selezione casuale di un sottoscrittore, un giudice o un medico a una lotteria che crea rumore sistemico. Il rumore occasionale deriva invece da un'altra lotteria, legata al momento in cui un professionista elabora un giudizio, al suo umore, alla serie di casi recenti che ha freschi nella memoria e a innumerevoli altri aspetti di una data situazione. Questa seconda lotteria di solito si situa su un piano molto più astratto della prima: capiamo, per esempio, come la prima avrebbe potuto estrarre un sottoscrittore diverso, ma le alternative possibili ai reali comportamenti del sottoscrittore selezionato restano astratte e controfattuali. Sappiamo solo che il giudizio espresso è stato estratto da un insieme di possibilità; il rumore occasionale rappresenta la variabilità tra queste possibilità invisibili.

Misurare il rumore occasionale

Misurare il rumore occasionale non è facile proprio perché la sua esistenza, una volta postulata, spesso ci sorprende. Quando una persona si forma un'opinione professionale dopo un'attenta riflessione, la associa alle motivazioni che giustificano il proprio punto di vista: se le si chiede di motivare il proprio giudizio, spesso lo difenderà con argomenti che le appaiono convincenti, e se si presenta lo stesso problema alla stessa persona per la seconda volta e questa lo riconosce, riproporrà la risposta

iniziale per ridurre al minimo lo sforzo e mantenere la massima coerenza. Facciamo un esempio: se un insegnante dà un voto eccellente all'elaborato di uno studente e poi rilegge quello stesso saggio una settimana dopo, conoscendo il voto originario è improbabile che ne dia uno molto diverso.

Pertanto, è difficile ottenere una misurazione diretta del rumore occasionale in presenza di casi facili da ricordare. Se, per esempio, mostrassimo a un sottoscrittore o a un giudice penale un caso su cui hanno già preso una decisione, probabilmente lo riconosceranno e riproporranno il giudizio precedente. In un'analisi delle ricerche sulla variabilità nei giudizi professionali (definita tecnicamente attendibilità test-retest, o anche solo attendibilità) si è riscontrato che in molti studi gli esperti esprimevano due volte lo stesso giudizio all'interno di una stessa sessione. Come era prevedibile, tendevano a essere d'accordo con se stessi.[5]

Gli esperimenti citati in precedenza hanno aggirato il problema impiegando stimoli che gli esperti non potevano riconoscere. Gli enologi sono stati sottoposti a una degustazione alla cieca (*blind testing*), ai criminologi sono state mostrate coppie di impronte già viste, mentre agli esperti informatici è stato chiesto di esprimersi su compiti affrontati settimane o mesi prima, senza dirgli che si trattava di casi che avevano già esaminato.

Esiste un altro metodo, meno diretto, per comprovare l'esistenza del rumore occasionale: l'impiego dei cosiddetti "big data" e di metodi econometrici. Potendo disporre di un ampio campione storico di decisioni professionali, in alcuni casi gli analisti sono in grado di controllare se tali decisioni siano state influenzate da fattori irrilevanti legati a una particolare circostanza come l'ora del giorno o la temperatura esterna. La presenza di effetti statisticamente significativi di simili fattori irrilevanti sui giudizi espressi costituisce una prova del rumore occasionale. A essere

realisti, non c'è speranza di scoprire tutte le fonti non pertinenti di rumore occasionale, ma quelle a cui si riesce a risalire ne attestano la grande varietà. Se vogliamo controllare il rumore occasionale, dobbiamo cercare di comprendere i meccanismi che ne sono all'origine.

La folla interna

Partiamo da una domanda concreta: in percentuale, quanti degli aeroporti presenti nel mondo si trovano negli Stati Uniti? Se ci riflettete, probabilmente vi verrà in mente una risposta, ma non sarà come rispondere a una domanda sulla vostra età attuale o sul vostro numero di telefono. Siete consapevoli che il numero a cui siete arrivati è una stima. Non è però un numero a caso: è chiaro che sia 1% sia 99% sarebbero risposte sbagliate. Il numero a cui siete arrivati si situa all'interno di una gamma di possibilità che non scartereste. Se qualcuno aggiungesse o sottraesse un punto percentuale alla vostra risposta, probabilmente non trovereste tale stima meno plausibile della vostra. (La risposta esatta, se ve lo state chiedendo, è 32%.)[6]

I ricercatori Edward Vul e Harold Pashler hanno pensato di porre questa stessa domanda (e molte altre simili) ad alcune persone, non una ma due volte.[7] La prima volta non veniva detto ai soggetti che avrebbero dovuto formulare una seconda stima. L'ipotesi era che la media delle due risposte sarebbe stata più accurata di ognuna di loro presa singolarmente.

I dati confermarono la loro intuizione: in generale, la prima risposta era più vicina alla verità della seconda, ma la media dei due tentativi dava una stima migliore.[8]

Vul e Pashler trassero ispirazione dal fenomeno noto come "effetto della saggezza della folla": in genere la media dei giudizi indipendenti di persone

diverse porta a un risultato più accurato. Nel 1907 Francis Galton, un cugino di Darwin famoso per la vastità dei suoi interessi scientifici, chiese a 787 avventori di una fiera di paese di ipotizzare quanto pesasse un bue, che sarebbe andato in premio a chi avesse indovinato. Nessuno dei presenti diede la risposta esatta, pari a 1198 libbre, ma la media delle risposte fu 1200, con uno scarto di sole due libbre, e anche la mediana (1207) era molto vicina. Quei paesani erano una "folla saggia", nel senso che le loro stime individuali erano affette da un certo grado di rumore, ma non da bias. La dimostrazione fu una sorpresa per lo stesso Galton, che nutriva uno scarso rispetto per il giudizio della gente comune e fu costretto a constatare a malincuore che quel risultato dava «più credito all'attendibilità di un giudizio democratico di quanto ci si potesse aspettare».

Si è arrivati a risultati simili in centinaia di situazioni diverse. Naturalmente, se le domande sono così difficili che solo gli esperti possono avvicinarsi alle risposte, non è detto che la folla sia altrettanto accurata, ma quando, per esempio, si chiede a un gruppo di persone di indovinare quante caramelle sono contenute in un barattolo trasparente, di prevedere che temperatura ci sarà nella propria città da qui a una settimana o di stimare la distanza tra due città di un certo stato, probabilmente la risposta media di un gran numero di individui si avvicinerà alla verità.

La ragione è puramente statistica: la media di diversi giudizi (o misure) indipendenti dà un nuovo giudizio che è meno rumoroso,[9] anche se non meno affetto da bias, dei singoli giudizi.

Vul e Pashler si chiesero se tale effetto si riscontrasse anche nel caso del rumore occasionale: ci si può avvicinare alla verità associando due risposte date dalla stessa persona, come per le risposte date da persone diverse? Ebbene, scoprirono che è possibile, e diedero a questo fenomeno un nome evocativo: *la folla interna*.

La media di due risposte date dalla stessa persona non porta a un miglioramento del giudizio finale pari a quello che si otterrebbe chiedendo una seconda opinione indipendente: per dirla con le parole degli autori dello studio, «ponendosi due volte la stessa domanda si guadagna circa un decimo di quanto si guadagnerebbe chiedendo una seconda opinione a qualcun altro».[10] Non è un grosso miglioramento. Ma l'effetto sarà molto più grande se si aspetta ad avanzare la seconda ipotesi: quando Vul e Pashler lasciarono passare tre settimane prima di porre la stessa domanda ai loro soggetti, ottennero un guadagno pari a un terzo di quello ottenibile chiedendo un secondo parere indipendente, il che non è poco, per una tecnica che non richiede informazioni aggiuntive o un aiuto esterno. Questo risultato fornisce senz'altro un fondamento logico alla classica raccomandazione che si dà a chi deve prendere una decisione: «Dormici su, la notte porta consiglio».

In uno studio indipendente contemporaneo a quello di Vul e Pashler, i due ricercatori tedeschi Stefan Herzog e Ralph Hertwig arrivarono a un'applicazione diversa dello stesso principio.[11] Invece di limitarsi a chiedere ai soggetti di avanzare una seconda stima, li incoraggiarono a fornirne una che, seppur plausibile, fosse la più diversa possibile dalla prima che avevano formulato. Questa richiesta implicava che i soggetti riconsiderassero attivamente le informazioni tralasciate la prima volta. I partecipanti ricevettero le seguenti istruzioni:

> Innanzitutto supponete che la vostra prima stima sia sbagliata. Poi pensate a quali potrebbero essere le motivazioni: quali presupposti, quali considerazioni saranno stati errati? Quindi chiedetevi cosa comportano queste nuove considerazioni: la prima stima era troppo alta o troppo bassa? Infine, partendo da questa nuova prospettiva, formulate una seconda stima alternativa.

Come Vul e Pashler, Herzog e Hertwig calcolarono la media delle due stime. La loro tecnica, definita *bootstrapping dialettico*, portò a un miglioramento dell'accuratezza superiore rispetto a una semplice richiesta di una seconda stima immediatamente successiva alla prima. Dal momento che i partecipanti erano costretti a riconsiderare il problema sotto una nuova luce, sperimentavano una versione ancora più diversa di loro stessi, come due "membri" della "folla interna" con posizioni più distanti. Di conseguenza, la media produceva una stima della risposta corretta più accurata. In termini di accuratezza, il guadagno delle due stime "dialettiche" immediatamente successive era pari alla metà del valore di una seconda opinione.

In conclusione, per Herzog e Hertwig si tratta di una semplice scelta tra due procedure: se si può ottenere un parere indipendente di altre persone, è bene chiederlo, perché è molto probabile che questa "saggezza della folla" si traduca in un giudizio migliore. Se non si può, ha senso esprimere due volte lo stesso giudizio da soli in modo da creare una "folla interna". Si può lasciar passare del tempo per stabilire una distanza con la propria opinione iniziale oppure si può cercare di avanzare delle controargomentazioni a se stessi per vedere il problema da una prospettiva diversa. Infine, indipendentemente dal tipo di folla, a meno che non abbiate fortissime motivazioni per dare un peso maggiore a una delle vostre stime, avrete un risultato migliore dalla media di entrambe.

Al di là delle considerazioni pratiche, questa linea di ricerca conferma un'intuizione essenziale sul giudizio: per citare Vul e Pashler, «più che essere deterministicamente selezionate sulla base di tutte le conoscenze possedute da un soggetto, le risposte date da tale soggetto costituiscono un campione di una distribuzione di probabilità interna».[12] Questa osservazione riflette l'esperienza fatta da voi stessi nel rispondere alla

domanda sugli aeroporti statunitensi: la vostra prima risposta non ha coinvolto tutte le vostre conoscenze, né le migliori a cui avreste potuto attingere, ma era solo una delle tante possibili risposte che avrebbe potuto generare la vostra mente. La variabilità osservabile nei giudizi dati da una stessa persona sullo stesso problema non è un caso osservato soltanto in problemi altamente specialistici: il rumore occasionale è insito in tutti i giudizi che prendiamo, in ogni momento.

Fonti di rumore occasionale

C'è almeno una fonte di rumore occasionale di cui tutti siamo consapevoli: l'umore. Avremo senz'altro osservato che i nostri giudizi possono dipendere da come ci sentiamo, e ci saremo accorti che anche quelli altrui variano a seconda del loro umore.

L'effetto dell'umore sul giudizio è oggetto di moltissime ricerche nel campo della psicologia. È molto facile rendere qualcuno temporaneamente felice o triste e misurare la variabilità dei suoi giudizi e delle sue decisioni una volta indotto un certo umore. I ricercatori impiegano varie tecniche a questo scopo: qualche volta, per esempio, viene chiesto ai partecipanti di scrivere un breve testo che rievochi un ricordo felice o triste, altre volte si chiede invece ai soggetti di guardare un filmato divertente o strappalacrime.

Vari psicologi studiano da decenni gli effetti della manipolazione dell'umore. Forse il più attivo è l'australiano Joseph Forgas, che ha pubblicato circa cento articoli scientifici sul tema.[13]

Alcune delle sue ricerche confermano ciò che tutti potevamo immaginare: in genere le persone sono più positive quando sono di buonumore. Trovano più facile richiamare ricordi felici che tristi, sono più

accondiscendenti, più generose e così via. L'umore negativo ha l'effetto opposto: come rileva Forgas, «lo stesso sorriso percepito come amichevole da una persona di buonumore può risultare impacciato quando l'osservatore è di cattivo umore; parlare di che tempo fa può essere considerato educato quando una persona è felice, ma noioso quando la stessa persona è abbattuta».[14]

In altri termini, l'umore ha un'influenza misurabile su cosa pensiamo: su ciò che notiamo in un certo ambiente, su ciò che ripeschiamo dalla memoria, sul significato che diamo a questi segnali. Ma ha anche un effetto ancora più sorprendente: cambia perfino il modo in cui pensiamo. E in questo caso gli effetti non sono quelli che potreste immaginare. Il buonumore ha degli svantaggi, il malumore dei risvolti positivi; i costi e i benefici di diversi umori sono legati alla situazione.

In un negoziato, per esempio, il buonumore aiuta. Le persone di buonumore sono più collaborative e gli altri lo sono con loro, dunque tendono a ottenere risultati migliori rispetto ai negoziatori tristi. Naturalmente i negoziati che vanno a buon fine sono essi stessi una fonte di gioia, ma in questi esperimenti l'umore non deriva dall'esito della trattativa: viene indotto prima. Inoltre, i negoziatori che passano dal buonumore alla rabbia nel corso del negoziato spesso ottengono buoni risultati – un dettaglio che vale la pena ricordare se si deve trattare con persone ostinate.[15]

D'altro canto, il buonumore ci rende più inclini a dare credito alla nostra prima impressione, senza metterla in discussione. In uno studio condotto da Forgas, ai partecipanti venne dato da leggere un breve saggio filosofico corredato da una foto dell'autore.[16] Alcuni lettori si trovarono davanti lo stereotipo del professore di filosofia, un uomo di mezza età con gli occhiali, mentre altri si ritrovarono la foto di una giovane donna. Come potrete

immaginare, questo test vuole dimostrare fino a che punto i lettori siano inclini agli stereotipi: daranno una valutazione più favorevole quando il saggio sarà attribuibile a un uomo di mezza età piuttosto che a una giovane donna? Come potrete immaginare, la risposta è sì. Ma attenzione: la differenza è maggiore quando i lettori sono di buonumore, e questo ci dice che chi è felice si lascia influenzare più facilmente dai propri pregiudizi.

Altri studi hanno esaminato l'effetto dell'umore sulla credulità. Gordon Pennycook e i suoi colleghi hanno condotto molti studi sulle reazioni della gente ad affermazioni insensate pseudoprofonde generate associando in maniera casuale nomi e verbi estrapolati da citazioni di guru molto noti e inserendoli in frasi grammaticalmente corrette, come: «Il tutto placa infiniti fenomeni», oppure «Il significato nascosto trasforma l'incomparabile bellezza astratta».[17] La propensione a concordare con queste affermazioni è nota come *bullshit receptivity* ("ricettività alle stronzate"). Per inciso, *bullshit* è diventato una sorta di termine tecnico dopo che Harry Frankfurt, un filosofo della Princeton University, ha pubblicato un libro sagace dal titolo *On Bullshit* (*Stronzate. Un saggio filosofico*, Rizzoli, Milano 2005), in cui operava una distinzione tra questo e altri tipi di travisamento.

Inevitabilmente certe persone sono più ricettive di altre alle "stronzate". Possono rimanere colpite da «asserzioni apparentemente convincenti presentate come vere e significative, quando in realtà sono prive di senso».[18] Ma, di nuovo, questa credulità non dipende solo da disposizioni permanenti e immutabili: una volta messe di buonumore, le persone sono più ricettive alle insulsaggini e in generale più credulone, meno inclini a individuare un raggiro o a identificare informazioni ingannevoli.[19] Per contro, un testimone oculare esposto a un'informazione fuorviante è più

incline a non prenderla sul serio – e a non testimoniare il falso – se è di cattivo umore.[20]

Perfino i giudizi morali sono fortemente influenzati dall'umore. In uno studio, alcuni ricercatori hanno sottoposto i soggetti coinvolti a una variante del dilemma del carrello, un classico della filosofia morale.[21] In questo esperimento mentale, cinque persone stanno per essere investite da un carrello ferroviario in corsa. I soggetti devono immaginarsi su un ponte sotto il quale passerà il carrello, e devono decidere se fermare la sua avanzata spingendo giù dal ponte un uomo molto grasso, in modo che il suo corpo funga da ostacolo. In questo modo, viene detto loro, l'uomo morirà, ma le altre cinque persone si salveranno.

Il dilemma del carrello illustra il conflitto tra diversi approcci al ragionamento morale. Il calcolo utilitario, associato al filosofo inglese Jeremy Bentham, suggerirebbe che sia preferibile perdere una vita che non cinque, mentre l'etica deontologica, riconducibile a Kant, proibisce di uccidere qualsiasi persona, anche con la finalità di salvarne tante altre. È chiaro che il dilemma del carrello contiene un forte elemento di pathos personale: spingere fisicamente un uomo giù da un ponte perché intralci il percorso di un carrello in corsa è un gesto ripugnante come pochi. Si tratta di una decisione utilitaria che richiede che una persona superi la propria avversione per un atto di violenza fisica nei confronti di uno sconosciuto. Solo una piccola parte dei soggetti (in questo studio, meno di uno su dieci) di solito opta per questa soluzione.

Tuttavia, quando i soggetti sono stati indotti al buonumore con la visione di un video di cinque minuti, la probabilità che decidessero di spingere l'uomo giù dal ponte è triplicata. Il fatto di considerare il "non uccidere" un principio assoluto o, al contrario, di essere disposti a far fuori un estraneo per salvarne cinque dovrebbe riflettere i nostri valori più

profondi; eppure, questa scelta sembra dipendere dal video che abbiamo appena guardato.

Se abbiamo descritto dettagliatamente questi studi sull'umore, è per sottolineare un'importante verità: *non siamo sempre la stessa persona*. Al variare del nostro umore (una cosa di cui siamo naturalmente consapevoli) variano anche alcuni tratti del nostro processo cognitivo (una cosa di cui *non* siamo pienamente consapevoli). Di fronte a un complesso problema di giudizio, l'umore del momento potrebbe influenzare il nostro approccio al problema e la conclusione a cui giungiamo, anche quando crediamo che il nostro umore non sia tanto influente o quando siamo in grado di giustificare con cognizione di causa la nostra risposta. Insomma, siamo soggetti al rumore.

Sono molti i fattori incidentali che introducono un rumore occasionale nel giudizio. Tra quelli non pertinenti, che quindi non dovrebbero influenzare i giudizi professionali (ma lo fanno), i maggiori indiziati sono due: lo stress e la stanchezza. Uno studio condotto su circa settecentomila visite di medici di base, per esempio, ha dimostrato che questi ultimi sono molto più propensi a prescrivere oppioidi alla fine di una lunga giornata di lavoro.[22] Naturalmente non è detto che un paziente che ha un appuntamento alle sedici soffra di più di uno che si presenta alle nove del mattino, né il fatto che il medico sia in ritardo sulla tabella di marcia dovrebbe influenzare le sue decisioni. Peraltro, le prescrizioni di altri analgesici come gli antinfiammatori non steroidei e il ricorso alla fisioterapia non seguono questa tendenza. Quando i medici hanno fretta, sembrano più inclini a optare per rimedi provvisori, senza tenere conto dei loro gravi effetti collaterali. Altri studi hanno evidenziato che, a fine giornata, i medici sono più inclini a prescrivere antibiotici[23] e meno inclini a prescrivere vaccini antinfluenzali.[24]

Anche il tempo atmosferico ha un'influenza misurabile sui giudizi professionali. Poiché spesso tali giudizi sono formulati in stanze climatizzate, l'effetto del tempo è probabilmente "mediato" dall'umore (cioè il meteo non agisce direttamente sulle decisioni, ma modifica l'umore del decisore, il che a sua volta ha un impatto sulle sue decisioni). Il maltempo è associato a una memoria migliore; le condanne sono tendenzialmente più severe quando fuori fa caldo; le prestazioni in Borsa sono influenzate dal sole.[25] In certi casi l'effetto del meteo è meno evidente: Uri Simonsohn ha rilevato che i responsabili delle ammissioni scolastiche prestano più attenzione alle credenziali accademiche dei candidati nei giorni nuvolosi e sono più sensibili a quelle non accademiche nelle giornate di sole. Il titolo dell'articolo che contiene questi risultati è di per sé memorabile: *Clouds Make Nerds Look Good* ("Le nuvole fanno bene ai nerd").[26]

Un'altra fonte di variabilità casuale nel giudizio è l'ordine in cui vengono esaminati i casi. Quando qualcuno considera un caso, le decisioni prese immediatamente prima fungono da riferimento implicito. I professionisti che prendono una serie di decisioni una dopo l'altra, compresi i giudici, i responsabili dei prestiti e gli arbitri di baseball sono portati a ristabilire una forma di equilibrio: dopo una sfilza di punti o una serie di decisioni che vanno nella stessa direzione, sono più propensi del dovuto a prendere una decisione che va nella direzione opposta. Di conseguenza gli errori (e le parzialità) sono inevitabili. La probabilità che i giudici americani accettino una richiesta di asilo, per esempio, si riduce del 19% se i due casi precedenti sono stati approvati. Una persona potrebbe ottenere un prestito se è stato negato ai due richiedenti che l'hanno preceduta, come potrebbe non ottenerlo se le due richieste precedenti sono state accolte. Questo comportamento riflette un bias cognitivo noto come *fallacia dello*

scommettitore: tendiamo a sottostimare la probabilità che si verifichi per caso una serie di eventi sfortunati o fortunati.[27]

Classificare il rumore occasionale

Quale peso ha il rumore occasionale sul rumore sistemico totale? Anche se non c'è una risposta valida per ogni situazione, è possibile individuare una regola generale. In termini di dimensioni, gli effetti che abbiamo descritto in questo capitolo sono inferiori rispetto alle differenze stabili tra diversi individui nel livello e nella struttura di giudizio.

Come si è visto, per esempio, la possibilità che un richiedente asilo venga ammesso negli Stati Uniti si riduce del 19% se lo stesso giudice ha appena concesso due ammissioni. Questa variabilità è senz'altro problematica, ma appare una minuzia se si confronta con la variabilità tra diversi giudici: Jaya Ramji-Nogales e i suoi colleghi hanno riscontrato che nel tribunale di Miami un giudice concedeva l'asilo all'88% dei richiedenti e un altro ad appena il 5%.[28] (Questo è un dato reale, non un controllo del rumore, quindi i richiedenti erano diversi, ma assegnati in maniera quasi casuale, e gli autori hanno riscontrato che le differenze dovute al paese di provenienza non giustificavano tali discrepanze.) Considerate queste disparità, ridurre i suddetti numeri del 19% non cambierebbe più di tanto la situazione.

Analogamente, talvolta gli esperti forensi e i medici sono in disaccordo perfino con se stessi, ma sempre meno di quanto non lo siano con gli altri. In ciascuno dei casi da noi esaminati in cui era possibile misurare la percentuale di incidenza del rumore occasionale sul totale, si è visto che questo contava molto meno delle differenze tra i singoli.

Se ne deduce che non solo non siamo sempre la stessa persona, ma siamo meno coerenti nel tempo di quanto non pensiamo. L'aspetto

rassicurante, però, è che siamo più simili ai noi stessi di ieri che a un'altra persona oggi.

Rumore occasionale, cause interne

L'umore, la stanchezza, il clima, gli effetti che si verificano a catena: molti fattori possono produrre variazioni indesiderate in un giudizio espresso sullo stesso caso dalla stessa persona. La speranza è di riuscire ad arrivare a uno scenario in cui tutti i fattori non pertinenti che pesano su una decisione siano noti e controllati. In teoria un tale scenario dovrebbe ridurre il rumore occasionale, ma è probabile che neanche questo sia sufficiente per eliminarlo del tutto.

Michael Kahana e i suoi colleghi della University of Pennsylvania si sono specializzati nello studio delle prestazioni mnemoniche.[29] (Stando alla nostra definizione, la memoria non è un'attività di giudizio, ma un'attività cognitiva in cui le condizioni possono essere rigorosamente controllate e le variazioni nelle prestazioni facilmente misurate.) In uno studio, hanno coinvolto settantanove soggetti in un'analisi straordinariamente approfondita delle loro capacità mnemoniche: i partecipanti sono stati sottoposti a ventitré sessioni distribuite in giorni diversi, in ognuna delle quali dovevano ricordare delle parole provenienti da ventiquattro diversi elenchi di ventiquattro parole ciascuno. La percentuale delle parole ricordate definiva la loro prestazione mnemonica.

Kahana e i suoi colleghi non erano interessati alle differenze tra i soggetti, ma ai predittori di variabilità nella prestazione di ciascun soggetto. La loro performance sarebbe stata condizionata da quanto si sentivano concentrati? Da quanto avevano dormito la notte prima? Dal momento della giornata? Sarebbe migliorata con la pratica, da una sessione

all'altra, o sarebbe peggiorata all'interno della stessa sessione al subentrare della stanchezza o della noia? Certi elenchi di parole si sarebbero dimostrati più facilmente memorizzabili di altri?

La risposta a tutte queste domande fu: "Sì, ma non di molto". Un modello che teneva conto di tutti questi predittori riusciva a spiegare solo l'11% della variazione nella prestazione di ciascun soggetto. I ricercatori si dissero «colpiti dal livello di variabilità che restava una volta eliminati gli effetti delle variabili predittive». Anche in un contesto così controllato, era un mistero quali fattori introducessero il rumore occasionale.

Tra tutte le variabili studiate dai ricercatori il predittore più forte della prestazione di un soggetto rispetto a un dato elenco non era un fattore esterno: la performance rispetto a un elenco di parole era prevedibile innanzitutto a partire da quella dello stesso soggetto rispetto all'elenco immediatamente precedente. Chi ricordava bene un elenco, ricordava bene anche il successivo; chi lo ricordava male, non ricordava neanche il successivo. La prestazione non variava in maniera casuale da un elenco all'altro: all'interno di ogni sessione aveva i suoi alti e bassi senza alcuna motivazione esterna evidente.

Questi dati indicano che le prestazioni mnemoniche sono condizionate, in gran parte, da quella che Kahana e i suoi colleghi definirono «l'efficienza di processi neurali endogeni che governano la funzione della memoria». Detto in altri termini, la variabilità dell'efficienza del cervello da un momento all'altro non è condizionata solo da fattori esterni come il meteo o una distrazione inattesa, ma è una caratteristica propria del modo in cui funziona la nostra mente.

È molto probabile che questa variabilità intrinseca del funzionamento del cervello incida anche sulla qualità dei nostri giudizi in modi che non riusciremo mai a controllare, e dovrebbe far riflettere chi pensa che il

rumore occasionale possa essere eliminato. L'analogia con il giocatore di basket sulla linea di tiro libero non era semplicistica come poteva sembrare: come i muscoli del giocatore non eseguono mai esattamente lo stesso movimento, così i nostri neuroni non operano mai esattamente nello stesso modo. La nostra mente è uno strumento di misurazione che non sarà mai perfetto.

Possiamo però cercare di controllare le pressioni indebite che si possono tenere a bada. Questo è particolarmente importante nei giudizi di gruppo, come vedremo nel capitolo 8.

A proposito del rumore occasionale

«Il giudizio è come un tiro libero: per quanto ci sforziamo di ripeterlo alla perfezione, non sarà mai esattamente identico.»

«Il giudizio dipende dal nostro umore, dai casi appena esaminati e perfino dal tempo atmosferico. Non siamo sempre la stessa persona.»

«Anche se forse non siamo la stessa persona di una settimana fa, siamo meno diversi da chi eravamo la settimana scorsa che da chiunque altro in questo stesso momento. Il rumore occasionale non è la maggiore fonte di rumore sistemico.»

[1] Vedi [www.iweblists.com/sports/basketball/FreeThrowPercent_c.html], consultato il 2 maggio 2021.

[2] Vedi [www.basketball-reference.com/players/o/onealsh01.html], consultato il 2 maggio 2021.

[3] R.T. Hodgson, *An Examination of Judge Reliability at a Major U.S. Wine Competition*, in "Journal of Wine Economics", 3(2008), n. 2, pp. 105-113.

[4] S. Grimstad, M. Jørgensen, *Inconsistency of Expert Judgment-Based Estimates of Software Development Effort*, in "Journal of Systems and Software", 80(2007), n. 11, pp. 1770-1777.

[5] R.H. Ashton, *A Review and Analysis of Research on the Test-Retest Reliability of Professional Judgment*, in "Journal of Behavioral Decision Making", 294(2000), n. 3, pp. 277-294. Per inciso, l'autore faceva notare che nessuno dei quarantuno studi da lui esaminati si proponeva di valutare il rumore occasionale: «In tutti i casi la misurazione dell'attendibilità era un prodotto secondario di qualche altro obiettivo di ricerca» (p. 279). Questo commento indica che l'interesse per lo studio del rumore occasionale è relativamente recente.

[6] Central Intelligence Agency, *The World Factbook*, Central Intelligence Agency, Washington, DC 2020. La cifra citata comprende tutti gli aeroporti e gli aerodromi riconoscibili dall'alto. La pista o le piste possono essere asfaltate o no, e comprendere impianti chiusi o abbandonati.

[7] E. Vul, H. Pashler, *Crowd Within: Probabilistic Representations Within Individuals*, in "Psychological Science", 19(2008), n. 7, pp. 645-647.

[8] Vedi anche J. Surowiecki, *La saggezza della folla*, Fusi orari, Roma 2007.

[9] La deviazione standard dei giudizi medi (la nostra misura del rumore) decresce in maniera proporzionale alla radice quadrata del numero di giudizi.

[10] E. Vul, H. Pashler, *Crowd Within*, cit., p. 646.

[11] S.M. Herzog, R. Hertwig, *Think Twice and Then: Combining or Choosing in Dialectical Bootstrapping?*, in "Journal of Experimental Psychology: Learning, Memory, and Cognition", 40(2014), n. 1, pp. 218-232.

[12] E. Vul, H. Pashler, *Crowd Within*, cit., p. 647.

[13] J.P. Forgas, *Affective Influences on Interpersonal Behavior*, in "Psychological Inquiry", 13(2002), n. 1, pp. 1-28.

[14] Ivi, p. 10.

[15] A. Filipowicz, S. Barsade, S. Melwani, *Understanding Emotional Transitions: The Interpersonal Consequences of Changing Emotions in Negotiations*, in "Journal of Personality and Social Psychology",

101(2011), n. 3, pp. 541-556.

[16] J.P. Forgas, *She Just Doesn't Look like a Philosopher...? Affective Influences on the Halo Effect in Impression Formation*, in "European Journal of Social Psychology", 41(2011), n. 7, pp. 812-817.

[17] G. Pennycook *et al.*, *On the Reception and Detection of Pseudo-Profound Bullshit*, in "Judgment and Decision Making", 10(2015), n. 6, pp. 549-563.

[18] Ivi, p. 549.

[19] J.P. Forgas, *Happy Believers and Sad Skeptics? Affective Influences on Gullibility*, in "Current Directions in Psychological Science", 28(2019), n. 3, pp. 306-313.

[20] Id., *Mood Effects on Eyewitness Memory: Affective Influences on Susceptibility to Misinformation*, in "Journal of Experimental Social Psychology", 41(2005), n. 6, pp. 574-588.

[21] P. Valdesolo, D. Desteno, *Manipulations of Emotional Context Shape Moral Judgment*, in "Psychological Science", 17(2006), n. 6, pp. 476-477.

[22] H.T. Neprash, M.L. Barnett, *Association of Primary Care Clinic Appointment Time with Opioid Prescribing*, in "JAMA Network Open", 2(2019), n. 8; L.M. Philpot *et al.*, *Time of Day Is Associated with Opioid Prescribing for Low Back Pain in Primary Care*, in "Journal of General Internal Medicine", 33(2018), p. 1828.

[23] J.A. Linder *et al.*, *Time of Day and the Decision to Prescribe Antibiotics*, in "JAMA Internal Medicine", 174(2014), n. 12, pp. 2029-2031.

[24] R.H. Kim *et al.*, *Variations in Influenza Vaccination by Clinic Appointment Time and an Active Choice Intervention in the Electronic Health Record to Increase Influenza Vaccination*, in "JAMA Network Open", 1(2018), n. 5, pp. 1-10.

[25] Sull'impatto sulla memoria, vedi J.P. Forgas, L. Goldenberg, C. Unkelbach, *Can Bad Weather Improve Your Memory? An Unobtrusive Field Study of Natural Mood Effects on Real-Life Memory*, in "Journal of Experimental Social Psychology", 45(2008), n. 1, pp. 254-257. Sulla luce del sole, vedi D. Hirshleifer, T. Shumway, *Good Day Sunshine: Stock Returns and the Weather*, in "Journal of Finance", 58(2003), n. 3, pp. 1009-1032.

[26] U. Simonsohn, *Clouds Make Nerds Look Good: Field Evidence of the Impact of Incidental Factors on Decision Making*, in "Journal of Behavioral Decision Making", 20(2007), n. 2, pp. 143-152.

[27] D. Chen *et al.*, *Decision Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires*, in "Quarterly Journal of Economics", 131(2016), n. 3, pp. 1181-1242.

[28] A.I. Schoenholtz, J. Ramji-Nogales, P.G. Schrag, *Refugee Roulette: Disparities in Asylum Adjudication*, cit.

[29] M.J. Kahana *et al.*, *The Variability Puzzle in Human Memory*, in "Journal of Experimental Psychology: Learning, Memory, and Cognition", 44(2018), n. 12, pp. 1857-1863.

8

Come i gruppi amplificano il rumore

Se il rumore nel giudizio individuale è già un bel problema, prendere decisioni in gruppo aggiunge ulteriori complicazioni. I gruppi possono andare in qualsiasi direzione, anche a causa di fattori che dovrebbero essere irrilevanti: chi parla per primo, chi parla per ultimo, chi parla con maggiore convinzione, chi è vestito di nero, chi è seduto accanto a una certa persona, chi sorride, aggrotta la fronte o gesticola al momento giusto – tutti questi fattori e tanti altri influenzano l'esito della decisione. Ogni giorno gruppi tra loro simili prendono decisioni molto diverse, che si tratti di assunzioni o promozioni, della chiusura di un ufficio o di una strategia di comunicazione, di norme ambientali, della sicurezza nazionale, delle ammissioni universitarie o del lancio di un nuovo prodotto.

Questa osservazione potrebbe sembrare strana, dal momento che, nel capitolo precedente, abbiamo detto che aggregare i giudizi di singoli individui riduce il rumore. Ma, a causa delle dinamiche di gruppo, anche i gruppi possono aggiungere una certa quota di rumore. Ci sono "folle sagge", il cui giudizio medio si avvicina alla risposta corretta, ma anche folle che sostengono i tiranni, che alimentano bolle speculative, che credono nella magia o cadono nell'incanto di un'illusione collettiva. Piccole differenze possono condurre un gruppo ad assentire con decisione e un altro essenzialmente identico a pronunciare un secco diniego. E, a causa delle dinamiche che si sviluppano tra i membri del gruppo – che qui analizzeremo – il livello di rumore può essere alto. Ciò vale sia per gruppi

simili sia per un unico gruppo il cui giudizio netto su una questione importante dovrebbe essere visto come una tra mille altre possibilità.

Rumore nella musica

Partiamo da un esempio che potrebbe sembrare fuori contesto: uno studio su vasta scala di download musicali condotto da Matthew Salganik e colleghi.[1] Gli studiosi hanno creato un gruppo di controllo di migliaia di persone, costituito dai visitatori di un sito web piuttosto noto, che potevano ascoltare e scaricare una o più canzoni di nuovi gruppi tra le settantadue presenti. Le canzoni avevano nomi a effetto come *Trapped in an Orange Peel*, *Gnaw*, *Eye Patch*, *Baseball Warlock v1* e *Pink Aggression*. (Alcuni dei titoli sembrano strettamente correlati al nostro tema: *Best Mistakes*, *I Am Error*, *The Belief Above the Answer*, *Life's Mystery*, *Wish Me Luck* e *Out of the Woods*.)

Ai partecipanti non veniva riferito niente di quanto era detto o fatto dagli altri, dovevano formarsi dei giudizi indipendenti su quali canzoni gradivano di più e desideravano scaricare. Salganik e i suoi colleghi crearono poi altri otto gruppi, a cui furono assegnati in maniera casuale migliaia di altri visitatori del sito web. Ai membri di questi gruppi veniva chiesto di fare la stessa cosa, ma con una differenza: potevano vedere quante persone all'interno del loro gruppo avevano già scaricato ogni canzone. Tutti i membri di un dato gruppo, per esempio, avrebbero visto se *Best Mistakes* era gettonata o se, al contrario, non l'aveva scaricata nessuno.

Poiché i vari gruppi non differivano per alcuna variabile importante, lo studio non faceva altro che riproporre per otto volte lo stesso copione. Era possibile prevedere che alla fine le canzoni migliori sarebbero sempre state

ai primi posti e quelle peggiori agli ultimi; in tal caso, i vari gruppi sarebbero arrivati a classifiche identiche o quantomeno simili, e tra loro non vi sarebbe stato alcun rumore. E in effetti era proprio questa l'ipotesi su cui Salganik e i suoi colleghi intendevano indagare. Volevano analizzare un particolare vettore di rumore: l'*influenza sociale*.

Il risultato fondamentale fu che le classifiche erano diversissime l'una dall'altra: tra gruppi diversi vi era molto rumore. In un gruppo, *Best Mistakes* era un successo strepitoso e *I Am Error* un grosso flop; in un altro, *I Am Error* si trovava ai primi posti e *Best Mistakes* non piaceva affatto. Se una canzone aveva un successo immediato, poteva piazzarsi molto bene; in caso contrario, l'esito poteva essere assai diverso.

È anche vero che le canzoni peggiori (o meglio, quelle giudicate come tali dal gruppo di controllo) non raggiungevano mai i primi posti, mentre le migliori non finivano mai agli ultimi. Ma, a parte questo, poteva succedere di tutto. Come sottolineano gli autori, «il grado di successo nella condizione di influenza sociale era più imprevedibile che nella condizione indipendente». Insomma, le influenze sociali creano un rumore significativo tra i gruppi. E, se ci pensate, anche i singoli gruppi erano rumorosi, nel senso che il loro giudizio pro o contro una canzone poteva essere ben diverso a seconda che questa avesse raggiunto un'immediata popolarità oppure no.

Come dimostrarono Salganik e i suoi colleghi, i risultati di gruppo possono essere facilmente manipolati, perché la popolarità si autoalimenta.[2] In un esperimento correlato un po' cattivello, gli autori invertirono le classifiche del gruppo di controllo (cioè mentirono sulla popolarità delle canzoni), per cui ai partecipanti veniva detto che le canzoni meno popolari erano le più apprezzate e viceversa. I ricercatori analizzarono i comportamenti dei visitatori del sito web: il risultato fu che

quasi tutte le canzoni meno popolari divennero piuttosto gettonate, e quasi tutte quelle più amate non ebbero alcun successo. All'interno di gruppi molto ampi, la popolarità e l'impopolarità si autoalimentavano, anche quando i partecipanti venivano ingannati sul successo delle canzoni. L'unica eccezione è che, nel tempo, le canzoni in testa alla classifica del gruppo di controllo divennero effettivamente le più popolari, quindi la classifica invertita non riuscì a tenere in fondo le canzoni migliori; tuttavia contribuì in gran parte a determinare la classifica finale.

Dovrebbe essere chiaro che questi studi ci dicono qualcosa sui giudizi di gruppo in generale. Poniamo che un gruppetto di una decina di persone stia decidendo se proporre o no un'iniziativa audace e fuori dagli schemi. Se un paio di promotori dell'iniziativa prendono la parola per primi, è ben possibile che portino dalla loro parte l'intero gruppo. Stessa cosa se a parlare per primi sono gli scettici. O almeno è così se le persone si influenzano a vicenda, come solitamente avviene. Per questo motivo, gruppi per altri versi simili potrebbero arrivare a giudizi molto diversi solo in virtù di chi si è fatto sentire per primo, avviando l'equivalente dei primi download musicali. La popolarità di *Best Mistakes* e *I Am Error* trova un correlativo in giudizi professionali di qualsiasi tipo. E se i gruppi non sono esposti all'analogo delle classifiche di gradimento delle canzoni – per esempio a un forte entusiasmo per una certa iniziativa – quell'iniziativa potrebbe non andare mai in porto semplicemente perché chi la sosteneva non ha espresso la propria opinione.

Oltre i download musicali

Se siete scettici, forse penserete che quello dei download musicali è un caso a sé, che ci dice poco sui giudizi di altri gruppi. Ma osservazioni simili sono

state avanzate anche in molti altri campi.[3] Poniamo, per esempio, la popolarità delle proposte di referendum nel Regno Unito. Per decidere se sostenere o no un referendum, naturalmente la gente deve giudicare se, a conti fatti, sia una buona idea. La traiettoria è simile a quella osservata da Salganik e dai suoi colleghi: un'esplosione di popolarità iniziale andrà ad autoalimentarsi, mentre se una proposta non è sostenuta sin dall'inizio, in pratica è condannata al fallimento. In politica, come in musica, molto dipende dalle influenze sociali e, in particolare, dalla possibilità che le persone vedano l'attrazione o la repulsione degli altri.

Sulla base dell'esperimento dei download musicali, il sociologo Michael Macy della Cornell University e i suoi colleghi si chiesero se le opinioni dichiarate di altre persone potessero d'un tratto rendere determinate posizioni politiche immediatamente riconoscibili, avvicinando così i democratici e allontanando i repubblicani, o viceversa.[4] In sintesi, la risposta è sì. Se in un gruppo online i democratici vedono che un certo punto di vista ha ricevuto una prima approvazione da altri democratici, sosterranno quel punto di vista, che alla fine verrà fatto proprio dalla maggior parte dei membri del suddetto gruppo.[5] Ma se in un altro forum online i democratici vedono che quello stesso punto di vista ha ricevuto una prima approvazione dai repubblicani, rifiuteranno quel punto di vista, che alla fine verrà rigettato dalla maggior parte dei democratici di quel forum. Un comportamento simile si è riscontrato anche tra i repubblicani. In sostanza, forse non c'è una grande differenza tra le posizioni politiche e le canzoni, nel senso che il loro destino ultimo dipende dal gradimento iniziale. Come notarono i ricercatori, «una variazione casuale di pochi che si spostano per primi su una certa posizione» può avere effetti considerevoli nel far pendere ampie fette della popolazione in una certa

direzione, e nel portare repubblicani e democratici a sostenere un insieme di idee che non hanno niente a che fare le une con le altre.

Consideriamo ora una questione direttamente correlata alle decisioni di gruppo in generale: in che modo la gente giudica i commenti lasciati sui siti internet. Lev Muchnik e i suoi colleghi della Hebrew University of Jerusalem hanno svolto un esperimento su un sito web che mostrava articoli di vario tipo e consentiva alla gente di lasciare dei commenti, che a loro volta potevano ricevere una valutazione positiva o negativa. In maniera del tutto automatica e artificiale i ricercatori diedero ad alcuni commenti un'immediata valutazione positiva, prima che ne ricevessero altre. Tenderemmo a pensare che dopo centinaia o migliaia di visitatori e valutazioni quell'unico voto iniziale dato a un commento non abbia alcun peso, ma questa riflessione, per quanto apparentemente sensata, è scorretta. Dopo aver visto una valutazione positiva (che, ricordate, era del tutto fasulla), la probabilità che un visitatore desse un voto favorevole a un commento aumentava del 32%.

L'aspetto interessante è che questo effetto perdurava nel tempo. Dopo cinque mesi un unico voto positivo iniziale aumentava artificialmente la valutazione media dei commenti del 25%. L'effetto di un unico voto positivo iniziale è un tipico esempio di rumore: qualunque sia il motivo alla base di quel voto, questo potrà produrre una grossa variazione del gradimento generale.

Lo studio offre qualche indizio su come si spostano i gruppi e sul perché di tanto rumore (sempre nel senso che gruppi simili possono esprimere giudizi molto diversi, e singoli gruppi possono esprimere giudizi che rappresentano soltanto una possibilità su varie altre). Spesso i loro componenti sono nella posizione di offrire l'equivalente funzionale di un voto positivo o negativo iniziale indicando accordo, neutralità o dissenso; e

se il membro di un gruppo ha dato la sua immediata approvazione, altri membri hanno motivo di fare altrettanto. Non c'è dubbio che quando i gruppi si spostano verso certi prodotti, persone, movimenti e idee, ciò potrebbe non essere dovuto ai loro meriti intrinseci ma all'equivalente funzionale dei primi voti positivi. Naturalmente lo studio di Muchnik coinvolgeva gruppi molto ampi, ma lo stesso può accadere in piccoli gruppi, anche in maniera più drastica, perché un voto positivo iniziale – a favore di un progetto, un prodotto o un verdetto – spesso ha un grande effetto sugli altri.

Vi è un altro aspetto correlato. Abbiamo fatto riferimento alla saggezza della folla: se prendiamo un ampio gruppo di persone e poniamo loro una domanda, c'è una buona possibilità che la risposta media si avvicini a quella giusta; aggregare i giudizi può essere un ottimo modo per ridurre il rumore, e quindi l'errore. Ma cosa accade se le persone ascoltano quello che dicono gli altri? Tenderemmo a supporre che questo atteggiamento possa essere d'aiuto: dopotutto, si può imparare gli uni dagli altri e, così facendo, pervenire alla soluzione corretta. In circostanze favorevoli, quando le persone condividono il proprio sapere, i gruppi di discussione possono pervenire a buoni risultati, ma l'indipendenza è un prerequisito della saggezza della folla. Se la gente, invece di esprimere un proprio giudizio, si basa su ciò che pensano gli altri, non è detto che la folla sia poi così saggia.

Questo problema è stato messo in luce dalla ricerca.[6] In semplici esercizi di stima – sul numero di reati commessi in una città, l'aumento della popolazione in periodi specifici o la lunghezza del confine tra due nazioni – la folla era effettivamente saggia finché manifestava le proprie opinioni in forma indipendente. Ma se si veniva a conoscenza delle stime altrui – per esempio, la stima media di un gruppo di dodici persone – il risultato della

folla era peggiore. Come sostengono gli autori dello studio, le influenze sociali sono un problema, perché «riducono la diversità all'interno dei gruppi senza ridurre l'errore collettivo». Il paradosso è che se opinioni indipendenti diverse, una volta aggregate, possono essere sorprendentemente accurate, anche la minima influenza sociale può produrre una sorta di effetto gregge che mina la saggezza della folla.

Cascata

Alcuni studi qui descritti offrono esempi di *cascate informative*. Queste cascate sono dilaganti e aiutano a spiegare perché gruppi simili in ambito aziendale, governativo e non solo possano seguire direzioni divergenti, e perché dei piccoli cambiamenti possano produrre effetti tanto diversi, e quindi rumore. Noi siamo in grado di osservare la storia solo per come è avvenuta, ma per molti gruppi e decisioni di gruppo esistono insiemi di possibilità di cui una sola troverà realizzazione.

Per comprendere il funzionamento delle cascate informative, immaginate che in un grande ufficio vi siano dieci persone che devono decidere chi assumere per un ruolo importante. I candidati papabili sono tre: Thomas, Sam e Julie. Poniamo che i membri del gruppo esprimano la propria opinione uno dopo l'altro, e che ognuno, com'è ragionevole, tenga conto del giudizio degli altri. Il primo a parlare è Arthur, il quale indica che il miglior candidato per quella posizione è Thomas. Ora che Barbara conosce il giudizio di Arthur, senz'altro dovrebbe concordare con il suo collega se anche a lei Thomas ha fatto un'ottima impressione. Ma poniamo che non sappia ancora con certezza qual è per lei il candidato migliore. Se si fida di Arthur, è possibile che convenga con lui: il migliore è Thomas. E infatti, poiché si fida di Arthur, conferma il suo giudizio.

Spostiamoci ora su una terza persona, Charles. Sia Arthur sia Barbara hanno detto di voler assumere Thomas, ma secondo Charles, sulla base delle limitate informazioni in suo possesso, Thomas non è la persona giusta per quel lavoro; la candidata migliore è Julie. Anche se questa è la sua posizione, Charles potrebbe benissimo metterla da parte e limitarsi a seguire il giudizio di Arthur e Barbara. Questo non perché Charles sia un codardo, ma perché è una persona rispettosa delle opinioni altrui. Forse penserà che Arthur e Barbara avranno avuto i loro buoni motivi per essere entusiasti di Thomas.

A meno che il quarto membro del gruppo, David, non creda davvero di avere informazioni migliori di chi lo ha preceduto, dovrebbe seguire la loro linea, e probabilmente così farà. In questo caso, David sarà coinvolto in un fenomeno a cascata. Certo, se ha motivazioni molto forti per pensare che Arthur, Barbara e Charles si sbagliano, resisterà. Ma, in caso contrario, probabilmente si dirà d'accordo con loro.

È importante sottolineare che Charles o anche David potrebbero avere qualche informazione o intuizione su Thomas (o sugli altri candidati) di cui Arthur e Barbara non sono a conoscenza. Se le avessero esposte, queste informazioni private avrebbero potuto indurre Arthur o Barbara a modificare la propria opinione, e se avessero parlato per primi, non solo avrebbero espresso le proprie opinioni sui candidati, ma avrebbero aggiunto delle informazioni che avrebbero potuto influenzare gli altri partecipanti. Ma poiché hanno parlato per ultimi, è ben possibile che le loro informazioni private restino tali.

Poniamo ora che tocchi a Erica, Frank e George esprimere il proprio parere. Se Arthur, Barbara, Charles e David hanno già detto che Thomas è il migliore, è probabile che anche loro dicano la stessa cosa, pur avendo magari buone ragioni per credere che sarebbe meglio scegliere qualcun

altro. Certo, potrebbero opporsi a questo crescente consenso, se è chiaro che è la scelta sbagliata. Ma se la decisione non è chiara? Il problema, in questo esempio, è che il giudizio iniziale di Arthur ha avviato un processo capace di coinvolgere diverse persone in un fenomeno a cascata che induce il gruppo a optare all'unanimità per Thomas, anche se alcuni dei suoi sostenitori in realtà non sono giunti a una precisa opinione, mentre altri ritengono perfino che non sia affatto il candidato migliore.

Questo esempio, naturalmente, è del tutto fittizio, ma in qualsiasi tipo di gruppo questi processi sono all'ordine del giorno: la gente apprende dagli altri, e se chi parla per primo sembra apprezzare qualcosa o voler fare qualcosa, è possibile che gli altri lo seguano. O almeno questo avviene se non hanno motivi di diffidare di coloro che hanno già espresso la propria opinione.

Per i nostri fini ciò che conta è che le cascate informative aumentano la possibilità che il rumore si insinui nei gruppi. Nell'esempio precedente Arthur ha parlato per primo e ha espresso la sua preferenza per Thomas, ma poniamo che avesse parlato per prima Barbara e che si fosse espressa a favore di Sam, o che Arthur avesse avuto un'impressione leggermente diversa e avesse preferito Julie: è plausibile supporre che il gruppo avrebbe optato per Sam o Julie, non perché i migliori siano loro ma perché la cascata si sarebbe sviluppata in un'altra direzione. È questo il dato centrale che emerge dall'esperimento sui download musicali (e da studi simili).

Il punto non è che seguire queste cascate informative sia irragionevole. Se si è in dubbio sulla persona da assumere, può avere senso seguire gli altri, e via via che aumenta il numero di persone con una certa opinione, fare affidamento su di loro diventa ancora più sensato. Si pongono, però, due problemi. Innanzitutto, si tende a trascurare la possibilità che anche in una folla la maggior parte delle persone è coinvolta in un fenomeno a

cascata e non esprime un giudizio indipendente. Vedendo che tre, dieci o venti persone convengono su una certa conclusione, potremmo sottovalutare fino a che punto stanno seguendo chi si è pronunciato prima di loro. Forse pensiamo che questa consonanza rifletta una saggezza collettiva, mentre in realtà riflette soltanto il parere iniziale di pochi. Secondo, le cascate informative possono portare i gruppi a esiti catastrofici. Dopotutto, Arthur potrebbe anche essersi sbagliato su Thomas.

Naturalmente non sono solo le informazioni a influenzare i membri del gruppo: contano anche le pressioni sociali. In ambito societario o governativo è possibile che le persone non si pronuncino per non sembrare antipatiche, torve, ottuse o sciocche. Vogliono mostrarsi collaborative, e per questo seguono le opinioni e le azioni altrui. Anche quando pensano di sapere cosa è giusto o è probabile che sia giusto, si accodano al consenso apparente del gruppo, o all'opinione di chi parla per primo, per essere benvisti dagli altri.

La storia appena raccontata potrebbe seguire una dinamica di questo tipo non perché i valutatori apprendono gli uni dagli altri quali siano le doti di Thomas, ma perché non vogliono sembrare sciocchi o sgradevoli. Il giudizio iniziale di Arthur a favore di Thomas potrebbe innescare una sorta di effetto traino e finire per imporre una forte pressione sociale su Erica, Frank o George semplicemente perché tutti gli altri si sono detti a favore di Thomas. Come nelle cascate informative, anche in quelle sociali gli interessati potrebbero esagerare la convinzione di chi ha parlato prima di loro. Se sostengono Thomas, è possibile che lo facciano non perché davvero preferiscono lui, ma perché chi ha parlato per primo, o una figura di potere, lo ha sostenuto; eppure i membri del gruppo finiscono per unirsi al coro generale, aumentando così il livello di pressione sociale. Ciò accade

spesso nelle società e negli uffici governativi, e può portare a fare affidamento su un giudizio del tutto sbagliato, e a sostenerlo all'unanimità.

Le influenze sociali producono rumore anche tra diversi gruppi. Se qualcuno apre una riunione promuovendo un grande cambiamento di linea all'interno della società, questa persona potrebbe far partire una discussione che porterà il gruppo a sostenere all'unanimità tale cambiamento. L'accordo unanime potrebbe essere frutto delle pressioni sociali, non di una convinzione. Se qualcun altro avesse aperto la riunione avanzando un'opinione diversa o se chi doveva parlare per primo avesse deciso di stare zitto, la discussione avrebbe potuto prendere una piega del tutto diversa, per la stessa ragione. Gruppi molto simili possono ritrovarsi in posizioni diverse a causa delle pressioni sociali.

Polarizzazione di gruppo

Negli Stati Uniti e in molti altri paesi le cause penali (e molte cause civili) in genere vengono giudicate da una giuria. Si spera che, attraverso il dibattito, le giurie pervengano a decisioni più sagge di quelle che prenderebbero i singoli individui coinvolti in questi organi deliberanti. Tuttavia, lo studio delle giurie rivela un tipo particolare di influenza sociale che è anch'esso una fonte di rumore: la *polarizzazione di gruppo*. L'idea è che quando la gente parla con gli altri, spesso tende a estremizzare le proprie inclinazioni originarie. Se, per esempio, la maggior parte dei componenti di un gruppo di sette persone pensa che aprire un nuovo ufficio a Parigi sia una buona idea, è possibile che il gruppo arrivi a concludere, dopo la discussione, che aprire quell'ufficio sia un'idea straordinaria. Le discussioni interne spesso creano maggiore fiducia, maggiore unità e maggiore estremismo, per lo più sotto forma di maggiore entusiasmo; ebbene, questa

polarizzazione di gruppo non si verifica soltanto nelle giurie, ma spesso anche nei gruppi di lavoro che esprimono giudizi professionali.

In una serie di esperimenti abbiamo analizzato le decisioni di alcune giurie che assegnano danni punitivi nelle cause concernenti la responsabilità di un prodotto. La decisione di ogni giuria, che si traduce nella richiesta di una certa somma, mira a punire la società per i suoi atti illeciti e funge da deterrente per altre aziende. (Torneremo ad approfondire questi studi nel capitolo 15.) Per le nostre finalità, consideriamo un esperimento che pone a confronto giurie deliberative reali e "giurie statistiche".[7] Per prima cosa abbiamo presentato agli 899 partecipanti coinvolti nel nostro studio delle illustrazioni di casi e abbiamo chiesto loro di esprimere un giudizio indipendente in merito, adottando una scala a sette livelli per indicare la propria indignazione e il proprio intento punitivo, e una scala monetaria per gli eventuali risarcimenti in denaro. Poi, con l'aiuto dell'informatica, abbiamo usato queste risposte individuali per creare milioni di giurie statistiche, cioè gruppi virtuali di sei persone (riunite in maniera casuale). In ogni giuria statistica abbiamo assunto come verdetto il valore mediano dei sei giudizi individuali.

Per farla breve, abbiamo scoperto che i giudizi di queste giurie statistiche erano molto più coerenti, e al loro interno il rumore si riduceva notevolmente. Questa riduzione era un effetto automatico dell'aggregazione statistica: il rumore presente in giudizi individuali indipendenti viene sempre ridotto quando si fa una media di tali giudizi.

Le giurie reali, però, non sono giurie statistiche: si incontrano e si scambiano le proprie opinioni sul caso. Ha senso quindi chiedersi se, di fatto, tendano ad arrivare allo stesso giudizio dei loro membri mediani. Per scoprirlo, abbiamo condotto un secondo esperimento, che questa volta

coinvolgeva più di tremila cittadini idonei e più di cinquecento giurie composte da sei persone.[8]

I risultati non lasciavano dubbi: analizzando i singoli casi si è riscontrato che le giurie deliberative erano molto più rumorose di quelle statistiche, un evidente riflesso del rumore causato dall'influenza sociale. Il dibattito aveva l'effetto di aumentare il rumore.

È emerso anche un altro dato interessante. Quando il membro mediano di un gruppo di sei persone era solo moderatamente indignato e si pronunciava a favore di una pena indulgente, in genere il verdetto della giuria deliberativa risultava ancora più indulgente. Quando, al contrario, il membro mediano di un gruppo di sei giurati era piuttosto indignato ed esprimeva un severo intento punitivo, in genere la giuria deliberativa finiva per essere ancora più indignata e più severa. E quando questa indignazione veniva espressa in termini di risarcimento monetario, vi era una tendenza sistematica a stabilire risarcimenti più elevati di quelli proposti dal membro mediano della giuria. Anzi, il 27% dei giurati sceglieva un risarcimento pari o perfino superiore a quello del membro più severo. Non solo le giurie deliberative erano più rumorose di quelle statistiche, ma accentuavano le opinioni dei loro membri.

Ricordiamo la conclusione sulla polarizzazione di gruppo espressa in precedenza: quando la gente si confronta con gli altri, in genere porta all'estremo le proprie inclinazioni originarie. Il nostro esperimento illustra questo effetto. Le giurie deliberative registravano uno spostamento verso una maggiore indulgenza (quando il membro mediano era indulgente) o verso una maggiore severità (quando il membro mediano era severo). Analogamente, le giurie inclini a imporre pene pecuniarie finivano per comminare pene più severe di quelle proposte dai loro membri mediani.

La polarizzazione di gruppo, a sua volta, ha una spiegazione simile a quella degli effetti a cascata. Anche qui le informazioni hanno un ruolo fondamentale: se la maggior parte delle persone propone una pena severa, nel gruppo emergeranno molti argomenti a favore della pena severa e pochi argomenti contrastanti; se i membri del gruppo ascoltano i pareri degli altri, si sposteranno verso la tendenza dominante, il che renderà il gruppo più compatto, più sicuro dei propri giudizi e più estremo, e se i singoli tengono alla propria reputazione all'interno del gruppo, si sposteranno nella direzione dominante, e anche questo porterà a una polarizzazione.

La polarizzazione di gruppo, naturalmente, può indurre in errore, come spesso accade. Ma a noi qui interessa la variabilità. Come abbiamo visto, un aggregato di giudizi ridurrà il rumore, e in quest'ottica più giudizi ci sono, meglio è. Allo stesso tempo, scopriamo che le giurie deliberative sono più soggette a rumore di quelle statistiche. Quando gruppi che partono da posizioni simili finiscono per divergere, spesso il motivo sta nella polarizzazione di gruppo; e il rumore che ne risulta può essere molto alto.

In ambito societario, governativo e non solo, le cascate e la polarizzazione possono produrre ampie disparità tra gruppi che esaminano lo stesso problema. Il fatto che gli esiti dei giudizi possano dipendere da alcuni individui – chi prende la parola per primo o chi ha più influenza – dovrebbe preoccuparci ulteriormente, ora che sappiamo quanto rumore possa esserci nei giudizi individuali. Abbiamo visto che il rumore di livello e quello strutturale portano a una differenza di opinioni più ampia del dovuto (e del previsto) tra i membri di un gruppo, e che il rumore occasionale – stanchezza, umore, precedenti – potrebbe influire sul giudizio di chi si pronuncia per primo. Le dinamiche di gruppo possono amplificare questo rumore. Di conseguenza, i gruppi deliberativi sono

tendenzialmente più rumorosi di quelli statistici che si limitano a presentare la media dei giudizi individuali.

Poiché molte delle decisioni aziendali e governative più importanti vengono prese a seguito di qualche tipo di processo deliberativo, è cruciale essere consapevoli di questo rischio. Le organizzazioni e i loro leader dovrebbero prendere provvedimenti per controllare il rumore nei giudizi individuali dei loro membri, e gestire i gruppi deliberativi in modo che tendano a ridurre il rumore, non ad amplificarlo. Le strategie di riduzione del rumore che proporremo puntano al raggiungimento di questo obiettivo.

A proposito delle decisioni di gruppo

«Tutto sembra dipendere dal gradimento iniziale. Bisogna lavorare per far sì che il lancio di un nuovo prodotto abbia successo nella prima settimana.»

«Come ho sempre sospettato, le idee politiche ed economiche sono come le stelle del cinema: se la gente pensa che agli altri piacciano, finiscono per imporsi.»

«Mi ha sempre preoccupato il fatto che quando il mio gruppo di lavoro si riunisce, ne esce più fiducioso e più compatto, e pronto a gettarsi a testa bassa nel piano d'azione stabilito. Presumo che nei nostri processi interni ci sia qualcosa che non va!»

[1] M.J. Salganik, P. Sheridan Dodds, D.J. Watts, *Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market*, in "Science", 311(2006), pp. 854-856. Vedi anche M.J. Salganik, D.J. Watts, *Leading the Herd Astray: An Experimental Study of Self-Fulfilling Prophecies in an Artificial Cultural Market*, in "Social Psychology Quarterly", 71(2008), pp. 338-355; Iid., *Web-Based Experiments for the Study of Collective Social Dynamics in Cultural Markets*, in "Topics in Cognitive Science", 1(2009), pp. 439-468.

[2] Iid., *Leading the Herd Astray*, cit.

[3] M. Macy *et al.*, *Opinion Cascades and the Unpredictability of Partisan Polarization*, in "Science Advances" (2019), pp. 1-8. Vedi anche H. Margetts *et al.*, *Political Turbulence*, Princeton University Press, Princeton 2015.

[4] M. Macy *et al.*, *Opinion Cascades*, cit.

[5] L. Muchnik *et al.*, *Social Influence Bias: A Randomized Experiment*, in "Science", 341(2013), n. 6146, pp. 647-651.

[6] J. Lorenz *et al.*, *How Social Influence Can Undermine the Wisdom of Crowd Effect*, in "Proceedings of the National Academy of Sciences", 108(2011), n. 22, pp. 9020-9025.

[7] D. Kahneman, D. Schkade, C. Sunstein, *Shared Outrage and Erratic Awards: The Psychology of Punitive Damages*, in "Journal of Risk and Uncertainty", 16(1998), pp. 49-86.

[8] Iid., *Deliberating about Dollars: The Severity Shift*, in "Columbia Law Review", 100(2000), pp. 1139-1175.

TERZA PARTE

Il rumore nei giudizi predittivi

Molti giudizi non sono altro che previsioni e, poiché le previsioni verificabili si possono valutare, studiandole possiamo imparare tanto sul rumore e sul bias. In questa parte ci concentreremo sui giudizi predittivi.

Nel capitolo 9 confronteremo l'accuratezza delle previsioni effettuate da esperti e da macchine, o derivanti da semplici regole. Ne concluderemo, non a sorpresa, che gli esperti si piazzano all'ultimo posto. Nel capitolo 10 indagheremo sulle ragioni alla base di questo risultato e mostreremo che il rumore incide pesantemente sull'inferiorità del giudizio umano.

Per arrivare a tali conclusioni dovremo valutare la qualità delle previsioni, per cui ci servirà una misura dell'accuratezza predittiva che ci permetta di rispondere alla seguente domanda: qual è il rapporto di *covarianza* tra previsioni ed esiti? Se la divisione risorse umane ha il compito di valutare il potenziale delle nuove assunzioni, per esempio, basterà aspettare qualche anno per conoscere le effettive prestazioni dei dipendenti e studiare il rapporto di covarianza tra le stime del potenziale e le valutazioni delle prestazioni. Le previsioni sono accurate nella misura in cui i dipendenti il cui potenziale è stato giudicato elevato al momento dell'assunzione ricevono valutazioni elevate per il proprio lavoro.

Una misura che sfrutta questa intuizione è la *percentuale di coppie concordanti* (PC),[1] che risponde a una domanda precisa. Prendiamo un paio di dipendenti a caso: qual è la probabilità che quello con un punteggio superiore nella valutazione del potenziale si dimostri poi migliore anche

nella prestazione lavorativa? Se l'accuratezza delle prime valutazioni fosse perfetta, avremmo una PC del 100%: la valutazione del potenziale dei due dipendenti sarebbe una previsione perfetta della valutazione successiva della loro performance. Se le previsioni fossero del tutto inefficaci, la concordanza si verificherebbe in maniera del tutto casuale e il dipendente "a elevato potenziale" avrebbe la medesima probabilità di raggiungere o non raggiungere prestazioni elevate: qui la PC sarebbe del 50%. Ci soffermeremo su questo esempio, che è stato oggetto di un'analisi approfondita, nel capitolo 9. Per fare un esempio più semplice, negli uomini adulti la PC tra la lunghezza dei piedi e l'altezza è del 71%. Guardando due persone dalla testa ai piedi, c'è il 71% di probabilità che il più alto dei due abbia anche i piedi più lunghi.

La PC ha il vantaggio di essere una misura della covariazione molto intuitiva, ma non è la più utilizzata nelle scienze sociali. La misura standard è invece il *coefficiente di correlazione* (*r*), che varia da 0 a 1 quando due variabili hanno una relazione positiva. Nell'esempio precedente, la correlazione tra altezza e lunghezza dei piedi è pari a 0,60.[2]

Si può pensare al coefficiente di correlazione da diversi punti di vista. Il primo, piuttosto intuitivo, è che la correlazione tra due variabili corrisponde alla percentuale di determinanti che hanno in comune. Poniamo, per esempio, che un certo tratto sia puramente genetico; per quel tratto ci aspetteremmo una correlazione di 0,50 tra fratelli, che hanno il 50% di geni in comune, e una correlazione di 0,25 tra cugini di primo grado, che hanno il 25% dei geni in comune. Potremmo anche dire che la correlazione 0,60 tra altezza e lunghezza dei piedi indica che il 60% dei fattori causali che determinano l'altezza determina anche il numero di scarpe.

Le due misure di covariazione che abbiamo descritto sono direttamente correlate tra loro. La tabella 1 presenta la PC per diversi valori del coefficiente di correlazione.[3] D'ora in avanti presenteremo sempre entrambe le misure quando analizzeremo le prestazioni di esseri umani e modelli informatici.

Tabella 1. Coefficiente di correlazione e percentuale di coppie concordanti (PC)

Coefficiente di correlazione	*Percentuale di coppie concordanti (PC)*
,00	50%
,10	53%
,20	56%
,30	60%
,40	63%
,60	71%
,80	79%
1,00	100%

Nel capitolo 11 affronteremo un limite importante dell'accuratezza predittiva: il fatto che quasi sempre i giudizi vengano espressi in uno stato che potremmo definire di *ignoranza oggettiva*, perché è impossibile conoscere molti degli elementi da cui dipende il futuro. Detto ciò, quasi sempre riusciamo a dimenticarci di questa limitazione e avanziamo previsioni con sicurezza (anzi, con *eccessiva* sicurezza). Infine, nel capitolo 12 mostreremo che l'ignoranza oggettiva influisce sulla nostra capacità

non solo di prevedere gli eventi ma anche di comprenderli. Questo ci aiuterà a sciogliere l'enigma dell'invisibilità del rumore.

[1] La percentuale di coppie concordanti (PC) è strettamente correlata alla *W* di Kendall, anche nota come coefficiente di concordanza.

[2] K. Kamboj *et al.*, *A Study on the Correlation Between Foot Length and Height of an Individual and to Derive Regression Formulae to Estimate the Height from Foot Length of an Individual*, in "International Journal of Research in Medical Sciences", 6(2018), n. 2, p. 528.

[3] La PC viene calcolata a partire dal presupposto che la distribuzione congiunta sia bivariata normale, e i valori mostrati nella tabella sono approssimazioni fondate su tale presupposto. Si ringrazia Julian Parris per l'elaborazione della tabella.

9

Giudizi e modelli

Molte persone hanno interesse a prevedere le prestazioni lavorative future, proprie e altrui, pertanto previsioni di questo tipo costituiscono un utile esempio di giudizio professionale predittivo. Prendiamo due dirigenti di una grande azienda. Al momento dell'assunzione, Monica e Nathalie sono state valutate da una società di consulenza specializzata, che ha assegnato loro un voto su una scala da uno a dieci nei seguenti campi: leadership, comunicazione, competenze interpersonali, competenze tecniche specifiche e motivazione rispetto alla posizione da ricoprire (tabella 2). Vi chiediamo di prevedere come verranno valutate le loro prestazioni, su una scala da uno a dieci, a distanza di due anni dall'assunzione.

Tabella 2. Due candidate per una posizione dirigenziale

	Leadership	*Comunicazione*	*Competenze interpersonali*	*Competenze tecniche*	*Motivazione*	*La vostra previsione*
Monica	4	6	4	8	8	
Nathalie	8	10	6	7	6	

Di fronte a un problema di questo tipo, la maggior parte delle persone dà un'occhiata alle varie voci ed esprime un giudizio immediato, a volte dopo aver calcolato a mente la media dei punteggi. Se l'avete fatto anche voi, probabilmente sarete giunti alla conclusione che la candidata più forte sia Nathalie, con una differenza di un paio di punti rispetto a Monica.

Giudizio o formula?

L'approccio informale a un problema come questo viene definito *giudizio clinico*. Si tiene conto di alcune informazioni, magari si effettua un rapido calcolo, si arriva a un'intuizione e si esprime un giudizio. In effetti il giudizio clinico è il processo che fin qui abbiamo chiamato semplicemente "giudizio".

Ora immaginiamo che abbiate svolto questa previsione in qualità di partecipanti a un esperimento. Monica e Nathalie sono state estratte da una banca dati di diverse centinaia di dirigenti assunti qualche anno fa e valutati sulla base di cinque dimensioni. Voi avete impiegato queste valutazioni per prevedere il loro successo lavorativo; ora sono disponibili le valutazioni delle loro effettive prestazioni nei rispettivi ruoli. Quanto si avvicinano queste valutazioni ai vostri giudizi clinici sul loro potenziale?

Questo esempio è liberamente basato su uno studio reale di previsioni delle prestazioni.[1] Se aveste partecipato a questo studio, probabilmente non sareste soddisfatti dei risultati: un gruppo di ricercatori in psicologia assunto da una società di consulenza internazionale per elaborare questo tipo di previsioni ha raggiunto una correlazione pari a 0,15 con le valutazioni delle prestazioni (PC = 55%). In altri termini, nel dare a un candidato una valutazione migliore rispetto a un altro – come nel caso di Monica e Nathalie – la probabilità che il favorito arrivasse a una valutazione delle prestazioni più elevata era del 55%, di poco superiore alla semplice casualità. Insomma, un risultato non certo brillante.

Forse penserete che le loro previsioni non siano state accurate perché gli indici presi in esame non consentivano di fare previsioni, ma questo fa nascere un'altra domanda: quante informazioni predittive utili contengono davvero gli indici dei candidati? Come si possono aggregare in un punteggio predittivo che dimostri di avere la più elevata correlazione possibile con le prestazioni future?

Per rispondere a queste domande consideriamo un metodo statistico classico, che nello studio in questione ha prodotto una correlazione ottimale di 0,32 (PC = 60%), non certo brillante ma significativamente più elevata dei risultati delle previsioni cliniche.

Questa tecnica, chiamata *regressione multipla*, produce un punteggio predittivo costituito dalla media ponderata dei predittori,[2] giungendo a una ponderazione ottimale che viene poi scelta per massimizzare la correlazione tra la previsione composita e la variabile di destinazione. La ponderazione ottimale minimizza il MSE (l'errore quadratico medio) delle previsioni, in quello che è un ottimo esempio del ruolo fondamentale del principio dei minimi quadrati in statistica. Come potrete immaginare, al predittore più strettamente correlato alla variabile di destinazione viene attribuito un peso maggiore,[3] a quelli ininfluenti un peso pari a zero. Il peso può anche essere negativo: il numero di multe inevase del candidato probabilmente avrà un peso negativo come predittore del successo manageriale.

L'uso della regressione multipla è un esempio di *previsione meccanica*. Ve ne sono diversi tipi, che vanno da semplici regole (come "assumere chiunque abbia un diploma") a sofisticati modelli di intelligenza artificiale, ma i modelli di regressione lineare, che sono stati definiti «l'asse portante della ricerca sul giudizio e sul processo decisionale»,[4] sono i più comuni. Per ridurre al minimo il gergo specialistico, li chiameremo *modelli semplici*.

Lo studio esemplificato con i casi di Monica e Nathalie proponeva, come tanti altri, un raffronto tra previsioni cliniche e meccaniche, tutte basate su una struttura molto semplice:[5]

- una serie di *variabili predittive* (nel nostro esempio, i punteggi dei candidati) viene impiegata per prevedere un *esito di destinazione* (le valutazioni lavorative delle stesse persone);
- dei giudici umani elaborano *previsioni cliniche*;
- una regola (come la regressione multipla) impiega gli stessi predittori per produrre *previsioni meccaniche* degli stessi risultati;

- viene posta a confronto l'accuratezza generale delle previsioni cliniche e delle previsioni meccaniche.

Meehl: il modello ottimale batte l'uomo

Quando si entra nel mondo delle previsioni cliniche e meccaniche, la prima cosa che si desidera sapere è in quale rapporto sono le une con le altre. Il giudizio umano è migliore di una formula?

Questa domanda, già sollevata in precedenza, attirò grande attenzione solo nel 1954, quando Paul Meehl, professore di psicologia alla University of Minnesota, pubblicò uno studio dai risultati sorprendenti.[6] Meehl passò in rassegna venti ricerche in cui venivano contrapposti un giudizio clinico e una previsione meccanica su esiti relativi al successo accademico e alle prognosi psichiatriche, giungendo alla conclusione sconvolgente che in genere alcune semplici regole meccaniche si rivelavano più efficaci del giudizio umano. Meehl scoprì che purtroppo i clinici, così come altri professionisti, sono molto deboli proprio in quello che ritengono il loro punto di forza, ovvero la capacità di integrare le informazioni.

Per comprendere la portata di questo risultato anche in rapporto al rumore, occorre spiegare come funziona un semplice modello di previsione meccanica. Il tratto essenziale è che applica la stessa regola a tutti i casi: ogni predittore ha un peso che resta invariato da caso a caso. Saremmo portati a pensare che un vincolo tanto rigido costituisca un grande svantaggio rispetto ai giudizi umani. Nel nostro esempio, forse avete pensato che l'insieme di motivazione e competenze tecniche di Monica costituisse una marcia in più, che avrebbe compensato i suoi limiti in altri ambiti. E magari avete anche pensato che le carenze di Nathalie in questi due campi non rappresentassero un serio problema, considerati gli altri suoi punti di forza. Implicitamente, dunque, avete immaginato diversi percorsi di successo per le due donne. Queste ipotesi cliniche plausibili, in effetti, assegnano un peso

diverso agli stessi predittori nei due casi, una sottigliezza a cui un modello semplice non arriverà mai.

Un altro vincolo del modello semplice è che un aumento di 1 unità in un predittore produce sempre il medesimo effetto (che sarà sempre la metà di quello prodotto da un aumento di 2 unità). Spesso l'intuizione clinica viola questa regola: se, per esempio, quel 10 di Nathalie nelle capacità comunicative vi ha colpito e ha fatto schizzare in alto le vostre previsioni, avete agito come un modello semplice non farebbe mai. In un calcolo della media ponderata la differenza tra i punteggi 10 e 9 sarà pari a quella tra 7 e 6. Il giudizio clinico non obbedisce a questa regola, ma al contrario riflette l'intuizione comune che la stessa differenza possa essere priva di conseguenze in un contesto e determinante in un altro. Sospettiamo, quindi, che nessun modello semplice perverrebbe esattamente ai vostri stessi giudizi su Monica e Nathalie.

Lo studio che abbiamo utilizzato per questi casi offre un chiaro esempio del modello di Meehl. Come abbiamo visto, le previsioni cliniche hanno raggiunto una correlazione di 0,15 (PC = 55%) con le prestazioni lavorative, mentre la previsione meccanica è arrivata a una correlazione di 0,32 (PC = 60%). Ripensate alla sicurezza con cui avete valutato le abilità di Monica e Nathalie. I risultati di Meehl indicano che la vostra soddisfazione sulla qualità dei giudizi da voi espressi era solo un'illusione: l'*illusione di validità*.

Questa illusione si ritrova ogniqualvolta vengano formulati dei giudizi predittivi, perché in genere non si opera una distinzione tra due fasi dell'attività predittiva: valutare i casi sulla base dei dati disponibili e prevedere gli esiti reali. Spesso vi sentirete sicuri della vostra valutazione sul candidato che *sembra* il migliore, ma indovinare chi di loro lo *sia davvero* è un altro paio di maniche. Si può tranquillamente asserire, per esempio, che Nathalie sembri una candidata più valida di Monica, ma non che Nathalie sarà

una dirigente migliore di Monica. Il motivo è evidente: avete quasi tutte le informazioni utili per valutare i due casi, ma nessuna certezza sul futuro.

Purtroppo questa differenza non ci è così chiara quando elaboriamo un pensiero. Se fate confusione tra casi e previsioni, siete in ottima compagnia: tutti noi troviamo difficile operare questa distinzione. Se siete sicuri delle vostre previsioni tanto quanto della vostra valutazione dei casi, però, incorrerete nell'illusione di validità.

Neanche i professionisti del campo sanitario ne sono immuni. Provate a immaginare la reazione degli psicologi ai risultati di Meehl, da cui emergeva che alcune formule banalissime, se applicate in maniera coerente, sono più attendibili dei giudizi clinici. Sconvolti da questi risultati, manifestarono incredulità e disprezzo per quella ricerca secondo loro superficiale, che pretendeva di analizzare al microscopio i prodigi dell'intuizione clinica. La loro reazione era comprensibilissima: il modello di Meehl contraddice l'esperienza di giudizio soggettiva, che la maggior parte di noi anteporrebbe anche all'autorità di uno studioso.

Lo stesso Meehl aveva un atteggiamento ambivalente nei confronti di quei risultati. Dal momento che il suo nome oggi viene associato all'idea della superiorità della statistica rispetto al giudizio clinico, potremmo pensare che fosse un critico agguerrito dell'intuizione umana o un progenitore delle teorie quantistiche, diremmo oggi. Eppure Meehl era lontanissimo da questa immagine caricaturale. Parallelamente alla carriera accademica, esercitava la professione di psicoanalista, e nel suo ufficio aveva un ritratto di Freud.[7] Era un personaggio eclettico che insegnava non solo psicologia, ma anche filosofia e diritto, e scrisse di metafisica, religione, scienze politiche e perfino parapsicologia.[8] (Insisteva per esempio che «la telepatia non è priva di fondamento».) Nessuna di queste caratteristiche avvalora lo stereotipo dell'uomo di scienza che bada solo alle nude cifre. Meehl non aveva niente contro i clinici, al contrario. Si limitò a registrare che le evidenze riguardo ai

vantaggi dell'approccio meccanico nel combinare gli input erano «imponenti e sistematiche».[9]

La sua è una descrizione impeccabile. Un'analisi di centotrentasei studi effettuata nel 2000 ha confermato in maniera schiacciante che l'aggregazione meccanica è più efficace del giudizio clinico.[10] Le ricerche esaminate nell'articolo toccavano un'ampia varietà di temi, come le diagnosi di itterizia, l'idoneità al servizio militare e la soddisfazione coniugale. La previsione meccanica era più accurata in sessantatré studi, in altri sessantacinque si evidenziava una parità statistica, mentre la previsione clinica dava risultati migliori in soli otto casi. Questi risultati semmai sottostimano i vantaggi della previsione meccanica, peraltro più rapida e più economica del giudizio clinico. Inoltre, in molti di questi studi i valutatori umani partivano da una situazione di vantaggio, perché avevano accesso a informazioni "private" non fornite al modello informatico.[11] I risultati conducono a una chiara conclusione: *i modelli semplici battono gli esseri umani.*

Goldberg: il modello dell'uomo batte l'uomo

I risultati di Meehl sollevano interrogativi importanti. Perché la formula è più precisa? Cosa riesce a fare meglio di noi? Forse sarebbe meglio chiedersi cosa fanno peggio gli esseri umani. La risposta è che siamo inferiori ai modelli statistici per molti versi, e una delle nostre maggiori debolezze sta proprio nel rumore.

Per suffragare questa conclusione passiamo a un'altra corrente di ricerca sui modelli semplici avviata nella cittadina di Eugene, in Oregon, da Paul Hoffman, un ricco psicologo visionario che non sopportava il mondo accademico. Hoffman fondò un istituto di ricerca che si avvaleva di studiosi di altissimo livello, grazie ai quali Eugene divenne un centro famoso in tutto il mondo per gli studi sul giudizio umano.

Tra questi studiosi vi era Lewis Goldberg, molto noto per il ruolo fondamentale che rivestì nello sviluppo della teoria dei Big Five, o cinque tratti di personalità. Alla fine degli anni sessanta, sulla scia di Hoffman, Goldberg studiò alcuni modelli statistici utili a descrivere i giudizi di un individuo.[12]

Elaborare un modello di questo tipo equivale a elaborare un modello di realtà: vengono impiegati gli stessi predittori. Nel nostro esempio iniziale i predittori erano i cinque indici relativi alle prestazioni di un dirigente. Anche qui viene adottato lo strumento della regressione multipla; l'unica differenza sta nella variabile di destinazione: invece di prevedere una serie di risultati reali, la formula viene applicata per prevedere una serie di giudizi, per esempio i *vostri* giudizi su Monica, Nathalie e altri dirigenti.

L'idea di elaborare un modello dei vostri giudizi sotto forma di media ponderata potrà sembrarvi strana, perché non è così che arrivate a formarvi le vostre opinioni. Nel vostro pensiero clinico su Monica e Nathalie, non avete applicato le stesse regole in entrambi i casi; anzi, non avete applicato alcuna regola. Il modello di un soggetto giudicante non offre una descrizione realistica del modo in cui questi di fatto giudica.

Tuttavia, anche se non impiegate una formula lineare, potete sempre esprimere un giudizio *come se* la impiegaste. Gli esperti di biliardo giocano come se avessero risolto le complesse equazioni che descrivono la meccanica di un particolare tiro, anche se ovviamente non è così.[13] Analogamente, anche voi potreste generare delle previsioni come se usaste una semplice formula, anche se in realtà fate qualcosa di molto più complesso. Un modello del tipo "come se", che preveda in modo ragionevolmente accurato come si comporterà qualcuno, è utile anche quando non descrive il processo in maniera corretta. I modelli di giudizio semplici rientrano in questa categoria. Da una rassegna completa delle ricerche sul giudizio è emerso che, in 237 studi, la correlazione media tra il modello del soggetto giudicante e i suoi

giudizi clinici era di 0,80 (PC = 79%). Questa correlazione, benché non perfetta, è sufficiente per giustificare il ricorso a una teoria del tipo "come se".[14]

Alla base della ricerca di Goldberg vi era il problema di verificare in quale misura un modello semplice del soggetto giudicante sarebbe stato in grado di prevedere degli esiti reali. Poiché il modello parte da un'approssimazione semplicistica del soggetto, potremmo ragionevolmente supporre che le sue prestazioni saranno inferiori: quanto si perde in termini di accuratezza quando si sostituisce il modello al soggetto giudicante?

La risposta potrebbe sorprendervi. Le previsioni generate dal modello non erano meno accurate; anzi, nella maggior parte dei casi il modello offriva previsioni migliori rispetto al professionista su cui era basato. Il surrogato era meglio dell'originale.

Questa conclusione trova conferma in molti studi effettuati in svariati campi. Una prima replica della ricerca di Goldberg fu applicata alla previsione del rendimento degli studenti universitari.[15] I ricercatori chiesero a novantotto partecipanti di prevedere la media finale dei voti di novanta studenti a partire da dieci indicazioni. Sulla base di queste previsioni i ricercatori costruirono un modello lineare dei giudizi di ciascun partecipante e confrontarono l'accuratezza con cui i partecipanti e i modelli degli stessi riuscirono a prevedere la media finale dei voti degli studenti. In tutti e novantotto i casi il modello diede un risultato migliore di quello del relativo partecipante! A distanza di decenni, una rassegna di cinquant'anni di studi giunse alla conclusione che le prestazioni dei modelli erano costantemente superiori a quelle dei soggetti su cui erano basati.[16]

Non sappiamo se i partecipanti abbiano ricevuto un riscontro sulle proprie prestazioni, ma possiamo benissimo immaginare il loro sgomento nell'apprendere che un modello rudimentale dei loro giudizi, quasi una caricatura, in realtà era più accurato di loro. Tutti noi riteniamo il giudizio un'attività complessa, stimolante e interessante proprio perché non

riducibile a qualche semplice regola. La nostra autostima in merito alla capacità di formulare giudizi sale quando inventiamo e applichiamo regole complesse, o quando abbiamo un'intuizione che rende un caso individuale diverso dagli altri: insomma, quando formuliamo dei giudizi non riducibili al semplice calcolo della media ponderata. Lo studio basato sui modelli dei soggetti giudicanti convalida la conclusione di Meehl secondo cui la finezza di pensiero va quasi sempre sprecata: in genere la complessità e la creatività non conducono a previsioni più accurate.

Per quale motivo succede? Per comprendere i risultati di Goldberg è necessario capire in cosa differiscono una persona e il suo modello. Da dove deriva questa discrepanza tra i giudizi reali e le previsioni di un modello semplice?

Un modello statistico dei vostri giudizi certo non potrà aggiungere nulla alle informazioni in essi contenute, ma solo sottrarre e semplificare. In particolare, non rappresenterà le regole complesse che seguite di volta in volta: se pensate che la differenza tra 10 e 9 in una valutazione delle competenze comunicative sia più significativa della differenza tra 7 e 6, o che un candidato con svariate abilità testimoniate da una sfilza di 7 in tutte le dimensioni considerate sia preferibile a un altro che ottiene la stessa media con chiari punti di forza ma anche notevoli lacune, il modello costruito su di voi non riprodurrà queste vostre regole complesse, anche se voi le applicate con una coerenza impeccabile.

Questa mancata riproduzione delle vostre regole più sofisticate comporterà una perdita di accuratezza quando la vostra finezza di pensiero è giustificata. Poniamo, per esempio, che dobbiate prevedere l'esito di un compito difficile a partire da due elementi, la competenza e la motivazione; in questo caso, una media ponderata non è la formula migliore, perché non c'è motivazione che possa supplire a una grave incompetenza e viceversa. Se impiegherete una combinazione più complessa dei due elementi,

l'accuratezza delle vostre previsioni aumenterà e supererà quella raggiunta da un modello che non riesce a cogliere queste sottigliezze. D'altro canto, spesso le regole complesse non faranno che darvi un'illusione di validità, inficiando la qualità dei vostri giudizi. Alcune sottigliezze sono giustificate, ma molte altre no.

Inoltre, un modello semplice costruito su voi stessi non rappresenterà il rumore strutturale presente nei vostri giudizi. Non potrà replicare gli errori positivi e negativi derivanti dalle reazioni arbitrarie suscitate da un caso particolare, né coglierà il condizionamento di un particolare contesto e del vostro stato mentale nel momento in cui esprimete un certo giudizio. Con ogni probabilità questi rumorosi errori di giudizio non hanno una correlazione sistematica con alcun elemento, per cui quasi sempre potranno essere considerati casuali.

La rimozione del rumore dal giudizio condurrà sempre a un miglioramento dell'accuratezza predittiva.[17] Poniamo, per esempio, che la correlazione tra le vostre previsioni e il risultato sia di 0,50 (PC = 67%), ma il 50% della varianza dei vostri giudizi sia rappresentata dal rumore. Eliminando quel rumore – come farebbe un modello basato su di voi – la correlazione dei giudizi con gli stessi risultati schizzerebbe a 0,71 (PC = 75%). Ridurre meccanicamente il rumore aumenta la validità dei giudizi predittivi.

In sostanza, sostituirvi con un modello di voi stessi avrebbe due effetti: eliminerebbe tanto la vostra finezza di pensiero quanto il rumore sistemico. Il dato comprovato che il modello del soggetto giudicante sia più valido del soggetto stesso ci porta a una conclusione importante: in genere i vantaggi delle regole sofisticate, ove presenti, non bastano a compensare gli effetti infausti del rumore nel giudizio umano. Forse vi riterrete più sagaci, più acuti e più sfumati della caricatura lineare del vostro pensiero, ma quasi sempre siete anche più rumorosi.

Perché le regole complesse vanno a discapito dell'accuratezza nelle previsioni, anche se a noi sembra che partano da intuizioni valide? Tanto per cominciare, molte di queste regole inventate da noi difficilmente saranno sempre vere. C'è poi un altro problema: anche quando, in teoria, sono valide, richiedono condizioni che non vengono quasi mai soddisfatte. Mettiamo, per esempio, che siate giunti alla conclusione che valga la pena di assumere dei candidati molto originali, anche quando il loro punteggio in altri parametri è mediocre. Il problema è che, per definizione, i candidati molto originali sono molto rari: poiché difficilmente una valutazione dell'originalità risulterà affidabile, un punteggio alto in quella variabile sarà in molti casi una coincidenza, mentre il vero talento originale passerà quasi sempre inosservato. Inoltre le valutazioni delle prestazioni che potrebbero confermare il grande successo delle persone originali sono anch'esse imperfette. Inevitabilmente, errori di misurazione su entrambi i fronti attenuano la validità delle previsioni, ed è probabile che un evento raro non venga notato. Così, i vantaggi della vera sottigliezza di giudizio vengono soffocati dagli errori di misurazione.

Una ricerca condotta da Martin Yu e Nathan Kuncel ha portato alle estreme conseguenze la dimostrazione di Goldberg.[18] In questo studio (che è alla base dell'esempio di Monica e Nathalie) sono stati utilizzati i dati di una società di consulenza internazionale che si è avvalsa di alcuni esperti per la valutazione di 847 candidati a ruoli dirigenziali, in tre campioni distinti. Gli esperti hanno espresso un voto sulla base di sette diverse dimensioni valutative e hanno impiegato il proprio giudizio clinico per assegnare un punteggio predittivo complessivo a ciascun candidato, con risultati piuttosto insoddisfacenti.

Yu e Kuncel hanno deciso di porre a confronto questi valutatori non con un semplice modello basato su di loro, ma con un modello lineare *casuale*. Hanno generato diecimila serie di pesi casuali per i sette predittori, per poi

applicare queste diecimila formule alla previsione delle prestazioni professionali.[19]

Il risultato è stato straordinario: *qualsiasi* modello lineare, se applicato coerentemente a tutti i casi, aveva un'alta probabilità di superare il giudizio umano nella previsione di un risultato, a partire dalle stesse informazioni. In uno dei tre campioni, il 77% dei diecimila modelli lineari con ponderazione casuale ha dato risultati migliori degli esperti, e negli altri due campioni i modelli casuali hanno fatto meglio degli esseri umani nel 100% dei casi. Per dirla senza troppi giri di parole, in quello studio si è dimostrato impossibile generare un modello semplice che dia risultati peggiori degli esperti.

Questa ricerca ci porta a una conclusione ancora più schiacciante di quella tratta dallo studio di Goldberg sul modello del soggetto giudicante, di cui rappresenta un esempio estremo. In questo contesto, gli esseri umani hanno espresso giudizi inefficaci in termini assoluti, il che spiega come mai anche degli insignificanti modelli lineari abbiano dato risultati migliori. Naturalmente non dobbiamo concluderne che qualsiasi modello farà meglio di qualsiasi essere umano. Detto ciò, il fatto che l'aderenza meccanica a una semplice regola (Yu e Kuncel l'hanno definita "coerenza cieca") conducesse a un giudizio decisamente più accurato su un problema difficile illustra l'enorme effetto del rumore sulla validità delle previsioni cliniche.

Questa rapida panoramica ha evidenziato in che modo il rumore compromette il giudizio clinico. Nei giudizi predittivi gli esperti umani vengono scalzati da semplici formule, modelli di realtà, modelli dei valutatori, o perfino modelli generati casualmente. Questi risultati depongono a favore dell'uso di metodi non soggetti a rumore, come regole e algoritmi, che saranno al centro del prossimo capitolo.

A proposito di giudizi e modelli

«La gente crede di cogliere la complessità e compiere ragionamenti più sottili quando esprime un giudizio, ma la complessità e la sottigliezza vanno quasi sempre sprecate: di solito non superano l'accuratezza dei modelli semplici.»

«A distanza di oltre sessant'anni dalla pubblicazione del libro di Paul Meehl, l'idea che le previsioni meccaniche siano superiori a quelle umane è ancora sconvolgente.»

«I giudizi sono così soggetti al rumore che un modello di un soggetto giudicante non intaccato dal rumore arriva a previsioni più accurate di quelle del soggetto stesso.»

[1] M.C. Yu, N.R. Kuncel, *Pushing the Limits for Judgmental Consistency: Comparing Random Weighting Schemes with Expert Judgments*, in "Personnel Assessment and Decisions", 6(2020), n. 2, pp. 1-10. La correlazione 0,15 raggiunta dagli esperti è la media non ponderata dei tre campioni studiati, per un totale di 847 casi. Lo studio reale differisce per molti aspetti da questa descrizione semplificata.

[2] Per il calcolo di una media ponderata è necessario che tutti i predittori siano misurati in unità comparabili. Il nostro esempio introduttivo soddisfa tale prerequisito, in quanto tutti i punteggi sono stati assegnati su una scala da 0 a 10, ma non sempre è così. I predittori delle prestazioni, per esempio, potrebbero consistere nella valutazione di un colloquio su una scala da 0 a 10, negli anni di esperienza nel settore e nei risultati di una prova valutativa. Il programma di regressione multipla trasforma tutti i predittori in *punteggi standard* prima di sommarli. Un punteggio standard misura la distanza di un'osservazione dalla media della popolazione, adottando la deviazione standard come unità di misura. Se, per esempio, la media della prova valutativa è di 55 e la deviazione standard è 8, un punteggio standard di +1,5 corrisponde a un risultato di 67. Da notare che la standardizzazione di ciascun dato del singolo elimina ogni traccia di errore nella media o nella varianza dei giudizi dei singoli.

[3] Un elemento importante della regressione multipla è che il peso ottimale di ogni predittore dipende dagli altri predittori: se uno è altamente correlato con un altro, non dovrebbe essergli attribuito un peso altrettanto elevato, per evitare una sorta di "doppio conteggio".

[4] R.M. Hogarth, N. Karelaia, *Heuristic and Linear Models of Judgment: Matching Rules and Environments*, in "Psychological Review", 114(2007), n. 3, p. 734.

[5] In questo contesto si ricorre spesso all'orizzonte di ricerca del *lens model of judgment*, su cui si basa questa disamina. Vedi K.R. Hammond, *Probabilistic Functioning and the Clinical Method*, in "Psychological Review", 62(1955), n. 4, pp. 255-262; N. Karelaia, R.M. Hogarth, *Determinants of Linear Judgment: A Meta-Analysis of Lens Model Studies*, in "Psychological Bulletin", 134(2008), n. 3, pp. 404-426.

[6] P.E. Meehl, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, University of Minnesota Press, Minneapolis 1954.

[7] Id., *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, Aronson, Northvale, NJ 1996, prefazione.

[8] "Paul E. Meehl", in E. Lindzey (a cura di), *A History of Psychology in Autobiography*, American Psychological Association, Washington, DC 1989.

[9] Ivi, p. 362.

[10] W.M. Grove *et al.*, *Clinical Versus Mechanical Prediction: A Meta-Analysis*, in "Psychological Assessment", 12(2000), n. 1, pp. 19-30.

[11] W.M. Grove, P.E. Meehl, *Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical-Statistical Controversy*, in "Psychology, Public Policy, and Law", 2(1996), n. 2, pp. 293-323.

[12] L. Goldberg, *Man Versus Model of Man: A Rationale, plus Some Evidence, for a Method of Improving on Clinical Inferences*, in "Psychological Bulletin", 73(1970), n. 6, pp. 422-432.

[13] M. Friedman, L.J. Savage, *The Utility Analysis of Choices Involving Risk*, in "Journal of Political Economy", 56(1948), n. 4, pp. 279-304.

[14] N. Karelaia, R.M. Hogarth, *Determinants of Linear Judgment*, cit., p. 411, tavola 1.

[15] N. Wiggins, E.S. Kohen, *Man Versus Model of Man Revisited: The Forecasting of Graduate School Success*, in "Journal of Personality and Social Psychology", 19(1971), n. 1, pp. 100-106.

[16] N. Karelaia, R.M. Hogarth, *Determinants of Linear Judgment*, cit.

[17] La correzione di un coefficiente di correlazione per l'affidabilità imperfetta del predittore è nota come *correzione per attenuazione*. La formula è: r^{xy} corretto = $r^{xy}/\sqrt{r^{xx}}$, dove r^{xx} è il coefficiente di affidabilità (la proporzione di varianza reale nella varianza osservata del predittore).

[18] M. Yu, N. Kuncel, *Pushing the Limits for Judgmental Consistency*, cit.

[19] Descriveremo in dettaglio i modelli di ponderazione equa e ponderazione casuale nel prossimo capitolo. I pesi si limitano a una gamma di numeri bassi che devono essere di segno positivo.

10

Regole prive di rumore

In anni recenti, l'intelligenza artificiale (IA) e in particolare le tecniche di apprendimento automatico hanno permesso alle macchine di svolgere molti compiti in precedenza ritenuti distintivi degli esseri umani. Gli algoritmi di apprendimento automatico vengono impiegati nel riconoscimento facciale, nella traduzione da una lingua all'altra e nella lettura delle radiografie, e sono in grado di risolvere problemi di calcolo – per esempio generare indicazioni di guida per migliaia di autisti allo stesso tempo – con una velocità e un'accuratezza sorprendenti. Svolgono inoltre difficili compiti predittivi: vi sono algoritmi in grado di prevedere le decisioni della Corte suprema americana, di indicare quali imputati è più probabile che non compaiano in tribunale dopo il rilascio su cauzione e di valutare quali chiamate alle associazioni a tutela dei minori richiedono un intervento urgente da parte di un assistente sociale.

Anche se oggi quando sentiamo parlare di "algoritmi" ci vengono in mente applicazioni di questo tipo, il termine ha un significato più ampio. Un dizionario definisce un algoritmo come «un processo o un insieme di regole seguite in operazioni di calcolo o nella risoluzione di altri problemi, specialmente da parte di un computer». Stando a questa definizione, anche i modelli semplici e le altre forme di giudizio meccanico descritti nel capitolo precedente possono essere classificati come algoritmi.

In effetti, molti approcci meccanici, dalle regole più banali agli algoritmi più sofisticati e impenetrabili, possono surclassare i giudizi umani. Un

motivo fondamentale alla base di questa prestazione superiore, anche se non l'unico, è che tutti gli approcci meccanici sono privi di rumore.

Per esaminare diversi tipi di approcci basati su regole e capire in che modo e in quali condizioni ciascuno di loro possa essere utile, partiamo dai modelli semplici illustrati nel capitolo 9, basati sulla regressione multipla (ovvero modelli di regressione lineare). Partendo da qui, seguiremo due percorsi opposti lungo lo spettro della complessità, prima per arrivare alla massima semplicità, poi per raggiungere una raffinatezza sempre maggiore (figura 11).

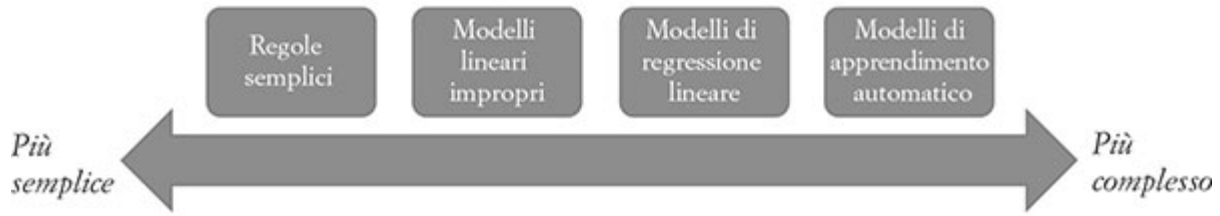


Figura 11. Quattro tipi di regole e algoritmi

Più semplicità: robusto e bello

Robyn Dawes era un altro ricercatore del gruppo di eccellenza di Eugene che negli anni sessanta e settanta si dedicò allo studio del giudizio. Nel 1974 giunse a una svolta nella semplificazione dei compiti predittivi grazie a un'idea stupefacente e quasi eretica: invece di usare la regressione multipla per determinare il peso esatto di ogni predittore, propose di dare a tutti lo stesso peso.

Dawes chiamò questa formula *modello lineare improprio*. A sorpresa, scoprì che questi modelli sono accurati quanto i modelli di regressione "veri e propri", e molto più dei giudizi clinici.[1]

Perfino i sostenitori dei modelli impropri ammettono che tale affermazione è implausibile e «statisticamente controintuitiva».[2] In effetti,

all'inizio Dawes e il suo assistente Bernard Corrigan ebbero difficoltà a pubblicare il loro articolo sulle riviste scientifiche: i redattori non ci credevano. Tornando al caso di Monica e Nathalie del capitolo precedente, probabilmente riterrete che alcuni predittori contino più di altri. Molte persone, per esempio, attribuirebbero alla leadership un peso maggiore rispetto alle competenze tecniche: come può una semplice media non ponderata prevedere le prestazioni di qualcuno meglio di una media attentamente ponderata o del giudizio di un esperto?

Oggi, molti anni dopo la scoperta sensazionale di Dawes, il fenomeno statistico che tanto sorprese i suoi contemporanei è stato pienamente compreso. Come si è detto in precedenza, con la regressione multipla si ottengono pesi "ottimali" che riducono al minimo gli errori quadratici, ma questa tecnica riduce al minimo l'errore *nei dati originari.* La formula, quindi, si adatta in modo da prevedere ogni coincidenza casuale nei dati: se, per esempio, il campione comprende qualche dirigente che ha elevate competenze tecniche e ha anche avuto prestazioni eccezionalmente buone per motivi non correlati a tali abilità, il modello esagererà il peso delle competenze tecniche.

L'obiezione è che quando questa formula viene applicata *fuori dal campione* – cioè quando è impiegata per prevedere un esito in un diverso insieme di dati – quei pesi non saranno più ottimali. Le coincidenze del campione originario, appunto in quanto coincidenze, ora non sono più presenti: nel nuovo campione i dirigenti con alte competenze tecniche non sono tutti fenomenali, e vi saranno altre coincidenze che la formula non è in grado di prevedere. La misura corretta dell'accuratezza predittiva di un modello è la sua prestazione quando viene applicato a un nuovo campione, chiamata *correlazione con convalida incrociata.* In effetti, un modello di regressione ha *troppo* successo nel campione originario, e la correlazione

con convalida incrociata è quasi sempre più bassa di quanto non lo fosse nei dati originari. Dawes e Corrigan hanno paragonato i modelli di ponderazione equa a quelli di regressione multipla (con convalida incrociata) in diverse situazioni. Hanno considerato, per esempio, le previsioni della media dei voti del primo anno di corso di novanta studenti di psicologia della University of Illinois, sulla base di dieci variabili legate al successo accademico: punteggi nei test attitudinali, voti delle superiori, varie valutazioni da parte di loro pari (per esempio riguardo a quanto fossero estroversi) e autovalutazioni (per esempio sul proprio grado di affidabilità). Il modello di regressione multipla standard raggiunse una correlazione di 0,69, che si riduceva a 0,57 (PC = 69%) nella convalida incrociata. La correlazione tra il modello di ponderazione equa e la media dei voti del primo anno era all'incirca la stessa: 0,60 (PC = 70%). Risultati simili sono emersi da molti altri studi.[3]

La perdita di accuratezza nella convalida incrociata è più marcata quando il campione originario è ristretto, perché in quel caso le coincidenze spiccano di più. Il problema evidenziato da Dawes è che i campioni impiegati nella ricerca sociale in genere sono così piccoli da annullare il vantaggio della cosiddetta ponderazione ottimale. Come recita il memorabile sottotitolo di un articolo dello statistico Howard Wainer sulla stima della giusta ponderazione, *It Don't Make No Nevermind* ("Fa lo stesso").[4] O meglio, nelle parole di Dawes, «non ci servono modelli più precisi delle nostre misurazioni».[5] I modelli di ponderazione equa funzionano perché non sono soggetti a problemi di campionamento.

Lo studio di Dawes ha una diretta implicazione su cui vale la pena soffermarsi: è possibile effettuare previsioni statistiche valide senza possedere alcun dato precedente sui risultati che si cerca di prevedere: basta avere un insieme di predittori affidabili da correlare ai risultati.

Poniamo di dover effettuare delle previsioni sulle prestazioni di dirigenti che sono stati valutati sulla base di una serie di dimensioni, come nell'esempio del capitolo 9. Siamo certi che questi punteggi misurino qualità importanti, ma non abbiamo alcun dato sull'accuratezza con cui ogni punteggio riesce a prevedere una determinata performance, né possiamo permetterci il lusso di aspettare anni per avere un riscontro sulle prestazioni di un ampio campione di dirigenti. Potremmo comunque considerare i sette punteggi, effettuare i calcoli statistici necessari per arrivare a un'equa ponderazione e impiegare questi risultati come previsioni. Quanto sarebbe accurato un simile modello di ponderazione equa? La sua correlazione con il risultato sarebbe di 0,25 (PC = 58%), ben superiore alle previsioni cliniche (r = 0,15, PC = 55%), e sicuramente abbastanza simile a un modello di regressione con convalida incrociata, senza richiedere dati che non abbiamo o calcoli complicati.[6]

Per ricorrere a un'espressione di Dawes, divenuta un meme tra chi si occupa dello studio del giudizio, c'è una «bellezza robusta» nella ponderazione equa.[7] La frase conclusiva dell'articolo di grande impatto che lanciò questa idea ne offriva un'altra sintesi efficace: «Il trucco sta nel decidere quali variabili considerare e poi essere in grado di fare la somma».[8]

Ulteriore semplicità: le regole semplici

Un altro stile di semplificazione è quello dei *modelli frugali*, o *regole semplici, ovvero* modelli di realtà simili a calcoli approssimativi ipersemplificati, che però in certi contesti possono condurre a previsioni sorprendentemente buone.

Questi modelli si basano su un tratto della regressione multipla che molti troveranno bizzarro. Poniamo di utilizzare due predittori la cui

correlazione con l'esito sia di 0,60 (PC = 71%) e 0,55 (PC = 69%), dunque un elevato grado di predittività. Ipotizziamo poi che i due predittori siano correlati l'uno all'altro, con una correlazione di 0,50. Secondo voi quanto sarà accurata la vostra previsione una volta combinati in maniera ottimale i due predittori? La risposta è piuttosto deludente: la correlazione è 0,67 (PC = 73%), più alta di prima, ma non di molto.

L'esempio illustra una regola generale: la combinazione di due o più predittori correlati non è molto più predittiva del migliore dei due preso singolarmente. Poiché, nella vita reale, i predittori sono quasi sempre correlati l'uno all'altro, questo dato statistico incoraggia l'uso di approcci predittivi *frugali*, che impiegano un numero ristretto di predittori. In certi contesti, alcune semplici regole applicabili con l'aiuto di pochi calcoli, o anche senza, hanno condotto a previsioni decisamente accurate, se paragonate a modelli più complessi.

Nel 2020 un gruppo di ricercatori ha pubblicato uno studio su larga scala basato sull'applicazione di un approccio frugale a vari problemi predittivi, compresa la scelta da parte di un giudice di rilasciare su cauzione o trattenere un imputato in attesa di processo.[9] Questa decisione implica una previsione sul comportamento dell'imputato: se si sbaglia a negare la libertà provvisoria, questa persona verrà detenuta inutilmente, con un costo notevole per l'individuo e per la società; se invece si concede erroneamente la libertà provvisoria all'imputato sbagliato, costui potrebbe fuggire prima del processo o anche commettere un altro reato.

Il modello elaborato dai ricercatori impiega solo due input noti per essere altamente predittivi rispetto alla probabilità che un imputato non compaia in tribunale dopo il rilascio su cauzione: la sua età (più è alta, minore è il rischio di fuga) e il numero di udienze già saltate (chi non è comparso in passato tende a essere recidivo). Il modello traduce questi due

input in un certo numero di punti, che possono essere utilizzati come un punteggio di rischio. Il calcolo del rischio per un imputato non richiede un computer, e a dire il vero neanche una calcolatrice.

Quando è stato sperimentato su un set di dati reali, il modello frugale ha dato un risultato analogo a quello dei modelli statistici che impiegavano un numero molto più ampio di variabili, e ha prodotto risultati migliori di tutte le previsioni del rischio di fuga effettuate dai giudici umani.

Lo stesso approccio frugale, con un massimo di cinque elementi ponderati sulla base di numeri interi bassi (tra –3 e +3), è stato applicato ai compiti più svariati, come la determinazione della gravità di un tumore a partire da dati mammografici, la diagnosi di una cardiopatia e la previsione di un rischio creditizio. In tutti questi compiti la regola frugale ha dato un risultato analogo a quello di modelli di regressione più complessi (anche se, di solito, inferiore a quello ottenuto tramite apprendimento automatico).

In un'altra dimostrazione del potere delle regole semplici, un diverso gruppo di ricercatori studiò un problema giudiziario simile ma distinto: la previsione della recidiva.[10] Impiegando solo due input, riuscirono ad arrivare alla stessa validità di uno strumento preesistente che impiegava 137 variabili per valutare il livello di rischio di un imputato.[11] Questi due predittori (età e numero di condanne precedenti), come era prevedibile, sono strettamente correlati ai due fattori utilizzati nel modello elaborato per la libertà provvisoria, e la loro associazione con il comportamento criminale è ben documentata.[12]

L'attrattiva di queste regole frugali sta nella loro trasparenza e facilità di applicazione. Peraltro, questi vantaggi comportano una perdita relativamente bassa in termini di accuratezza rispetto a modelli più complessi.

Maggiore complessità: verso l'apprendimento automatico

Giunti alla seconda parte del nostro percorso, invertiamo la direzione di marcia lungo lo spettro della complessità. Sarebbe possibile usare molti più predittori, raccogliere molti più dati su ciascuno, individuare schemi relazionali inaccessibili a qualsiasi essere umano e modellarli per giungere a previsioni migliori? Ebbene, questo promette di fare l'intelligenza artificiale.

Per analisi raffinate occorrono set di dati molto ampi,[13] e la loro crescente disponibilità è uno dei principali fattori alla base del rapido progresso dell'intelligenza artificiale in anni recenti. Questi grandi set di dati, per esempio, rendono possibile gestire in forma meccanica le cosiddette "eccezioni della gamba rotta", una denominazione piuttosto criptica che si riferisce a un esempio immaginato da Meehl. Consideriamo un modello elaborato per prevedere la probabilità che una certa sera un individuo vada al cinema: a prescindere da quanto si ritenga affidabile questo modello, se si viene a sapere che una determinata persona si è appena rotta una gamba, probabilmente si sarà in grado di immaginare meglio del modello come quella persona passerà la serata.

Quando si utilizza un modello semplice, il principio della gamba rotta è di grande utilità per chi è chiamato a decidere: indica quando è meglio mettere da parte il modello e quando no. Se si è in possesso di informazioni determinanti a cui il modello non potrebbe attingere, è giusto applicare il principio della gamba rotta e quindi scavalcare le sue indicazioni. D'altro canto, talvolta si è in disaccordo con le indicazioni del modello anche senza disporre di informazioni private, e, in quei casi, la tentazione di scavalcarle riflette l'applicazione di uno schema personale agli stessi predittori. Poiché è molto probabile che questo schema personale sia privo di valore, sarebbe

meglio trattenersi dall'agire di testa propria, perché probabilmente scavalcare il modello porterà a una previsione meno accurata.

Uno dei motivi alla base del successo dei modelli di apprendimento automatico nei compiti di previsione è la loro capacità di scoprire queste gambe rotte – che sono molte di più di quanto potremmo aspettarci. In presenza di una grande massa di dati su un vasto numero di casi, un modello che monitora il comportamento degli spettatori cinematografici potrebbe apprendere, per esempio, che è improbabile che chi è stato in ospedale il giorno in cui di solito va al cinema quella sera vedrà un film. Un simile miglioramento nelle previsioni di eventi rari riduce la necessità di una supervisione umana.

L'intelligenza artificiale non ha niente a che fare né con la magia, né con la comprensione: si basa unicamente sulla ricerca di uno schema. Se non possiamo che ammirare la forza dell'apprendimento automatico, dovremmo però ricordare che probabilmente ci vorrà del tempo prima che un'intelligenza artificiale capisca *perché* una persona che si è rotta una gamba quella sera non andrà al cinema.

Un esempio: decisioni migliori sulla libertà provvisoria

Più o meno nello stesso periodo in cui il gruppo di ricercatori appena menzionato applicava le regole semplici al problema delle decisioni sulla libertà provvisoria, un altro gruppo guidato da Sendhil Mullainathan elaborò dei sofisticati modelli di intelligenza artificiale per eseguire il medesimo compito.[14] Il gruppo aveva accesso a un set di dati ancora più grande, ovvero 758 027 decisioni sulla libertà provvisoria. Per ogni caso il gruppo aveva accesso alle stesse informazioni di cui era in possesso il giudice: il reato attuale, i precedenti penali e le mancate comparizioni

precedenti dell'imputato; esclusa l'età, nell'elaborazione dell'algoritmo non venne impiegata nessun'altra informazione demografica. I ricercatori sapevano anche, per ciascuno degli imputati, se era stato rilasciato, e in tal caso se non si era presentato in giudizio o era stato arrestato di nuovo. (Era stato rilasciato il 74% degli imputati, il 15% dei quali non era ricomparso in giudizio, mentre il 26% era stato arrestato.) A partire da questi dati i ricercatori elaborarono un algoritmo di apprendimento automatico, di cui valutarono poi le prestazioni.[15] Poiché il modello era stato costruito attraverso l'apprendimento automatico, non si limitava alle combinazioni lineari: se individuava una regolarità più complessa nei dati, poteva impiegare quello schema per migliorare le sue previsioni.

Il modello fu impostato in modo da arrivare a una previsione del rischio di fuga quantificabile in un punteggio numerico, piuttosto che in una decisione a favore o contro la libertà provvisoria. Questo approccio riconosce che la soglia massima di rischio accettabile, cioè il livello di rischio sopra il quale andrebbe negata la libertà provvisoria a un imputato, richiede un giudizio valutativo che un modello non è in grado di esprimere. Tuttavia, i ricercatori calcolarono che, a prescindere dalla soglia di rischio stabilita, l'impiego del punteggio predittivo del loro modello avrebbe portato a un miglioramento rispetto alle prestazioni dei giudici. Se la soglia fosse stata stabilita in modo che il numero di persone a cui veniva negata la libertà provvisoria restasse invariato rispetto a quello emerso dalle decisioni dei giudici, stando ai calcoli del gruppo di ricerca di Mullainathan il tasso di criminalità si sarebbe ridotto fino al 24%, perché dietro le sbarre sarebbero finiti gli imputati con più elevata probabilità di recidiva. Per contro, se la soglia di rischio fosse stata stabilita in modo da ridurre il più possibile il numero di persone a cui veniva negata la libertà provvisoria senza aumentare i reati, i ricercatori calcolarono che il numero di detenuti

si sarebbe ridotto fino al 42%. In sostanza, il modello di apprendimento automatico dava risultati decisamente migliori di quelli dei giudici quando si trattava di prevedere quali fossero gli imputati ad alto rischio, ed era inoltre molto più efficace dei modelli lineari che impiegavano le stesse informazioni. Il motivo è affascinante: «L'algoritmo di apprendimento automatico trova segnali significativi in combinazioni di variabili che altrimenti potrebbero passare inosservate».[16] Questa capacità di trovare schemi che possono facilmente sfuggire ad altri metodi è particolarmente marcata nel caso degli imputati che l'algoritmo definisce ad altissimo rischio. In altre parole, alcune configurazioni dei dati, per quanto rare, riescono efficacemente a predire un rischio elevato. La scoperta che gli algoritmi riescono a individuare configurazioni rare ma determinanti ci riporta al principio della gamba rotta.

I ricercatori impiegarono inoltre l'algoritmo per costruire un modello di ciascun giudice analogo a quello descritto nel capitolo 9 (ma non limitato a semplici combinazioni lineari). L'applicazione di tali modelli all'intero set di dati permise al gruppo di simulare le decisioni che i giudici avrebbero preso se avessero gestito gli stessi casi, e di confrontarle. I risultati indicarono una presenza notevole di rumore sistemico nelle decisioni sulla libertà provvisoria. In parte si tratta di rumore di livello: se si classificano i giudici sulla base della loro indulgenza, il quintile più indulgente (ovvero il 20% dei giudici con il più alto tasso di rilascio) ha rilasciato l'83% degli imputati, il quintile meno indulgente solo il 61%. I giudici hanno inoltre diversi schemi di giudizio per decidere quali imputati siano più a rischio di fuga: un imputato che per un giudice è a basso rischio può essere considerato ad alto rischio da un altro che, in generale, non è più severo. Da questi risultati emerge chiaramente il rumore strutturale. Un esame più dettagliato ha messo in luce che alle differenze tra i casi era attribuibile il

67% della varianza, al rumore sistemico il restante 33%. Il rumore sistemico comprendeva anche una certa quantità di rumore di livello, ovvero differenze nella severità media, ma per la maggior parte (79%) si trattava di rumore strutturale.[17]

Infine, la maggiore accuratezza del programma di apprendimento automatico non intacca altri obiettivi che i giudici potrebbero aver perseguito, in particolare l'equità razziale. In teoria, benché l'algoritmo non si serva di dati legati alla razza, il programma potrebbe inavvertitamente aggravare le disparità che potrebbero sorgere qualora il modello impiegasse predittori altamente correlati con la razza (come il codice postale), o la fonte dei dati su cui viene elaborato l'algoritmo fosse già affetta da bias. Se, per esempio, venisse impiegato come predittore il numero di arresti precedenti, e tali arresti fossero stati viziati dalla discriminazione razziale, anche l'algoritmo finale sarebbe discriminatorio.

Questo tipo di discriminazione, in linea di principio, è certamente un rischio, ma le decisioni di questo algoritmo sono, per molti versi, meno soggette a distorsioni di tipo razziale rispetto a quelle dei giudici. Se, per esempio, la soglia di rischio viene fissata in modo da raggiungere lo stesso tasso di criminalità cui pervengono le decisioni dei giudici, l'algoritmo manderà in galera il 41% in meno di persone di colore. Risultati simili si ottengono anche in altri scenari: non è detto che una maggiore accuratezza debba esacerbare le disparità razziali, anzi, come ha dimostrato il gruppo di ricerca, l'algoritmo può facilmente ridurle.

Un altro studio svolto in un campo diverso illustra come gli algoritmi possano arrivare al contempo a una maggiore accuratezza e a una minore discriminazione. Il professor Bo Cowgill della Columbia Business School ha effettuato uno studio sul reclutamento degli ingegneri informatici di una grande società high-tech.[18] Invece di avvalersi di valutatori (umani) per

scremare i curricula prima dei colloqui, Cowgill ha sviluppato a questo scopo un algoritmo di apprendimento automatico che ha poi "addestrato" con oltre trecentomila candidature ricevute e valutate dalla società. I candidati selezionati tramite l'algoritmo avevano il 14% di probabilità in più di ricevere un'offerta di lavoro dopo il colloquio rispetto a quelli scelti dal personale, e il 18% di probabilità in più di accettare tale offerta. L'algoritmo, inoltre, individuava un gruppo di candidati più vario in termini di razza, genere e altri parametri: era molto più probabile che venissero selezionati candidati "non tradizionali", che per esempio non provenivano da università d'élite e non avevano esperienze lavorative pregresse né referenze. Il personale tendeva a favorire i curricula che rispondevano al profilo "tipico" di un ingegnere informatico, mentre l'algoritmo dava a ciascun predittore pertinente il giusto peso.

Sia chiaro, questi esempi non dimostrano che gli algoritmi sono sempre imparziali, obiettivi e non discriminanti. Ne è un classico esempio un algoritmo che dovrebbe prevedere quali candidati otterranno un lavoro, ma è stato programmato per prendere in considerazione unicamente le promozioni ottenute dai candidati nei lavori precedenti. Naturalmente un simile algoritmo replicherà tutte le distorsioni umane presenti in quelle decisioni.

Costruire un algoritmo che perpetui disparità di razza o di genere è possibile, forse anche troppo facile, e si attestano molti casi di questo tipo; la visibilità data a tali casi spiega le crescenti preoccupazioni per le distorsioni insite nel processo decisionale degli algoritmi. Prima di trarre conclusioni generali, tuttavia, dovremmo ricordare che alcuni algoritmi non solo sono più accurati dei valutatori umani, ma anche più equi.

Perché usiamo le regole così raramente?

Per fare il punto su questa nostra breve disamina del processo decisionale meccanico, riprendiamo due delle ragioni della superiorità di regole di qualsiasi tipo rispetto al giudizio umano. Primo: come descritto nel capitolo 9, tutte le tecniche predittive meccaniche, non soltanto le più recenti e sofisticate, rappresentano un notevole miglioramento rispetto al giudizio umano. Dato il peso che la combinazione di schemi personali e rumore occasionale ha sulla qualità del giudizio umano, semplicità e assenza di rumore costituiscono due vantaggi considerevoli. Regole semplici e minimamente sensate in genere portano a risultati migliori.

Secondo: talvolta si hanno dati abbastanza variegati da consentire a tecniche sofisticate di intelligenza artificiale di individuare schemi validi e superare notevolmente la forza predittiva di un modello semplice. Quando l'intelligenza artificiale arriva a tanto, il vantaggio di tali modelli sul giudizio umano non risiede solo nell'assenza di rumore, ma anche nella capacità di sfruttare molte più informazioni.

Visti i vantaggi e le evidenze scientifiche in loro favore, ci si chiede perché gli algoritmi non vengano usati di più per esprimere giudizi professionali simili a quelli trattati in questo libro. Malgrado il gran parlare di algoritmi e apprendimento automatico, se si escludono importanti eccezioni in particolari ambiti il loro uso è ancora limitato. Molti esperti ignorano il dibattito che contrappone giudizi clinici e meccanici, preferendo fidarsi dei propri: hanno fiducia nelle proprie intuizioni, e dubbi sul fatto che le macchine possano fare di meglio. Ritengono disumanizzante l'idea di adottare algoritmi nel processo decisionale, come fosse una rinuncia alle proprie responsabilità.

Il ricorso agli algoritmi nelle diagnosi mediche, per esempio, non è entrato nella prassi, malgrado i grandi progressi in tal senso; poche organizzazioni impiegano algoritmi per prendere decisioni in merito ad

assunzioni e promozioni; i produttori di Hollywood danno il via libera a un film sulla base del proprio giudizio e della propria esperienza, non di una formula, e lo stesso vale per gli editori. E se gli Oakland Athletics, la squadra di baseball americana ossessionata dalle statistiche al centro del bestseller di Michael Lewis *Moneyball* (da cui è stato tratto il film *L'arte di vincere*), hanno fatto scalpore, è proprio perché il rigore degli algoritmi è sempre stato l'eccezione più che la regola nel processo decisionale degli sportivi. Ancora oggi gli allenatori, i commissari tecnici e i loro collaboratori spesso agiscono d'istinto e insistono che l'analisi statistica non potrà mai sostituirsi a un buon giudizio.

In un articolo del 1996, Meehl e un suo collega elencarono (e confutarono) ben diciassette categorie di obiezioni ai giudizi meccanici formulate da psichiatri, medici, giudici e altri professionisti.[19] Gli autori conclusero che la resistenza dei clinici fosse dovuta a una concomitanza di fattori sociopsicologici, compresi la «paura della disoccupazione tecnologica», il «basso livello di istruzione» e una «generale avversione per i computer».

Da allora i ricercatori hanno identificato ulteriori fattori che contribuiscono a tale resistenza. Non intendiamo illustrare nel dettaglio i risultati dei loro studi: in questo libro ci proponiamo di dare suggerimenti per migliorare i giudizi umani, non per promuovere la «sostituzione delle persone con le macchine», come avrebbe detto il giudice Frankel.

Ma alcuni dati sulle motivazioni alla base delle resistenze nei confronti della previsione meccanica sono inerenti al nostro discorso sul giudizio umano. Da ricerche recenti è emerso un aspetto rilevante: la gente non è sistematicamente sospettosa nei confronti degli algoritmi; quando le si dà la possibilità di chiedere un parere a un essere umano o a un algoritmo, per esempio, spesso sceglie l'algoritmo.[20] La resistenza agli algoritmi, o

avversione agli algoritmi, non sempre si manifesta sotto forma di rifiuto generalizzato verso l'adozione di nuovi strumenti di supporto decisionale. Spesso si è disposti a dare una possibilità all'algoritmo, ma al primo errore si perde ogni fiducia.[21]

Da una parte questa reazione sembra sensata: perché perdere tempo con un algoritmo inaffidabile? Come esseri umani siamo profondamente consapevoli di sbagliare, ma non siamo pronti a concedere ad altri questo privilegio. Dalle macchine pretendiamo la perfezione; se deludono le nostre aspettative, le scartiamo.[22]

Proprio in virtù di quest'aspettativa istintiva, tuttavia, le persone tendono a non fidarsi degli algoritmi e continuano a ricorrere ai propri giudizi, anche quando questa scelta porta a risultati evidentemente inferiori. È un atteggiamento molto radicato, ed è difficile che cambi finché non si arriverà a un'accuratezza predittiva prossima alla perfezione.

Fortunatamente, ciò che rende regole e algoritmi così efficaci può essere replicato nel giudizio umano. Non pretendiamo di arrivare all'efficienza dei modelli di intelligenza artificiale, ma possiamo sforzarci di emulare la semplicità e l'assenza di rumore dei modelli semplici. Adottando metodi in grado di ridurre di una certa misura il rumore sistemico, dovremmo riscontrare dei miglioramenti nella qualità dei nostri giudizi predittivi. Parleremo di come arrivare a giudizi migliori nella parte 5.

A proposito di regole e algoritmi

> *«Quando si dispone di molti dati, gli algoritmi di apprendimento automatico daranno risultati migliori sia rispetto agli esseri umani sia ai modelli semplici, ma anche le regole e gli algoritmi più semplici presentano grandi vantaggi rispetto ai giudizi umani: non sono soggetti a rumore, e non cercano di applicare ai predittori intuizioni complesse e spesso inefficaci.»*

«Dal momento che non abbiamo dati sui risultati delle nostre previsioni, perché non adottare un modello di ponderazione equa? I risultati saranno validi quasi come quelli di un vero modello, e sicuramente migliori di quelli di un giudizio umano formulato caso per caso.»

«Non siete d'accordo con le previsioni del modello, e questo lo capisco. Ma si tratta dell'eccezione della gamba rotta, o quella previsione non vi piace punto e basta?»

«L'algoritmo può sbagliare, naturalmente. Ma se gli esseri umani sbagliano di più, di chi dovremmo fidarci?»

[1] R.M. Dawes, B. Corrigan, *Linear Models in Decision Making*, in "Psychological Bulletin", 81(1974), n. 2, pp. 95-106. Dawes e Corrigan hanno inoltre proposto l'uso della ponderazione casuale. Lo studio delle previsioni sulle prestazioni dei dirigenti descritto nel capitolo 9 è un'applicazione di questo concetto.

[2] J. Dana, *What Makes Improper Linear Models Tick?*, in J.I. Krueger (a cura di), *Rationality and Social Responsibility: Essays in Honor of Robyn M. Dawes*, Psychology Press, New York 2008, pp. 71-89. La definizione è a pagina 73.

[3] J. Dana, R.M. Dawes, *The Superiority of Simple Alternatives to Regression for Social Sciences Prediction*, in "Journal of Educational and Behavior Statistics", 29(2004), pp. 317-331; J. Dana, *What Makes Improper Linear Models Tick?*, cit.

[4] H. Wainer, *Estimating Coefficients in Linear Models: It Don't Make No Nevermind*, in "Psychological Bulletin", 83(1976), n. 2, pp. 213-217.

[5] J. Dana, *What Makes Improper Linear Models Tick?*, cit., p. 72.

[6] M.C. Yu, N.R. Kuncel, *Pushing the Limits for Judgmental Consistency: Comparing Random Weighting Schemes with Expert Judgments*, in "Personnel Assessment and Decisions", 6(2020), n. 2, pp. 1-10. Come nel capitolo precedente, la correlazione qui riportata è la media non ponderata dei tre campioni studiati. Il confronto regge in tutti e tre: la validità del giudizio clinico degli esperti era di 0,17, 0,16 e 0,13, mentre la validità dei modelli di ponderazione equa era rispettivamente di 0,19, 0,33 e 0,22.

[7] R.M. Dawes, *The Robust Beauty of Improper Linear Models in Decision Making*, in "American Psychologist", 34(1979), n. 7, pp. 571-582.

[8] R.M. Dawes, B. Corrigan, *Linear Models in Decision Making*, cit., p. 105.

[9] J. Jung *et al.*, *Simple Rules to Guide Expert Classifications*, in "Journal of the Royal Statistical Society, Statistics in Society", 183(2020), pp. 771-800.

[10] J. Dressel, H. Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, in "Science Advances", 4(2018), n. 1, pp. 1-6.

[11] Questi sono due esempi di modelli lineari basati su un insieme di variabili molto ristretto (e, nel caso del modello elaborato per la cauzione, su un'approssimazione dei pesi lineari ottenuta mediante un metodo di arrotondamento che trasforma il modello in un calcolo approssimativo). Un altro tipo di "modello improprio" è la *regola della variabile singola*, che considera soltanto un

predittore e ignora tutti gli altri. Vedi P.M. Todd, G. Gigerenzer, *Précis of Simple Heuristics That Make Us Smart*, in "Behavioral and Brain Sciences", 23(2000), n. 5, pp. 727-741.

[12] P. Gendreau, T. Little, C. Goggin, *A Meta-Analysis of the Predictors of Adult Offender Recidivism: What Works!*, in "Criminology", 34(1996).

[13] L'ampiezza in questo contesto è da intendersi come il rapporto tra il numero delle osservazioni e i predittori. In *Robust Beauty*, cit., Dawes indica che deve essere almeno di 15 o 20 a 1 perché nella convalida incrociata la ponderazione ottimale dia risultati migliori dei pesi unitari. In *Superiority of Simple Alternatives*, cit., Dana e Dawes, sulla base di un numero molto più elevato di casi di studio, parlano di un rapporto di 100 a 1.

[14] J. Kleinberg *et al.*, *Human Decisions and Machine Predictions*, in "Quarterly Journal of Economics", 133(2018), pp. 237-293.

[15] L'algoritmo venne addestrato con un sottoinsieme di dati utilizzati per fargli fare pratica, e poi valutato sulla base della sua capacità di prevedere i risultati di un diverso sottoinsieme selezionato in modo casuale.

[16] J. Kleinberg *et al.*, *Human Decisions*, cit., p. 16.

[17] G. Stoddard, J. Ludwig, S. Mullainathan, scambio di email con gli autori, giugno-luglio 2020.

[18] B. Cowgill, *Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening*, comunicazione presentata alla Smith Entrepreneurship Research Conference, College Park, MD, 21 aprile 2018.

[19] W.M. Grove, P.E. Meehl, *Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical-Statistical Controversy*, in "Psychology, Public Policy, and Law", 2(1996), n. 2, pp. 293-323.

[20] J.M. Logg, J.A. Minson, D.A. Moore, *Algorithm Appreciation: People Prefer Algorithmic to Human Judgment*, in "Organizational Behavior and Human Decision Processes", 151(2019), pp. 90-103.

[21] B.J. Dietvorst, J.P. Simmons, C. Massey, *Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err*, in "Journal of Experimental Psychology General", 144(2015), pp. 114-126. Vedi anche A. Prahl, L. Van Swol, *Understanding Algorithm Aversion: When Is Advice from Automation Discounted?*, in "Journal of Forecasting", 36(2017), pp. 691-702.

[22] M.T. Dzindolet *et al.*, *The Perceived Utility of Human and Automated Aids in a Visual Detection Task*, in "Human Factors: The Journal of the Human Factors and Ergonomics Society", 44(2002), n. 1, pp. 79-94; K.A. Hoff, M. Bashir, *Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust*, in "Human Factors: The Journal of the Human Factors and Ergonomics Society", 57(2015), n. 3, pp. 407-434; P. Madhavan, D.A. Wiegmann, *Similarities and Differences Between Human-Human and Human-Automation Trust: An Integrative Review*, in "Theoretical Issues in Ergonomics Science", 8(2007), n. 4, pp. 277-301.

11

Ignoranza oggettiva

Spesso ci è capitato di esporre a una platea di dirigenti aziendali il contenuto dei due capitoli precedenti, che inducono a una riflessione sui limiti del giudizio umano. Poiché il messaggio che intendiamo far passare circola ormai da più di mezzo secolo, presumiamo che quasi tutti coloro che per lavoro sono portati a dover prendere decisioni importanti ne siano al corrente. Ciò non toglie che continuino a opporre resistenza.

Alcuni dei dirigenti con cui ci confrontiamo sostengono orgogliosamente di fidarsi più del proprio istinto che di qualsiasi analisi, e molti altri sono meno espliciti ma la pensano allo stesso modo. Le ricerche sui processi decisionali in ambito aziendale mostrano che i dirigenti, soprattutto quelli più anziani ed esperti, ricorrono ampiamente a quello che chiamano, di volta in volta, *intuito*, *istinto* o, semplicemente, *giudizio* (con un'accezione diversa da quella adottata in questo libro).[1]

Insomma, i decisori ascoltano il proprio istinto e se ne compiacciono. La domanda è: cosa dice l'istinto a queste persone che godono del privilegio dell'autorità e di una grande autostima?

In uno studio compilativo sull'intuizione nelle decisioni dei vertici aziendali, questa viene definita «un giudizio su una data linea d'azione che salta alla mente avvolto da una certa aura, convinzione di correttezza o plausibilità, senza però basarsi su motivazioni o giustificazioni chiare e articolate: "sapere" senza sapere perché».[2] Riteniamo che questa

impressione non sia altro che il *segnale interno* del completamento di un giudizio cui si è accennato nel capitolo 4.

Il segnale interno è un autoriconoscimento che ci si sforza (a volte neanche tanto) di ottenere quando si giunge a una conclusione su un certo giudizio. È un'esperienza emotivamente appagante, la piacevole sensazione di aver raggiunto una coerenza tra le informazioni esaminate e il giudizio a cui si è arrivati. Tutti i tasselli sembrano combaciare. (Vedremo poi che spesso questa impressione di coerenza viene rafforzata nascondendo o trascurando le informazioni che non si riescono a integrare.)

Il segnale interno è importante – e fuorviante – perché non viene interpretato come un'impressione, ma come una convinzione: un'esperienza emotiva («Ho l'impressione che sia giusto») si fa passare per fiducia razionale nella validità del proprio giudizio («Non so perché, ma lo so»).

La fiducia, però, non garantisce l'accuratezza, e molte previsioni fiduciose si dimostrano sbagliate. Se il bias e il rumore contribuiscono a generare errori nelle previsioni, ciò non dipende tanto da quanto *sono* validi i giudizi predittivi, ma dal fatto che c'è un limite a quanto *possano esserlo*. In questo capitolo ci concentreremo proprio su questo limite, che definiamo *ignoranza oggettiva*.

Ignoranza oggettiva

Se vi trovate spesso nella posizione di esprimere giudizi predittivi, dovreste riflettere su un punto, a prescindere della vostra attività, che si tratti di selezionare titoli azionari o di prevedere le prestazioni di atleti professionisti. Per semplicità, tornerò sull'esempio del capitolo 9: la

selezione del personale. Immaginate di aver valutato, negli anni, cento candidati. Ora potete avere un riscontro sulla validità delle vostre decisioni, confrontando le vostre previsioni con la valutazione oggettiva delle prestazioni successive delle persone selezionate. Scegliendo un paio di candidati a caso, fino a che punto il vostro giudizio *ex ante* e le valutazioni *ex post* coincidono? In altri termini, confrontando una qualsiasi coppia di candidati, qual è la probabilità che quello a cui voi attribuivate un più alto potenziale si sia dimostrato effettivamente il più capace?

Spesso rivolgiamo questa domanda in maniera informale a dei gruppi di dirigenti. Nella maggior parte dei casi le risposte rientrano nella fascia di probabilità del 75-85%, e sospettiamo che siano temperate da una certa dose di modestia e dalla volontà di non apparire presuntuosi. Dalle conversazioni private a tu per tu si ha l'impressione che il vero senso di fiducia nelle proprie decisioni sia ancora più alto.

Poiché ormai avete dimestichezza con la percentuale di coppie concordanti, capirete subito qual è il problema posto da questa valutazione. Una PC dell'80% corrisponde grosso modo a una correlazione di 0,80: è raro che nel mondo reale si arrivi a un simile potere predittivo. Nel campo della selezione del personale, un recente studio riporta che le prestazioni dei valutatori sono ben lontane da questo valore: in media raggiungono una correlazione predittiva di 0,28 (PC = 59%).[3]

Questi risultati deludenti non sorprendono, se teniamo conto delle difficoltà insite nella selezione del personale. Un individuo che oggi avvii una nuova carriera lavorativa si troverà ad affrontare molte sfide e opportunità, e il caso cambierà in vari modi la direzione del suo percorso: forse incontrerà un supervisore che crederà in lui o in lei, creerà nuove opportunità, promuoverà il suo lavoro e stimolerà la sua fiducia in se stesso e la sua motivazione, o forse sarà meno fortunato e, senza averne

colpa, partirà con un fallimento che lo demoralizzerà. Anche alcuni eventi della vita privata potrebbero influire sulle sue prestazioni lavorative. Oggi nessuna di queste circostanze può essere prevista, né da voi né da nessun altro, e neanche dal miglior modello predittivo del mondo. Questa incertezza irrisolvibile si estende a tutto ciò che in questo momento non è possibile sapere sul risultato che state cercando di prevedere.

Peraltro, molto di ciò che in teoria si potrebbe sapere sul candidato non è noto nel momento in cui si esprime un giudizio su di lui. Ai fini del nostro discorso non importa se queste lacune siano dovute alla mancanza di test sufficientemente predittivi, a una rinuncia da parte vostra a ottenere ulteriori informazioni per via dei costi o al fatto che non abbiate svolto le adeguate ricerche: in un modo o nell'altro, disponete di informazioni incomplete.

Tanto l'incertezza irrisolvibile (ciò che non si può sapere) quanto l'incompletezza delle informazioni (ciò che si potrebbe sapere ma non si sa) rendono impossibile arrivare a una perfetta previsione. Queste mancate conoscenze non dipendono da bias o rumore nel vostro giudizio, ma sono caratteristiche oggettive del vostro compito, e una tale ignoranza oggettiva di dati importanti limita gravemente l'accuratezza finale. Ci siamo presi la libertà di sostituire il termine d'uso comune *incertezza* con *ignoranza* per non rischiare di fare confusione tra l'incertezza, legata al mondo e al futuro, e il rumore, ovvero la variabilità di giudizi che dovrebbero essere identici.

La quantità di informazioni possedute (e di conseguenza il grado più o meno elevato di ignoranza oggettiva) varia da caso a caso. I giudizi professionali sono quasi sempre piuttosto buoni: le previsioni dei medici, per esempio, sono eccellenti per tante malattie, e in molte controversie

legali un avvocato vi saprà dire in maniera molto accurata a quale decisione è probabile che perverrà il giudice.

In generale, però, c'è da aspettarsi che chi effettua una previsione sottovaluti la propria ignoranza oggettiva. L'eccessiva sicurezza è uno dei bias cognitivi più documentati.[4] Tocca, in particolare, i giudizi sulla propria capacità di arrivare a previsioni precise, anche a partire da informazioni limitate. Quanto si è detto sul rumore nei giudizi predittivi vale anche per l'ignoranza oggettiva: dove c'è una previsione c'è ignoranza, e più di quanto non si pensi.

L'eccessiva sicurezza dei guru

Un nostro amico, lo psicologo Philip Tetlock, è armato di una forte passione per la verità e uno spiccato senso dell'umorismo. Nel 2005 ha pubblicato *Expert Political Judgment* ("Il giudizio politico degli esperti"), un libro in cui, malgrado il titolo in apparenza neutrale, sferrava un attacco devastante alla capacità degli esperti di giungere a previsioni accurate sugli eventi politici.

Tetlock ha studiato le previsioni politiche, economiche e sociali di quasi trecento esperti – insigni giornalisti, stimati accademici e alti consulenti dei leader nazionali – e si è chiesto se si fossero avverate. La ricerca è durata vent'anni; per scoprire se le previsioni a lungo termine si rivelano esatte ci vuole pazienza.

Tetlock è giunto alla conclusione che le previsioni di questi presunti esperti in merito a eventi politici di grande rilievo sono tutt'altro che brillanti. Il libro è diventato famoso per il suo finale col botto: «L'accuratezza dell'esperto medio è paragonabile a quella di uno scimpanzé a cui venga chiesto di scegliere tra varie opzioni lanciando delle

freccette». Più precisamente, il messaggio del libro era il seguente: gli esperti che si guadagnano da vivere «commentando o offrendo consulenze sui trend politici ed economici» non sono più preparati «di un giornalista o di un lettore attento del "New York Times" quando si tratta di interpretare un particolare scenario in fase di sviluppo».[5] Sicuramente gli esperti erano bravi a raccontare le loro storie: riuscivano ad analizzare la situazione attuale, tratteggiare un quadro convincente della sua evoluzione e confutare con grande sicurezza le obiezioni di chi li contestava negli studi televisivi. Ma sapevano davvero cosa sarebbe accaduto? Non si direbbe.

Tetlock giunse a questa conclusione smontando le loro narrazioni. Per ogni tema chiese agli esperti di assegnare un grado di probabilità a tre esiti possibili: status quo, crescita o riduzione. Lo scimpanzé citato dall'autore "sceglierebbe" uno di questi tre esiti con la stessa probabilità – una su tre – a prescindere dalla realtà dei fatti. Gli esperti di Tetlock non superarono di molto questo infimo standard: in media assegnarono a eventi poi verificatisi probabilità leggermente più alte rispetto ad altri che non si verificarono, ma il punto cruciale di questa indagine si rivelò il loro eccesso di sicurezza riguardo alle proprie previsioni. I guru che vantavano una loro teoria su come girasse il mondo erano i più sicuri e i meno accurati.

I risultati di Tetlock indicano che è praticamente impossibile fare previsioni dettagliate a lungo termine su eventi specifici. Il mondo è il regno del caos, e anche un piccolo evento può avere enormi conseguenze. Consideriamo, per esempio, che nell'istante del concepimento c'era un'identica probabilità che tutte le più grandi figure della storia (ma anche quelle più irrilevanti) nascessero con un genere diverso. Accadono gli eventi più imprevedibili, con conseguenze altrettanto imprevedibili; quindi, più lontano ci si proietta nel futuro, più aumenta l'ignoranza oggettiva. La debolezza del giudizio politico degli esperti non dipende da

una limitazione cognitiva dei previsori, ma dall'irrisolvibile ignoranza oggettiva degli eventi futuri.

Possiamo concludere, quindi, che i guru non andrebbero biasimati per le loro previsioni sul futuro, ma ciò non toglie che vadano comunque criticati per essersi cimentati in un'impresa impossibile, credendo di riuscire nel loro intento.

Qualche anno dopo queste scoperte sconcertanti sulla futilità di gran parte delle previsioni a lungo termine, Tetlock e sua moglie Barbara Mellers studiarono la validità delle previsioni su eventi mondiali a breve termine, cioè di solito a una distanza inferiore a un anno, scoprendo che previsioni del genere sono difficili ma non impossibili, e che alcuni, che Tetlock e Mellers definirono "superprevisori", sono sistematicamente più bravi di quasi tutti gli altri, compresi i professionisti dell'intelligence. Per ricorrere alla terminologia adottata in questo libro, queste nuove conclusioni sono compatibili con l'idea che l'ignoranza oggettiva aumenti man mano che ci si allontana nel futuro. Torneremo a parlare dei superprevisori nel capitolo 21.

Chi giudica ha i suoi limiti, ma i modelli non sono tanto meglio

La prima ricerca di Tetlock dimostrava la generale incapacità delle persone di effettuare valide previsioni politiche a lungo termine. Le sue conclusioni sarebbero state completamente ribaltate se avesse trovato anche una sola persona la cui sfera di cristallo non fosse appannata; un compito si può definire impossibile solo dopo che molti sfidanti credibili ci hanno provato e hanno fallito. Così come si è visto che un aggregato meccanico di informazioni spesso conduce a risultati migliori rispetto al giudizio umano,

l'accuratezza predittiva di regole e algoritmi fornisce una migliore dimostrazione dell'intrinseca prevedibilità o imprevedibilità dei risultati.

Dai capitoli precedenti vi sarete fatti l'idea che gli algoritmi siano di gran lunga superiori ai giudizi predittivi; questa impressione, tuttavia, potrebbe essere fuorviante. I modelli sono sistematicamente migliori degli esseri umani, ma non di molto. Non si danno situazioni in cui a un pessimo risultato umano si contrappone un ottimo risultato dei modelli a partire dalle stesse informazioni.

Nel capitolo 9 abbiamo citato una rassegna di 136 studi che dimostrano come l'aggregazione meccanica sia superiore al giudizio clinico.[6] Se le prove di tale superiorità sono, effettivamente, «imponenti e sistematiche», la differenza non è poi tanta. Novantatré di quegli studi prendevano in considerazione decisioni binarie e misuravano il "tasso di successo" di clinici e formule. Nello studio mediano gli umani hanno indovinato il 68% delle volte, contro il 73% delle formule. Nel piccolo sottoinsieme di trentacinque studi che impiegava il coefficiente di correlazione come misura dell'accuratezza, i clinici arrivavano a una correlazione mediana con il risultato pari a 0,32 (PC = 60%), mentre le formule arrivavano a 0,56 (PC = 69%). Secondo entrambi gli indicatori, le formule danno risultati sistematicamente migliori degli umani, ma colpisce, di nuovo, la limitata validità delle previsioni meccaniche. Le prestazioni dei modelli non consentono di andare oltre una soglia di prevedibilità alquanto bassa.

Come si comporta, invece, l'intelligenza artificiale? Come abbiamo osservato, spesso dà risultati migliori rispetto ai modelli più semplici; tuttavia, in molte applicazioni le sue prestazioni sono tutt'altro che perfette. Consideriamo, per esempio, l'algoritmo applicato alle decisioni sulla libertà provvisoria discusso nel capitolo 10. Abbiamo visto come, lasciando inalterato il numero di persone a cui viene negata la libertà

provvisoria, l'algoritmo è in grado di ridurre il tasso di criminalità del 24%. Si tratta di un miglioramento notevole rispetto alle previsioni dei giudici, ma se l'algoritmo sapesse predire in maniera perfettamente accurata quali imputati reitereranno il reato, il tasso di criminalità si potrebbe ridurre molto di più. Non a caso le previsioni soprannaturali di crimini futuri del protagonista di *Minority Report* rientrano nella sfera della fantascienza: c'è una gran quantità di ignoranza oggettiva nelle previsioni dei comportamenti umani.

In un altro studio, Sendhil Mullainathan e Ziad Obermeyer hanno elaborato un modello delle diagnosi di infarto.[7] Quando i pazienti presentano sintomi di un possibile attacco cardiaco, i medici del pronto soccorso devono decidere se prescrivere ulteriori esami. In linea di principio, andrebbe prescritto un esame solo quando il paziente è ad alto rischio, dal momento che gli esami sono non soltanto costosi, ma anche invasivi e pericolosi. Pertanto, la decisione di un medico di prescrivere o meno un esame richiede una valutazione preliminare del rischio di infarto. Per effettuare tale valutazione, i ricercatori hanno costruito un modello di intelligenza artificiale che impiega oltre duemilaquattrocento variabili ed è basato su un ampio numero di casi (4,4 milioni di visite nell'ambito del programma sanitario pubblico Medicare,[8] per un totale di 1,6 milioni di pazienti). Con una simile quantità di dati, è probabile che il modello si avvicini ai limiti dell'ignoranza oggettiva.

Non sorprenderà che l'accuratezza del modello di intelligenza artificiale sia decisamente superiore a quella dei medici. Per valutarne le prestazioni, consideriamo i pazienti che il modello ha piazzato nel decile di rischio più alto: quando sono stati esaminati, si è riscontrato che il 30% di loro aveva avuto un infarto, mentre solo il 9,3% di quelli che si collocavano nell'area centrale della distribuzione del rischio ne aveva avuto uno. Questo livello di

precisione è ragguardevole, ma non certo perfetto. Possiamo pertanto concludere che le prestazioni dei medici sono limitate dai vincoli dell'ignoranza oggettiva non meno che dall'imperfezione dei propri giudizi.

La negazione dell'ignoranza

L'idea che sia impossibile arrivare a una previsione perfetta forse sembrerà scontata: che il futuro sia imprevedibile non è certo una scoperta. Tuttavia, l'ovvietà di questo dato di fatto è pari soltanto alla regolarità con cui viene ignorato, come dimostrano in maniera sistematica i dati sull'eccesso di fiducia nelle previsioni.

La pervasività dell'eccesso di fiducia getta una nuova luce sulla nostra indagine informale sui decisori che si fidano del proprio istinto. Abbiamo notato che spesso le persone scambiano la fiducia soggettiva in se stesse per un'indicazione di validità predittiva. Dopo aver considerato le informazioni su Nathalie e Monica fornite nel capitolo 9, per esempio, il segnale interno che avete percepito una volta raggiunto un giudizio coerente vi ha resi sicuri del fatto che Nathalie fosse la candidata migliore. Se vi siete fidati della vostra previsione, però, siete incorsi nell'illusione di validità: l'accuratezza a cui potete arrivare con le informazioni di cui siete in possesso è piuttosto bassa.

Chi si ritiene capace di un'accuratezza predittiva inverosimilmente alta non è solo troppo sicuro di sé, né si limita a negare che nel suo giudizio rischi di essere affetto da rumore e bias, né semplicemente si ritiene superiore ai comuni mortali: oltre a tutto questo, crede nella prevedibilità di eventi che sono a tutti gli effetti imprevedibili, negando implicitamente

la realtà dell'incertezza. Per ricorrere alla terminologia impiegata in questo libro, il suo atteggiamento comporta una *negazione dell'ignoranza*.

La negazione dell'ignoranza offre una risposta alla domanda che si sono posti Meehl e i suoi seguaci: perché le loro conclusioni sono cadute nel vuoto, e perché i decisori continuano a fare affidamento sull'intuito? Fidandosi del proprio istinto, chi prende una decisione ascolta il segnale interno e ne riceve una ricompensa di tipo emotivo. Questo segnale interno che indica che si è arrivati al giudizio giusto è la voce della fiducia in se stessi, quel «"sapere" senza sapere perché». Ma una valutazione oggettiva del vero potere predittivo dei giudizi raramente giustificherà un tale livello di fiducia.

Non è facile rinunciare alla ricompensa della certezza intuitiva sul piano emotivo. È significativo che molti leader affermino di essere più inclini a prendere decisioni intuitive in situazioni che percepiscono come altamente incerte.[9] Quando i fatti negano loro quel senso di comprensione e fiducia che desiderano, lo cercano nell'intuito. Più è grande l'ignoranza, più allettante sarà la sua negazione.

In questo fenomeno possiamo trovare la risposta anche a un altro enigma. Di fronte ai dati oggettivi che abbiamo presentato qui, molti leader traggono una conclusione apparentemente paradossale: forse le loro decisioni istintive non saranno perfette, dicono, ma se le alternative più sistematiche sono anch'esse lontane dalla perfezione, allora non vale la pena adottarle. Ricordiamo, per esempio, che la correlazione media tra le previsioni dei valutatori e le prestazioni degli impiegati è 0,28 (PC = 59%), e secondo lo stesso studio, in accordo con quanto detto sinora, le previsioni meccaniche saranno migliori ma non di molto, raggiungendo un'accuratezza predittiva pari a 0,44 (PC = 65%). Un dirigente potrebbe dirsi: chi me lo fa fare?

La risposta è che in decisioni importanti, per esempio la scelta su chi assumere, un tale aumento nella validità della previsione ha un grande valore. Gli stessi dirigenti modificano di continuo i propri comportamenti professionali per inseguire vantaggi assai meno significativi. Sul piano razionale, si rendono conto che il successo non è mai garantito, e che devono puntare a decisioni che aumentino le probabilità di raggiungerlo: nessuno di loro comprerebbe un biglietto della lotteria che ha il 59% di probabilità di essere quello vincente, se allo stesso prezzo potesse acquistarne uno con il 65% di probabilità.

Il problema è che in questo caso il "prezzo" non è lo stesso. Il giudizio intuitivo ha la sua ricompensa, il segnale interno: le persone sono disposte a fidarsi di un algoritmo che raggiunge un altissimo livello di accuratezza perché ne traggono un senso di sicurezza pari o superiore a quello del loro segnale interno,[10] ma rinunciare alla ricompensa emotiva del segnale interno è un prezzo alto da pagare, se l'alternativa è una sorta di procedimento meccanico che non vanta nemmeno un'alta validità.

Questa osservazione ha un risvolto importante in termini di miglioramento del giudizio. Malgrado le evidenze a favore dei metodi predittivi meccanici e algoritmici, e malgrado il calcolo razionale che mostra chiaramente il valore dei miglioramenti incrementali nell'accuratezza predittiva, molti decisori rifiuteranno approcci decisionali che li privino della facoltà di esercitare il proprio intuito. Finché gli algoritmi non raggiungeranno la perfezione – e in molti campi l'ignoranza oggettiva negherà per sempre questa possibilità – il giudizio umano non verrà sostituito. Ed è questo il motivo per cui va migliorato.

A proposito di ignoranza oggettiva

«Dove c'è previsione, c'è ignoranza, e probabilmente più di quanto non si pensi. Siamo sicuri che gli esperti di cui ci fidiamo siano più accurati di uno scimpanzé che lancia freccette a caso?»

«Quando ci si fida dell'istinto a causa di un segnale interno, non perché davvero si abbia qualche informazione specifica, si sta negando la propria ignoranza oggettiva.»

«I modelli sono migliori degli esseri umani, ma non di molto: a giudizi umani mediocri corrispondono per lo più modelli leggermente più efficaci. In ogni caso, è sempre da preferire il metodo più efficace, e i modelli sono più efficaci.»

«Forse non ci abitueremo mai a impiegare un modello per prendere decisioni di questo tipo: ci basta il segnale interno per avere fiducia in noi stessi. Quindi cerchiamo di seguire il miglior processo decisionale possibile.»

[1] E. Dane, M.G. Pratt, *Exploring Intuition and Its Role in Managerial Decision Making*, in "Academy of Management Review", 32(2007), n. 1, pp. 33-54; C. Akinci, E. Sadler-Smith, *Intuition in Management Research: A Historical Review*, in "International Journal of Management Reviews", 14(2012), pp. 104-122; G.P. Hodgkinson *et al.*, *Intuition in Organizations: Implications for Strategic Management*, in "Long Range Planning", 42(2009), pp. 277-297.

[2] G.P. Hodgkinson *et al.*, *Intuition in Organizations*, cit., p. 279.

[3] N. Kuncel *et al.*, *Mechanical Versus Clinical Data Combination in Selection and Admissions Decisions: A Meta-Analysis*, in "Journal of Applied Psychology", 98(2013), n. 6, pp. 1060-1072. Vedi anche il capitolo 24 per un ulteriore approfondimento sulle decisioni legate al personale.

[4] D.A. Moore, *Perfectly Confident: How to Calibrate Your Decisions Wisely*, HarperCollins, New York 2020.

[5] P.E. Tetlock, *Expert Political Judgment: How Good Is It? How Can We Know?*, Princeton University Press, Princeton, NJ 2005, pp. 233, 239.

[6] W.M. Grove *et al.*, *Clinical Versus Mechanical Prediction*, cit., pp. 19-30.

[7] S. Mullainathan, Z. Obermeyer, *Who Is Tested for Heart Attack and Who Should Be: Predicting Patient Risk and Physician Error*, NBER Working Paper 26168, National Bureau of Economic Research, 2019.

[8] Medicare è un programma di assicurazione sanitaria statunitense che copre essenzialmente i pazienti di età superiore ai sessantacinque anni e le persone affette da disabilità, a prescindere dal reddito dichiarato. (*n.d.t.*)

[9] W. Agor, *The Logic of Intuition: How Top Executives Make Important Decisions*, in "Organizational Dynamics", 14(1986), n. 3, pp. 5-18; L.A. Burke, M.K. Miller, *Taking the Mystery Out of Intuitive Decision Making*, in "Academy of Management Perspectives", 13(1999), n. 4, pp. 91-99.

[10] P. Madhavan, D.A. Wiegmann, *Effects of Information Source, Pedigree, and Reliability on Operator Interaction with Decision Support Systems*, in "Human Factors: The Journal of the Human Factors and Ergonomics Society", 49(2007), n. 5.

12

La valle della normalità

Passiamo ora a una questione più ampia: come essere in pace con se stessi in un mondo in cui molti problemi sono di facile risoluzione, ma molti altri sono soggetti a un'ignoranza oggettiva? Dopotutto, di fronte a un'ignoranza oggettiva insanabile, a un certo punto ci rendiamo conto dell'inutilità di qualsiasi previsione sulle vicende umane; ma non è così che normalmente ci appare il mondo. Al contrario, come suggerito nel capitolo precedente, continuiamo imperterriti a esprimere previsioni audaci sul futuro a partire da piccole informazioni utili. In questo capitolo analizzeremo la sensazione erronea, seppur diffusissima, che sia possibile comprendere eventi impossibili da prevedere.

Cosa comporta questa convinzione? Solleveremo il problema in due contesti diversi: i procedimenti impiegati nelle scienze sociali e l'esperienza quotidiana.

Prevedere un percorso di vita

Nel 2020 un gruppo di centododici ricercatori diretto da Sara McLanahan e Matthew Salganik, entrambi professori di sociologia alla Princeton University, ha pubblicato un articolo insolito sulla rivista scientifica "Proceedings of the National Academy of Sciences".[1] I ricercatori intendevano scoprire quanto capiscano davvero gli scienziati sociali di ciò che accade nei percorsi di vita delle famiglie socialmente fragili. In virtù

delle loro competenze, fino a che punto sono in grado di prevedere gli eventi che si verificheranno in una famiglia? E, nello specifico, quale livello di accuratezza possono raggiungere gli esperti nel prevedere eventi famigliari impiegando le informazioni normalmente raccolte e applicate dai sociologi nelle proprie ricerche? Per tornare alla nostra terminologia, lo studio era volto a misurare il livello di ignoranza oggettiva che permane rispetto a questi eventi una volta che i sociologi hanno svolto le loro indagini.

Gli autori si sono basati sullo studio *Fragile Families and Child Wellbeing* ("Famiglie fragili e benessere dei minori"), un'indagine longitudinale su larga scala condotta su bambini seguiti dalla nascita fino ai quindici anni di età. Questa enorme banca dati contiene diverse migliaia di informazioni sulle famiglie di circa cinquemila minori, molti dei quali nati in grandi città americane da coppie non sposate: grado di istruzione e occupazione dei nonni, dettagli sulle condizioni sanitarie di tutti i membri della famiglia, indicatori del loro stato sociale ed economico, risposte a vari questionari e test attitudinali e della personalità, e molto altro ancora. Gli scienziati hanno fatto buon uso di questa straordinaria mole di dati: più di settecentocinquanta articoli scientifici si rifanno a questo studio. Molti partono da informazioni di base sui bambini e le loro famiglie per spiegare esiti personali come il rendimento scolastico o la fedina penale.

Lo studio condotto dal gruppo di ricerca di Princeton si è concentrato sulla prevedibilità di sei esiti osservati al raggiungimento dei quindici anni, come il verificarsi di un recente episodio di sfratto, la media dei voti dell'adolescente e una valutazione generale delle condizioni materiali della famiglia. Gli organizzatori hanno impiegato il cosiddetto *common task method*, cioè hanno lanciato una sfida invitando alcuni gruppi di ricercatori a elaborare previsioni accurate dei sei esiti selezionati, impiegando i dati

disponibili per ciascuna famiglia analizzata nello studio originario. Una sfida come questa, che costituiva una novità assoluta nel mondo delle scienze sociali, è invece comune nell'informatica, dove spesso più gruppi vengono invitati a sfidarsi in compiti come la traduzione automatica di un insieme standard di testi o l'individuazione da parte di un'intelligenza artificiale di un particolare animale in un ampio campionario di fotografie. Il risultato del gruppo vincitore va a costituire lo stato dell'arte della disciplina in quel dato momento, destinato a essere puntualmente superato nella sfida successiva. Nei compiti predittivi delle scienze sociali, in cui non ci si aspettano rapidi miglioramenti, è ragionevole adottare la previsione più accurata raggiunta nella competizione come misura della prevedibilità degli esiti a partire da tali dati, ovvero come misura del livello di ignoranza oggettiva residua.

Questa sfida ha suscitato un notevole interesse tra i ricercatori. La relazione finale ha presentato i risultati del lavoro di centosessanta gruppi iperqualificati, selezionati da un bacino molto più ampio di candidati provenienti da tutto il mondo. Molti degli sfidanti si definivano esperti di analisi dei dati e si avvalevano di strumenti di apprendimento automatico.

Nella prima fase della competizione le squadre partecipanti avevano accesso a tutti i dati di metà del campione totale, tra i quali i sei esiti di cui abbiamo parlato sopra.[2] Hanno impiegato questi "dati di addestramento" per creare degli algoritmi predittivi, che sono stati poi applicati a un campione di conferma di famiglie non incluse nella fase di addestramento. I ricercatori ne hanno misurato l'accuratezza utilizzando l'errore quadratico medio: l'errore di previsione per ciascun caso era il quadrato della differenza tra l'esito reale e la previsione dell'algoritmo.

Quanto erano affidabili i modelli vincenti? I sofisticati algoritmi di apprendimento automatico addestrati su un vasto insieme di dati

effettuarono, come ci si poteva aspettare, previsioni superiori a quelle dei modelli lineari semplici (e quindi, per estensione, ai giudizi umani). Ma il miglioramento dei modelli informatici rispetto a un modello semplicissimo era minimo, e la loro accuratezza predittiva restava purtroppo molto bassa. Nelle previsioni sugli sfratti, il modello migliore arrivava a una correlazione di 0,22 (PC = 57%).[3] Risultati simili vennero raggiunti per altri esiti singoli, come la possibilità che la figura di accudimento primaria del minore fosse stata licenziata o avesse avviato un percorso di formazione professionale, oppure il punteggio assegnato dal bambino nell'autovalutazione della "grinta", un tratto della personalità che associa passione e perseveranza per il raggiungimento di un particolare obiettivo. Per questi fattori la correlazione cadeva tra 0,17 e 0,24 (PC = 55-58%).

Due dei sei esiti di destinazione erano aggregati, quindi molto più prevedibili. Le correlazioni predittive erano di 0,44 (PC = 65%) con la media dei voti del bambino e 0,48 (PC = 66%) con una misura riassuntiva delle difficoltà materiali che aveva vissuto nei dodici mesi precedenti, basata sulle risposte a undici domande, come: «Hai mai sofferto la fame?» e «A casa tua hanno tagliato la linea telefonica?». È risaputo che le misure aggregate sono più predittive e più prevedibili delle misure dei singoli esiti. La conclusione principale che si può trarre da questa sfida è che una grande massa di informazioni predittive non è sufficiente per prevedere singoli eventi nella vita di ognuno, e anche la previsione degli aggregati è piuttosto limitata.

I risultati osservati in questa ricerca si riscontrano in molte delle correlazioni rilevate dagli scienziati sociali. Un'ampia rassegna delle ricerche condotte nel campo della psicologia sociale, che considera venticinquemila studi e otto milioni di soggetti coinvolti nell'arco di un secolo, è giunta alla conclusione che «in genere gli effetti psicosociali

producono un valore di *r* [coefficiente di correlazione] pari a 0,21».[4] Correlazioni molto più alte, come il già menzionato 0,60 tra l'altezza e il numero di scarpe di un adulto, sono comuni nelle misure corporee ma molto rare nelle scienze sociali. Una rassegna di settecentootto studi nell'ambito delle scienze cognitive e comportamentali ha rilevato che solo il 3% delle correlazioni riportate era pari o superiore a 0,50.[5]

Coefficienti di correlazione tanto bassi forse vi sorprenderanno, se siete abituati a leggere di risultati definiti «statisticamente significativi», o anche «altamente significativi». Spesso i termini statistici sono ingannevoli per i lettori comuni: il termine "significativo" ne è l'esempio peggiore. Quando un risultato è definito "significativo", non dobbiamo trarre la conclusione che l'effetto che descrive sia rilevante: vuol dire soltanto che difficilmente sarà un mero frutto del caso. Con un campione sufficientemente ampio, una correlazione può essere allo stesso tempo "significativa" e troppo debole per meritare attenzione.

La prevedibilità limitata dei singoli esiti emersa dalla sfida fa riflettere sulla differenza tra comprensione e previsione. Lo studio sulle famiglie fragili è considerato una miniera d'oro nel campo delle scienze sociali e, come abbiamo visto, i suoi dati sono stati impiegati in numerosissime ricerche. Gli studiosi coinvolti in tali ricerche sicuramente ritenevano che il loro lavoro avrebbe portato a una maggiore conoscenza della vita delle famiglie fragili, ma purtroppo questa percezione ottimistica non ha trovato riscontro in una reale capacità di arrivare a previsioni precise su singoli eventi relativi a singoli individui. Gli autori della relazione conclusiva sulla sfida hanno inserito già nell'abstract un severo ammonimento: «I ricercatori dovranno riconsiderare la propria convinzione che sia possibile comprendere i percorsi di vita delle famiglie,

alla luce del fatto che nessuna delle loro previsioni si è dimostrata molto accurata».[6]

Comprensione e previsione

La logica soggiacente a questa conclusione pessimistica richiede una spiegazione. Quando gli autori della sfida sulle famiglie fragili assimilano la comprensione alla previsione (o l'assenza di una all'assenza dell'altra), usano il termine *comprensione* in un'accezione specifica. Questa parola ha infatti anche altri significati: se diciamo che comprendiamo un concetto matematico, o che comprendiamo cos'è l'amore, probabilmente non ci riferiamo alla capacità di effettuare alcuna specifica previsione.

Tuttavia, nelle scienze sociali, e molto spesso nelle conversazioni di ogni giorno, dire che si comprende qualcosa equivale a dire che se ne comprende la *causa*. I sociologi che hanno raccolto e studiato le migliaia di variabili esaminate nello studio sulle famiglie fragili stavano cercando di risalire alle cause degli esiti osservati. I medici che comprendono cosa affligge un paziente sostengono che la patologia da loro diagnosticata sia la causa dei sintomi osservati. Comprendere, in questa accezione, vuol dire descrivere una catena causale.[7] La capacità di effettuare una previsione indica se e quanto tale catena sia stata effettivamente identificata. Allo stesso modo, la correlazione, ovvero la misura dell'accuratezza predittiva, indica fino a che punto siamo in grado di spiegare una causazione.

Forse vi sorprenderà se conoscete i fondamenti della statistica, e quindi il classico avvertimento secondo cui "correlazione non implica causazione". Prendiamo, per esempio, la correlazione tra il numero di scarpe e le competenze matematiche nei bambini: com'è ovvio, una variabile non causa l'altra. La correlazione deriva dal fatto che nei bambini

sia i piedi sia le competenze matematiche crescono con l'età, il che può condurre a formulare una previsione: sapendo che un bambino ha i piedi grandi, dovremmo prevedere che avrà un livello di competenze matematiche superiore rispetto a un bambino che ha i piedi piccoli. Ma non bisognerebbe dedurre un nesso causale da questa correlazione.

Occorre però ricordare che, se la correlazione non implica la causazione, è vero l'inverso: in presenza di un legame causale, dovremmo trovare una correlazione. Se non ne troviamo una tra età e numero di scarpe negli adulti, possiamo tranquillamente concludere che dopo l'adolescenza i piedi non crescono con l'età, quindi bisognerà cercare altrove la causa delle differenze nella taglia del piede.

In sintesi: dove c'è causazione, c'è correlazione. Ne consegue che dove c'è causazione, dovremmo essere in grado di effettuare una previsione – e la correlazione, cioè la sua accuratezza, è una misura di quanto capiamo questo nesso di causalità. Pertanto, le conclusioni dei ricercatori di Princeton sono le seguenti: la capacità dei sociologi di predire eventi come lo sfratto, pari a una correlazione di 0,22, ci dà un'indicazione di quanto – o meglio quanto poco – capiscano il percorso di vita di queste famiglie. L'ignoranza oggettiva pone un limite non solo alle nostre previsioni, ma anche alla nostra comprensione.

Cosa intendono allora tutti gli esperti che si dicono sicuri di capire il proprio settore? Come possono pronunciarsi sulle cause dei fenomeni che osservano e arrivare a previsioni convinte in tal senso? Insomma, perché gli esperti, come del resto tutti noi, sembrano sottovalutare la propria ignoranza oggettiva del mondo?

Il pensiero causale

Se leggendo la prima parte di questo capitolo vi siete chiesti cosa porti le famiglie fragili allo sfratto e ad altri esiti sul piano personale, avete fatto lo stesso ragionamento dei ricercatori di cui abbiamo parlato. Avete cioè applicato il *pensiero statistico*: vi siete soffermati sull'insieme, vale a dire sulla popolazione delle famiglie fragili, e sugli indicatori statistici impiegati per descriverle, come la media, la varianza, la correlazione eccetera. Non avete considerato i casi singoli.

Un modo diverso di pensare, che adottiamo in maniera più naturale, è quello che qui chiameremo *pensiero causale*.[8] Questo tipo di pensiero crea delle narrazioni in cui eventi, persone e oggetti specifici si influenzano a vicenda. Per capire di che si tratta, mettetevi nei panni di un assistente sociale che segue molte famiglie disagiate. Supponete di avere appena appreso che una di queste famiglie, i Jones, ha subìto uno sfratto: la vostra reazione all'evento è influenzata da ciò che sapete dei Jones. Si dà il caso che Jessica Jones, responsabile del mantenimento della famiglia, abbia perso il lavoro qualche mese fa e non sia riuscita a trovarne un altro; da allora, non è stata in grado di pagare l'affitto per intero. Ha effettuato pagamenti parziali, ha supplicato più volte l'amministratore del condominio di venirle incontro, arrivando anche a chiedervi di aiutarla (ma anche davanti al vostro intervento lui è rimasto inamovibile). Alla luce del contesto, lo sfratto della famiglia Jones, per quanto triste, non ci stupisce. Al contrario, sembra la conclusione logica di una catena di eventi, l'inevitabile epilogo di una tragedia annunciata.

Quando cediamo a questo senso di inevitabilità, dimentichiamo che le cose sarebbero potute benissimo andare in tutt'altro modo, che a ogni bivio della vita il destino avrebbe potuto prendere una strada diversa. Jessica avrebbe potuto continuare a lavorare, o trovare subito un altro lavoro; un parente sarebbe potuto giungere in suo soccorso; voi, in qualità

di assistenti sociali, avreste potuto convincere l'amministratore, il quale a sua volta avrebbe potuto essere più comprensivo e concedere alla famiglia qualche settimana di proroga, in modo che intanto Jessica potesse trovare un lavoro che le permettesse di saldare gli arretrati.

Queste storie alternative sono plausibili quanto la prima, una volta che se ne conosce la fine. Qualunque sia l'esito (che si arrivi o no allo sfratto), una volta che questo si verifica, in virtù del pensiero causale diventa del tutto spiegabile, anzi perfino prevedibile.

La comprensione nella valle della normalità

C'è una spiegazione psicologica a questa osservazione. Alcuni eventi ci sorprendono: una pandemia letale, un attacco alle Torri Gemelle, un eccellente fondo speculativo che si rivela uno schema Ponzi. Anche nel privato possono capitare degli eventi sconcertanti: innamorarsi di un estraneo, perdere un fratello più giovane di punto in bianco, ricevere un'eredità inaspettata; altri eventi invece sono del tutto attesi, come il ritorno da scuola di un bambino all'orario previsto.

Ma le esperienze degli esseri umani cadono per lo più nello spazio intermedio tra questi due estremi. Talvolta ci aspettiamo che un certo evento si verifichi, altre volte arriva di sorpresa, ma quasi tutto accade nella grande valle della normalità, in cui gli eventi non sono né del tutto attesi né particolarmente sorprendenti. In questo momento, per esempio, non avete un'aspettativa precisa su ciò che leggerete nel prossimo paragrafo. Vi sorprenderebbe se all'improvviso ci mettessimo a scrivere in turco, ma ci sono molte cose che potremmo dire senza sconcertarvi.

Nella valle della normalità, gli eventi si svolgono come lo sfratto dei Jones: sembrano normali col senno di poi, anche se non erano attesi e noi

non avremmo potuto prevederli. Questo accade perché il processo di comprensione della realtà è retrospettivo.[9] Un evento non previsto (lo sfratto della famiglia Jones) attiva la ricerca mentale di una possibile causa (la congiuntura economica sfavorevole, l'amministratore inflessibile). Questa ricerca si interrompe una volta trovata una narrazione soddisfacente. Se l'esito fosse stato l'esatto opposto, la ricerca avrebbe condotto a cause altrettanto stringenti (la tenacia di Jessica, l'amministratore comprensivo).

Come illustrano questi esempi, in qualsiasi storia molti eventi si spiegano da soli, nel senso letterale dell'espressione. Forse avrete notato che nelle due versioni del racconto dello sfratto l'amministratore non è la stessa persona: il primo era insensibile, il secondo gentile. Ma in realtà l'unico indizio in nostro possesso sul carattere di questa persona è il comportamento che ha dimostrato. Sulla scorta di ciò che ora sappiamo di lui, il suo comportamento ci appare coerente. È il verificarsi di un evento che ne rivela la causa.

Quando un esito inatteso ma non sorprendente viene spiegato in questi termini, le conclusioni che se ne possono trarre sono sempre sensate. È questo che vuol dire, per tutti noi, comprendere una storia, e che ci fa sembrare la realtà prevedibile, col senno di poi. Poiché l'evento si spiega da sé una volta verificatosi, ci illudiamo che potesse essere previsto.

In generale, la nostra comprensione del mondo dipende dalla straordinaria capacità che abbiamo di costruire storie in grado di spiegare gli eventi che osserviamo. La ricerca delle cause viene quasi sempre soddisfatta, perché queste possono essere tratte da un repertorio illimitato di fatti e di idee del mondo. Come sa chiunque segua il telegiornale della sera, per esempio, sono pochi i grandi movimenti di borsa che non trovano una spiegazione. La stessa notizia potrà "spiegare" sia una caduta (gli

investitori più nervosi sono preoccupati per ciò che hanno saputo!) sia un aumento degli indici (gli investitori più ottimisti non hanno perso la fiducia!).

Quando non riusciamo a trovare una causa lampante, per prima cosa fabbrichiamo una spiegazione colmando un vuoto nel nostro modello del mondo. In questo modo arriviamo a una deduzione che prima ci era ignota (per esempio, che l'amministratore era un uomo gentilissimo). Solo quando non riusciamo a modificare il nostro modello del mondo per arrivare a un certo risultato etichettiamo quest'ultimo come un fatto sorprendente e iniziamo a costruirne un resoconto più elaborato. La vera sorpresa si verifica solo quando un evento non è spiegabile con l'abituale senno di poi.

Questa interpretazione causale della realtà è il nostro modo abituale di "comprendere" il mondo: abbiamo l'impressione di comprendere la vita nel suo svolgimento grazie a questo flusso costante di valutazioni a posteriori nella valle della normalità. È un processo fondamentalmente causale: i nuovi eventi, una volta noti, eliminano ogni alternativa, e la narrazione non lascia spazio all'incertezza. Come sappiamo dagli studi classici sul senno di poi, anche quando vi è una prima incertezza soggettiva, questa viene per lo più dimenticata una volta risolta.[10]

Dentro e fuori

Abbiamo posto a confronto due diversi modi di pensare agli eventi della vita: il pensiero statistico e quello causale. La modalità causale ci fa risparmiare molta energia categorizzando gli eventi sul momento come normali o anormali. Quelli anormali richiedono da subito uno sforzo notevole per trovare informazioni pertinenti, sia nell'ambiente circostante

sia nella nostra memoria. Anche l'aspettativa attiva, cioè il restare in solerte attesa che qualcosa accada, prevede uno sforzo; per contro, il flusso di eventi della valle della normalità richiede poco lavoro mentale. Che il vostro vicino di casa vi sorrida o appaia distratto e si limiti ad annuire incrociandovi per strada, non farete molto caso a nessuno dei due eventi, se in passato è già successo con una certa frequenza. Se il sorriso è insolitamente smagliante o il cenno del capo insolitamente sbrigativo, ne cercherete la causa nella vostra memoria. Il pensiero causale evita gli sforzi superflui, pur mantenendo il grado di allerta necessario per individuare gli eventi anormali.

Al contrario, il pensiero statistico è laborioso, e richiede un'attenzione che soltanto il sistema 2,[11] la modalità di pensiero lenta e intenzionale, può mobilitare. Oltre una certa soglia di base, inoltre, necessita di competenze specialistiche. Questo tipo di pensiero parte dagli insiemi e considera i casi individuali alla stregua di esemplificazioni di categorie più ampie: lo sfratto dei Jones non è inteso come il risultato di una catena di eventi specifici, ma come un esito statisticamente probabile (o improbabile) a partire da osservazioni preliminari di casi che condividono alcune caratteristiche predittive con quello in esame.

Torneremo più volte sulla distinzione tra queste due prospettive. Affidarsi al pensiero causale per un caso singolo induce a errori prevedibili; assumere la prospettiva statistica, che qui chiameremo la *visione esterna*, permette di evitare tali errori.

Per adesso ci interessa soltanto sottolineare che la modalità causale è quella che ci viene più naturale. Perfino spiegazioni che andrebbero considerate più propriamente come statistiche spesso vengono presentate come narrazioni causali. Pensiamo ad affermazioni come: «Hanno fallito perché non avevano abbastanza esperienza», oppure «Hanno avuto

successo perché avevano un leader geniale»; sarebbe facile portare dei controesempi di aziende alle prime armi che hanno avuto successo e leader geniali che hanno fallito. La correlazione tra il successo e l'esperienza o la genialità è al massimo moderata, ma più probabilmente bassa. Eppure si fa presto a trovare un nesso causale. Ogni volta che la causalità è plausibile, la nostra mente dà a una correlazione, per quanto bassa, una forza causale esplicativa: la genialità di un leader viene accettata come spiegazione soddisfacente del successo di un'azienda, l'inesperienza come spiegazione di un fallimento.

Affidarsi a spiegazioni fallaci è forse inevitabile, se l'alternativa è rinunciare a comprendere il mondo. Detto ciò, il pensiero causale e l'illusione di capire il passato inducono ad avanzare previsioni eccessivamente fiduciose. Come vedremo, la preferenza accordata al pensiero causale contribuisce inoltre alla mancata individuazione del rumore come fonte di errore, essendo il rumore un concetto essenzialmente statistico.

Il pensiero causale ci aiuta a dare senso a un mondo che è molto meno prevedibile di quanto non si pensi, e spiega perché lo riteniamo molto più prevedibile di quanto non sia. Nella valle della normalità, non vi sono sorprese o incongruenze: il futuro non sembra meno prevedibile del passato, e il rumore resta invisibile.

A proposito dei limiti della comprensione

«Negli eventi umani, correlazioni intorno allo 0,20 (PC = 56%) sono piuttosto comuni.»

«La correlazione non implica causazione, ma la causazione implica correlazione.»

«Quasi tutti gli eventi quotidiani non sono né attesi né sorprendenti, e non richiedono alcuna spiegazione.»

«Nella valle della normalità, gli eventi non sono né attesi né sorprendenti: semplicemente, si spiegano da soli.»

«Pensiamo di capire quello che succede, ma saremmo riusciti a prevederlo?»

[1] M.J. Salganik *et al.*, *Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration*, in "Proceedings of the National Academy of Sciences", 117(2020), n. 15, pp. 8398-8403.

[2] Il campione totale comprendeva 4242 famiglie, in quanto una parte di quelle incluse nello studio originario è stata esclusa dall'analisi per motivi di privacy.

[3] Per valutare l'accuratezza, gli organizzatori della gara hanno impiegato lo stesso indicatore presentato nella parte 1: l'errore quadratico medio, o MSE. Allo scopo di favorire la comparazione, inoltre, hanno confrontato l'MSE di ogni modello con una strategia di previsione "inutile", ovvero una previsione generica secondo cui ciascun caso singolo non è diverso dalla media dei dati di addestramento. Per comodità, abbiamo convertito i risultati in coefficienti di correlazione. L'MSE e la correlazione vengono messi in rapporto nell'espressione r^2 = (Var (Y) – MSE) / Var (Y), dove Var (Y) è la varianza della variabile del singolo esito e (Var (Y) – MSE) è la varianza degli esiti previsti.

[4] F.D. Richard *et al.*, *One Hundred Years of Social Psychology Quantitatively Described*, in "Review of General Psychology", 7(2003), n. 4, pp. 331-663.

[5] G.E. Gignac, E.T. Szodorai, *Effect Size Guidelines for Individual Differences Researchers*, in "Personality and Individual Differences", 102(2016), pp. 74-78.

[6] Occorre fare una precisazione. Questo studio si propone di impiegare un set di dati descrittivi preesistente che è molto ampio ma non adattato appositamente per prevedere esiti specifici. Questo costituisce una differenza importante rispetto al caso degli esperti dello studio di Tetlock, che erano liberi di avvalersi di qualsiasi informazione ritenessero opportuna. Sarebbe possibile, per esempio, identificare dei predittori dello sfratto non presenti nella banca dati, ma che potrebbero verosimilmente essere raccolti. Pertanto, lo studio non dimostra l'*intrinseca* imprevedibilità dello sfratto e di altri esiti, ma la loro imprevedibilità *sulla base di questo set di dati*, che viene utilizzato da numerosi scienziati sociali.

[7] J.M. Hofman *et al.*, *Prediction and Explanation in Social Systems*, in "Science ", 355(2017), pp. 486-488; D.J. Watts *et al.*, *Explanation, Prediction, and Causality: Three Sides of the Same Coin?*, ottobre 2018, pp. 1-14, disponibile sul sito del Center for Open Science, [osf.io/bgwjc].

[8] Una distinzione simile a questa pone in contrasto il pensiero *estensivo* e quello *non estensivo* o *intenzionale*. A. Tversky, D. Kahneman, *Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment*, in "Psychological Review", 4(1983), pp. 293-315.

[9] D. Kahneman, D.T. Miller, *Norm Theory: Comparing Reality to Its Alternatives*, in "Psychological Review", 93(1986), n. 2, pp. 136-153.

[10] B. Fischhoff, *An Early History of Hindsight Research*, in "Social Cognition", 25(2007), n. 1, pp. 10-13, doi:10.1521/soco.2007.25.1.10; Id., *Hindsight Is Not Equal to Foresight: The Effect of Outcome Knowledge on Judgment Under Uncertainty*, in "Journal of Experimental Psychology: Human Perception and Performance", 1(1975), n. 3, p. 288.

[11] D. Kahneman, *Pensieri lenti e veloci*, Mondadori, Milano 2012.

QUARTA PARTE

Come nasce il rumore

Cosa c'è all'origine del rumore, e anche del bias? Quali sono i meccanismi mentali alla base della variabilità dei nostri giudizi e degli errori comuni che li influenzano? Insomma, cosa sappiamo sulla psicologia del rumore? Qui affronteremo queste questioni.

Innanzitutto descriveremo come alcune delle operazioni proprie del sistema 1, quello dei pensieri veloci, siano responsabili di molti errori di giudizio. Nel capitolo 13 presenteremo tre importanti euristiche di giudizio su cui poggia il sistema 1, e mostreremo come provochino errori direzionali prevedibili (bias statistico), nonché rumore.

Il capitolo 14 si concentrerà sul *matching*, una particolare operazione del sistema 1, analizzando gli errori che ne derivano.

Nel capitolo 15 prenderemo in considerazione uno strumento indispensabile per qualsiasi giudizio: la scala di riferimento che viene impiegata. Mostreremo che la scelta di una scala appropriata è un prerequisito per arrivare a un buon giudizio, e che le scale indefinite o inadeguate sono una grande fonte di rumore.

Il capitolo 16 esplorerà la radice psicologica di quella che è forse la più affascinante tipologia di rumore: gli schemi di risposta di persone diverse a casi diversi. Come le personalità individuali, questi schemi non sono casuali e restano per lo più stabili nel tempo, ma i loro effetti non sono facilmente prevedibili.

Infine, nel capitolo 17, riassumeremo ciò che abbiamo imparato sul rumore e sulle sue componenti. Questa esplorazione ci porterà ad avanzare

una risposta al rebus sollevato in precedenza: perché, pur essendo onnipresente, raramente il rumore è ritenuto un problema serio?

13

Euristiche, bias e rumore

Questo libro si pone nel solco delle ricerche sul giudizio intuitivo umano sviluppate nell'arco di mezzo secolo nel cosiddetto Heuristics and Biases Program ("Programma sulle euristiche e i bias"). I risultati raggiunti nei primi quarant'anni sono illustrati nel libro *Pensieri lenti e veloci*,[1] un'indagine sui meccanismi psicologici volta a spiegare miserie e splendori del pensiero intuitivo. L'idea centrale di questo programma di ricerca era che, di fronte a domande difficili, la gente ricorre a scorciatoie semplificatrici, chiamate *euristiche*. In generale le euristiche, prodotte da un pensiero veloce e intuitivo, che definiamo "sistema 1", sono piuttosto utili e pervengono a risposte adeguate, ma talvolta conducono a sviluppare dei bias, che qui abbiamo descritto come errori di giudizio sistematici e prevedibili.

Il programma di ricerca sulle euristiche e i bias si è concentrato su ciò che accomuna le persone, più che sulle loro differenze, per mostrare come i processi che producono errori di giudizio siano comuni ai più. Anche per questo, chi è avvezzo al concetto di bias psicologico spesso presuppone che sfoci sempre in un *bias statistico*, termine usato in questo libro per riferirci alle misure o ai giudizi che quasi sempre deviano dalla verità in una certa direzione. Certo, i bias psicologici creano un bias statistico quando sono ampiamente diffusi, ma quando invece chi giudica è affetto da bias di diverso tipo, o in misura diversa, sviluppano rumore sistemico. In ogni caso, che causino bias statistico o rumore, i bias psicologici inducono sempre in errore.

Diagnosticare i bias

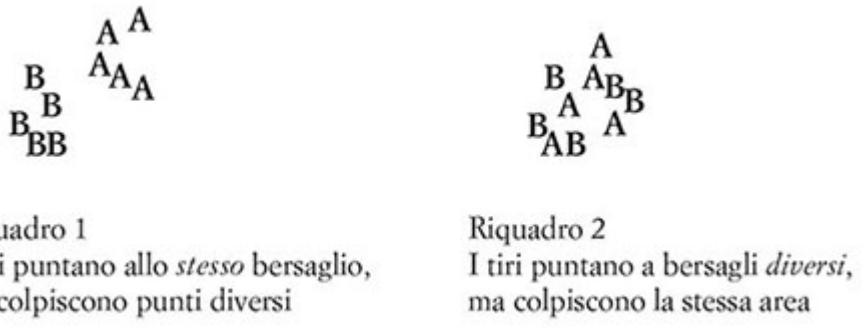
I bias di giudizio vengono spesso identificati in riferimento a un valore reale. Nei giudizi predittivi vi è un bias se gli errori vanno tendenzialmente in una determinata direzione piuttosto che in un'altra: per esempio, quando viene chiesto ad alcune persone di prevedere quanto tempo impiegheranno per completare un progetto, la media delle loro stime di solito è molto più bassa rispetto al tempo realmente necessario. Questo noto bias psicologico è conosciuto come *fallacia della pianificazione*.

Spesso, però, non vi è un valore reale con cui poter confrontare i giudizi. Vista la nostra insistenza sul fatto che si possa individuare il bias statistico solo quando si conosce il valore reale, forse vi starete chiedendo come si possano studiare i bias psicologici quando non si conosce la verità. La risposta è che i ricercatori confermano la presenza di un bias psicologico nel caso in cui un fattore che non dovrebbe influenzare il giudizio ha un effetto statisticamente rilevante sullo stesso, oppure un fattore che dovrebbe influenzarlo non lo fa.

Per illustrare questo metodo torniamo all'analogia del tiro a segno. Immaginiamo che le squadre A e B abbiano sparato i loro colpi, ora visibili sul retro del bersaglio (figura 12). In questo esempio non sappiamo dov'è il centro (il valore reale non è noto), pertanto non conosciamo la deviazione delle due squadre rispetto a quest'ultimo. Tuttavia, ci viene detto che, nel riquadro 1, le due squadre puntavano allo stesso obiettivo, mentre, nel riquadro 2, la squadra A puntava a un obiettivo e la squadra B a un altro.

Malgrado l'assenza di un bersaglio, in entrambi i riquadri è evidente una deviazione sistematica (o bias). Nel primo, i tiri delle due squadre differiscono tra loro, quando dovrebbero essere identici; questa configurazione è simile a quella che vedremmo in un esperimento in cui due gruppi di investitori leggessero due piani aziendali sostanzialmente

identici ma stampati in un carattere diverso su un diverso tipo di carta: se questi dettagli irrilevanti fanno la differenza nel giudizio degli investitori, si è in presenza di bias psicologico. Non sappiamo se gli investitori sedotti dal carattere elegante e dalla carta patinata siano troppo ottimisti o se quelli che hanno letto la versione più grezza siano troppo pessimisti, ma sappiamo che hanno un giudizio diverso, anche se non dovrebbero.



Riquadro 1
I tiri puntano allo *stesso* bersaglio, ma colpiscono punti diversi

Riquadro 2
I tiri puntano a bersagli *diversi*, ma colpiscono la stessa area

Figura 12. Uno sguardo al retro dei bersagli in un esperimento sul bias

Il riquadro 2 illustra il fenomeno opposto. Poiché le squadre puntavano a bersagli diversi, i cluster dei tiri dovrebbero essere distinti, mentre si concentrano nella stessa area. Immaginate, per esempio, che a due gruppi di persone venga posto il quesito su Michael Gambardi del capitolo 4, ma con una piccola variazione: a un gruppo viene chiesto, come è stato chiesto a voi, di stimare la probabilità che tra due anni Gambardi ricopra lo stesso ruolo aziendale; all'altro viene fatta la stessa richiesta, ma allungando l'arco temporale a tre anni. I due gruppi dovrebbero giungere a conclusioni diverse, perché ovviamente ci sono più possibilità di perdere il lavoro in tre anni che in due. Tuttavia, i dati indicano che le stime della probabilità nei due gruppi varieranno poco o niente. È chiaro che le risposte dovrebbero essere diverse, ma non lo sono, a indicare che uno dei fattori che dovrebbe influenzare il giudizio non è stato preso in considerazione. (Questo bias

psicologico viene chiamato tecnicamente *insensibilità alle dimensioni del campione*.)[2]

Gli errori sistematici di giudizio sono stati dimostrati in molti campi, e il termine *bias* oggi è adottato in vari ambiti, compresi quello aziendale, politico, legislativo e giuridico. Trattandosi di una parola di uso comune, ha assunto un significato molto ampio. Oltre alla definizione cognitiva qui impiegata (in riferimento a un meccanismo psicologico e all'errore che questo tipicamente produce), spesso questa parola viene usata per indicare che qualcuno ha un pregiudizio nei confronti di un certo gruppo (per esempio il bias di gender o il bias razziale). Può anche significare che qualcuno cerca di favorire un certo risultato, come chi è sviato da un conflitto di interessi o da una certa opinione politica. Abbiamo inserito questi bias nel nostro discorso sulla psicologia degli errori di giudizio perché tutti i bias psicologici producono sia bias statistico sia rumore.

Vi è invece un uso del termine a cui ci opponiamo con forza, ovvero quando si attribuiscono dei clamorosi fallimenti a non specificati "bias", e il riconoscimento dell'errore si accompagna alla promessa di un grande impegno per eliminarli nel processo decisionale. Affermazioni del genere equivalgono alla mera constatazione che è stato fatto un errore, e alla promessa che si farà di tutto per migliorare. In effetti alcuni fallimenti sono davvero causati da errori prevedibili associati a precisi bias psicologici, e crediamo fermamente nella validità delle strategie volte a ridurre il bias (e il rumore) nei giudizi e nelle decisioni, ma attribuire a un bias ogni esito sfavorevole non serve a spiegare nulla. Raccomandiamo di riservare la parola "bias" a errori specifici che possono essere identificati, e ai meccanismi che ne sono alla base.

Sostituzione

Per sperimentare il processo euristico, provate voi stessi ad affrontare il seguente quesito, che illustra vari temi fondamentali legati all'approccio ai bias e alle euristiche. Come sempre, l'esempio sarà più utile se tenterete di avanzare una risposta.

> Bill è un trentatreenne intelligente ma con poca fantasia, abitudinario e nel complesso spento. A scuola era bravo in matematica, ma non brillava nelle materie umanistiche.
>
> Di seguito trovate un elenco di otto proposizioni associabili alla situazione attuale di Bill.
>
> Leggetele e scegliete le due che ritenete *più probabili.*
>
> ❑ Bill è un medico con l'hobby del poker.
> ❑ Bill è un architetto.
> ❑ Bill è un ragioniere.
> ❑ Bill suona il jazz per hobby.
> ❑ Bill ha l'hobby del surf.
> ❑ Bill è un reporter.
> ❑ Bill è un ragioniere che suona il jazz per hobby.
> ❑ Bill ha l'hobby dell'alpinismo.
>
> Ora tornate all'elenco e selezionate le due categorie di cui Bill *sembra* un tipico rappresentante. Potete scegliere le stesse di prima o altre.

Siamo piuttosto certi che abbiate riscontrato la più alta probabilità e la più alta somiglianza nelle stesse categorie; la nostra fiducia deriva dal fatto che molti esperimenti hanno evidenziato che le persone danno una risposta identica alle due domande.[3] Ma somiglianza e probabilità sono due concetti piuttosto diversi. Provate a chiedervi, per esempio, quale delle seguenti affermazioni ha più senso:

> ❑ Bill rispecchia la mia idea di una persona che suona il jazz per hobby.
> ❑ Bill rispecchia la mia idea di un ragioniere che suona il jazz per hobby.

Nessuna di queste due frasi è del tutto calzante, ma una è chiaramente meno terribile dell'altra: Bill ha più elementi in comune con un ragioniere che suona il jazz per hobby che con una persona che suona il jazz per

hobby. Ora riflettete su quest'altra domanda: quale di queste affermazioni è più probabile?

- ❑ Bill suona il jazz per hobby.
- ❑ Bill è un ragioniere che suona il jazz per hobby.

Forse sarete tentati di scegliere la seconda, ma la logica ve lo vieta. Non può che essere più probabile che Bill suoni il jazz per hobby, rispetto all'eventualità che sia un ragioniere con l'hobby del jazz. Ricordate i diagrammi di Venn? Se Bill è un ragioniere che suona il jazz per hobby, allora è in ogni caso qualcuno che suona il jazz. Aggiungere dettagli a una descrizione non fa altro che renderla meno probabile, benché forse più rappresentativa e quindi, in questo caso, più "calzante".

La teoria delle euristiche del giudizio suggerisce che talvolta la gente ricorre alla risposta a una domanda semplice per rispondere a una più difficile. Dovremmo chiederci a quale di queste due domande è più facile rispondere: «Qual è il grado di somiglianza tra Bill e un tipico jazzista dilettante?» o «Qual è la probabilità che Bill sia un jazzista dilettante?». Tutti riconoscerebbero che è più facile indicare la somiglianza, per cui è ben possibile che, di fronte alla richiesta di una valutazione sulla probabilità, la gente risponda invece alla prima domanda.

Questa è l'idea essenziale del programma di ricerca sulle euristiche e i bias: una tipica euristica utilizzata per rispondere a una domanda difficile è rispondere a una più semplice. Tale sostituzione di un quesito con l'altro produce errori prevedibili, chiamati bias psicologici.

Un bias di questo tipo è evidente nel caso di Bill: sostituendo un giudizio di somiglianza con uno di probabilità si incorrerà in un errore, perché la probabilità segue una sua precisa logica. In particolare, i diagrammi di

Venn sono applicabili solo alla probabilità, non alla somiglianza. Da qui derivano i prevedibili errori logici in cui molti cadono.

Per fare un altro esempio di una proprietà statistica spesso trascurata, torniamo all'esempio di Gambardi del capitolo 4. Se avete agito come quasi tutte le persone a cui è stato sottoposto il quesito, la vostra valutazione in merito alle possibilità di successo di Michael Gambardi si è basata sulla descrizione fornita, che avete poi confrontato con l'immagine che avete di un amministratore delegato di successo.

Ma avete pensato di tener conto di quante probabilità ha un amministratore delegato a caso di occupare la stessa posizione lavorativa a distanza di due anni? Probabilmente no. Potete immaginare queste *informazioni sul tasso di base* come una misura di quanto sia difficile sopravvivere come amministratore delegato. Se questo approccio vi sembra bizzarro, pensate a come valutereste la probabilità che un certo studente superi un esame: di certo la percentuale di studenti bocciati sarà un dato rilevante, in quanto fornirà un'indicazione della difficoltà dell'esame. Analogamente, il tasso di base della permanenza degli amministratori delegati è un'informazione rilevante nel caso di Gambardi. Entrambi i quesiti possono essere considerati esempi di quella che abbiamo definito la visione esterna: se si adotta questa visione, si considera lo studente, o Gambardi, all'interno di una categoria di casi simili, e si fa riferimento a questa categoria in termini statistici, invece di pensare al singolo caso in un'ottica causale.

Assumere la visione esterna può fare una grande differenza nel prevenire errori gravi. Basterà una rapida ricerca per scoprire che tra gli amministratori delegati delle società americane si stima un ricambio annuo del 15%, e che in media un nuovo amministratore delegato ha circa il 72% di probabilità di mantenere la sua posizione per almeno due anni.[4]

Naturalmente questo è solo un punto di partenza, e le informazioni più specifiche su Gambardi incideranno sulla valutazione finale. Ma concentrandosi esclusivamente su ciò che ci viene detto di lui, si trascura un dato fondamentale. (Confessiamo di avere allestito il caso Gambardi per illustrare il rumore nel giudizio; solo dopo diverse settimane ci siamo resi conto di come fornisse anche un ottimo esempio del bias qui descritto, chiamato *fallacia del tasso di base*. Evidentemente, anche per gli autori di questo libro non è automatico prendere in considerazione il tasso di base.)

La sostituzione di un quesito con un altro non riguarda solo la somiglianza e la probabilità: si incorre nello stesso problema quando si scambia l'impressione della facilità con cui vengono in mente esempi di determinati avvenimenti con un giudizio sulla loro frequenza. La percezione del rischio di incidenti aerei o di uragani, per dirne una, è più elevata subito dopo il verificarsi di eventi di questo tipo con un'ampia copertura mediatica. In teoria, un giudizio di rischio dovrebbe essere basato sulla media a lungo termine; in pratica, invece, si dà molto più peso agli incidenti recenti perché vengono subito in mente. Sostituire una stima della frequenza con un giudizio sulla facilità di richiamare alla memoria casi simili è un esempio di quella che viene definita *euristica della disponibilità*.

La sostituzione di un giudizio semplice con uno più complesso non si limita a questi esempi; al contrario, è molto comune. Rispondere a una domanda più semplice è una procedura che in genere adottiamo quando ci troviamo di fronte a quesiti che ci mettono in difficoltà. Pensate a come tendiamo a rispondere alle seguenti domande attraverso dei loro sostituti più semplici:

Credo al cambiamento climatico?
Mi fido di chi ne dichiara l'esistenza?

Ritengo che questo chirurgo sia competente?

Questa persona si presenta come autorevole e sicura di sé?

Il progetto verrà completato nei tempi previsti?
In questo momento è in linea con i tempi?

L'energia nucleare è necessaria?
La parola "nucleare" mi fa saltare sulla sedia?

In generale sono soddisfatto della mia vita?
Di che umore sono in questo momento?

A prescindere dal tema, sostituire una domanda con un'altra condurrà a una risposta che non dà un peso appropriato ai diversi aspetti da considerare, e questa errata ponderazione dei dati fattuali porta inevitabilmente all'errore. Per dare una piena risposta alla domanda sulla soddisfazione personale, per esempio, non bisognerebbe limitarsi a considerare il proprio umore attuale, ma le ricerche indicano che, in effetti, si dà un peso eccessivo proprio a questo aspetto.

Allo stesso modo, sostituire la probabilità con la somiglianza porta a trascurare i tassi di base, del tutto ininfluenti nei giudizi relativi a quest'ultima. E fattori come variazioni irrilevanti nella forma esteriore del documento di presentazione di un piano aziendale non dovrebbero avere alcun peso nella valutazione di una società. Il loro eventuale impatto sul giudizio probabilmente riflette una ponderazione errata dei dati fattuali, che condurrà a un errore.

Saltare alle conclusioni

In un momento chiave dello sviluppo della sceneggiatura di *Il ritorno dello Jedi*,[5] il terzo film in ordine di produzione della saga di *Guerre stellari*, George Lucas, il suo ideatore, ebbe un acceso dibattito con il suo eccellente collaboratore Lawrence Kasdan, che gli aveva consigliato di uccidere Luke

e dare più spazio a Leila.[6] Lucas si disse fermamente contrario all'idea, al che Kasdan propose di far morire un altro personaggio centrale, ma Lucas ribadì la sua contrarietà, aggiungendo: «Non si ammazza la gente così a caso». L'altro replicò con una sentita riflessione sul cinema, spiegando che «il film avrà una maggiore carica emotiva se viene meno un personaggio amato; il viaggio avrà un impatto più forte».

Lucas rispose senza mezzi termini: «L'idea non mi piace, e non credo sia così».

Qui il processo mentale sembra del tutto diverso da quello sperimentato a proposito di Bill, il ragioniere con il pallino del jazz. Torniamo alla risposta del creatore di *Guerre stellari*: «non mi piace» precede «non credo sia così». Di fronte al suggerimento di Kasdan, Lucas ha avuto una reazione automatica che ha contribuito a motivare il suo giudizio (poco importa che poi si sia rivelato corretto).

Questo esempio illustra un diverso tipo di bias, che chiamiamo *pregiudizio*, o *saltare alle conclusioni*. Come Lucas, tutti noi, quando iniziamo a elaborare un giudizio, spesso abbiamo già un'inclinazione verso una particolare conclusione. Così facendo, lasciamo che sia il sistema 1, cioè il pensiero rapido e intuitivo, a indicare la strada. Dopodiché, o saltiamo a quella conclusione, eludendo il processo di raccolta e integrazione delle informazioni, o mobilitiamo il sistema 2, il pensiero intenzionale, per trovare argomentazioni a sostegno del nostro pregiudizio. In tal caso, le prove che troveremo saranno selettive e distorte: a causa del *bias di conferma* e del *bias di desiderabilità*, tenderemo a raccogliere e a interpretare le informazioni in maniera selettiva per favorire, rispettivamente, un giudizio in cui crediamo già o che vorremmo fosse vero.[7]

Spesso le persone elaborano razionalizzazioni plausibili per i propri giudizi e arrivano a pensare che siano la causa dei propri convincimenti.

Per mettere alla prova il ruolo dei pregiudizi, basta immaginare che le argomentazioni a sostegno delle nostre convinzioni all'improvviso si dimostrino infondate. Kasdan, per esempio, avrebbe potuto far notare a Lucas che «non si ammazza la gente così a caso» non è proprio un'argomentazione stringente: l'autore di *Romeo e Giulietta* forse non sarebbe stato d'accordo con Lucas, e se gli sceneggiatori dei *Soprano* e del *Trono di spade* avessero escluso le uccisioni, con ogni probabilità entrambe le serie sarebbero state cancellate già alla prima stagione. Ma siamo certi che neanche una solida controargomentazione avrebbe fatto cambiare idea a George Lucas; al contrario, il regista avrebbe trovato altri argomenti a sostegno del suo giudizio (per esempio, «*Guerre stellari* è diversa dalle altre saghe!»).

I pregiudizi sono evidenti in qualsiasi campo, e, proprio come la reazione di Lucas, spesso hanno una componente emotiva. Lo psicologo Paul Slovic l'ha chiamata *euristica dell'affetto*: a determinare i pensieri delle persone sono i loro sentimenti. Ci piace quasi tutto dei politici che sosteniamo, mentre troviamo sgradevoli perfino l'aspetto e la voce di quelli che osteggiamo, che è lo stesso motivo per cui le società più astute fanno di tutto per associare un sentimento positivo al loro marchio. Spesso i professori notano che se alla fine dell'anno gli studenti attribuiscono al loro insegnamento un giudizio positivo, danno una valutazione alta anche ai materiali del corso; se invece agli studenti quel professore non piace molto, danno agli stessi materiali una valutazione più bassa. Lo stesso meccanismo si attiva anche in assenza di un fattore emotivo: a prescindere dalle vere ragioni alla base di una certa credenza, sarete più inclini ad accettare gli argomenti che sembrano confermarla, anche quando il ragionamento è sbagliato.[8]

Un esempio più sottile di questa tendenza a saltare alle conclusioni è rappresentato dall'*ancoraggio*, ovvero l'effetto che ha un numero arbitrario su chi deve esprimere un giudizio quantitativo. Un tipico esempio potrebbe essere una situazione in cui vi vengono mostrati oggetti di cui è difficile indovinare il prezzo, come una bottiglia di un vino che non conoscete.[9] Vi viene chiesto di trascrivere le ultime due cifre del vostro numero di telefono e di indicare se paghereste tale somma per quella bottiglia, poi vi si chiede di stabilire l'importo massimo che sareste disposti a pagare per quel vino. I risultati dimostrano che l'ancoraggio a cifre così casuali incide sul prezzo di acquisto finale: in uno studio americano, alcune persone con un alto ancoraggio (più di ottanta dollari) legato al loro numero di previdenza sociale dichiararono di essere disposte a pagare una somma tre volte superiore rispetto a chi aveva un basso ancoraggio (meno di venti dollari).

Chiaramente, il nostro numero di telefono non dovrebbe influenzare a tal punto il nostro giudizio sul valore di una bottiglia di vino, eppure è così. L'ancoraggio è un effetto estremamente robusto, spesso usato di proposito nelle contrattazioni.[10] Che stiate tirando sul prezzo al mercato o siate coinvolti in una complessa operazione commerciale, probabilmente fare la prima mossa tornerà a vostro vantaggio, perché, in virtù dell'effetto ancoraggio, la controparte sarà involontariamente indotta a pensare a come la vostra offerta possa risultare ragionevole. Le persone cercano sempre una logica in ciò che viene detto: di fronte a un numero inverosimile, giungono quasi senza volerlo a considerazioni atte a ridurne l'inverosimiglianza.

Eccesso di coerenza

Ora presenteremo un altro esperimento che vi aiuterà a riconoscere un terzo tipo di bias. Leggerete la descrizione di un candidato a un ruolo dirigenziale che consta di quattro aggettivi, ciascuno scritto su una carta. Le carte vengono mischiate, e le prime due estratte dal mazzo contengono i seguenti descrittori:

Intelligente, Tenace.

Per logica, bisognerebbe sospendere il giudizio finché non si possiedono tutte le informazioni, ma non è questo che accade: vi sarete già fatti un'idea del candidato, un'idea positiva. Ciò è avvenuto senza che aveste alcun controllo del vostro processo mentale, e senza che poteste sospendere il giudizio.

A questo punto, scoprite le altre due carte. Ecco la descrizione completa:

Intelligente, Tenace, Scaltro, Spregiudicato.

La vostra valutazione non è più ugualmente favorevole, ma non è cambiata del tutto. Ora paragonate la precedente descrizione con questa, che sarebbe potuta capitare se solo aveste estratto le quattro carte in un ordine diverso:

Spregiudicato, Scaltro, Tenace, Intelligente.

Questa seconda descrizione comprende gli stessi aggettivi, eppure, per via dell'ordine in cui vengono presentati, è chiaramente molto meno attraente della prima. La parola "scaltro" era solo leggermente negativa, quando era preceduta da "intelligente" e "tenace", perché continuavamo a credere (senza alcuna ragione valida) che il dirigente fosse mosso da buone intenzioni. Eppure, posta dopo "spregiudicato", la stessa parola ha un effetto terribile. In questo contesto, tenacia e intelligenza non sono più qualità positive: rendono una persona cattiva ancora più pericolosa.

L'esperimento illustra l'*eccesso di coerenza*: ci formiamo velocemente delle impressioni coerenti e siamo lenti ad abbandonarle.[11] In questo esempio abbiamo subito assunto un atteggiamento positivo verso il candidato, pur avendo scarsi riscontri. Il bias di conferma – la tendenza che ci porta, in presenza di un pregiudizio, a ignorare ogni dato contraddittorio – ci ha indotti ad attribuire minore importanza del dovuto ai dati successivi. (Un altro termine per indicare questo fenomeno è *effetto alone*, perché il candidato è stato valutato sulla base dell'"alone" positivo della prima impressione. Nel capitolo 24 vedremo come l'effetto alone ponga seri problemi nelle decisioni relative al reclutamento del personale.)

Facciamo un altro esempio. Negli Stati Uniti le catene di ristorazione sono obbligate per legge a inserire etichette nutrizionali per comunicare ai consumatori quante calorie contengono cheeseburger, hamburger, insalate e tutti gli altri piatti serviti. Dopo aver letto le etichette, i consumatori compiranno scelte diverse? I risultati sono controversi e contrastanti, ma uno studio rivelatore ha evidenziato come vi sia una più elevata probabilità che queste etichette influenzino i consumatori se poste a sinistra del prodotto invece che a destra.[12] Quando l'indicazione è posta sulla sinistra, i consumatori ricevono subito l'informazione ed evidentemente pensano: "Quante calorie!" o "Mica tante calorie!" ancor prima di vedere il prodotto, e questa reazione iniziale positiva o negativa influenza notevolmente le loro scelte. Per contro, quando le persone vedono prima il prodotto, a quanto pare pensano: "Che buono!" o "Insomma...", prima di leggere l'etichetta. Anche in questo caso, la reazione iniziale influisce notevolmente sulla decisione finale. Questa ipotesi trova conferma in un altro risultato riportato dagli autori: per i parlanti di lingua ebraica, che leggono da destra a sinistra, l'etichetta

nutrizionale ha un impatto di gran lunga più significativo se posta a destra piuttosto che a sinistra del prodotto.

In generale, saltiamo alle conclusioni e non le molliamo: pensiamo che le nostre opinioni si fondino sui fatti, ma i fatti che consideriamo e la nostra interpretazione degli stessi probabilmente verranno distorti, almeno in una certa misura, perché si adattino al nostro giudizio sommario iniziale. Di conseguenza, nel complesso restiamo coerenti alla storia che ha preso il sopravvento nella nostra mente. Nel caso in cui le conclusioni si dimostrino corrette, questo procedimento è ancora accettabile, ma quando la valutazione iniziale è errata, la tendenza a non staccarsene contro ogni evidenza che ne dimostri l'infondatezza ha molte probabilità di amplificare l'errore. E questo effetto è difficile da controllare, perché i fatti che abbiamo visto o sentito sono impossibili da ignorare, e quasi sempre difficili da dimenticare. In tribunale spesso i giudici chiedono ai giurati di non tenere in considerazione una prova inammissibile, ma la richiesta è poco realistica (per quanto possa risultare utile nella fase in cui la giuria delibera sul caso, quando le argomentazioni basate esplicitamente su quella prova possono essere rigettate).

I bias psicologici producono rumore

Abbiamo presentato per sommi capi tre tipi di bias che operano in modi diversi: i bias di sostituzione, che portano ad attribuire un peso sbagliato ai fatti; il saltare alle conclusioni (o pregiudizio), un bias che ci porta ad aggirare i fatti o a considerarli in maniera distorta; l'eccesso di coerenza, che amplifica l'effetto delle impressioni iniziali e riduce l'impatto delle informazioni contraddittorie. Tutti e tre questi tipi di bias, naturalmente, possono portare a un bias statistico, ma anche creare rumore.

Partiamo dalla sostituzione. La maggior parte delle persone valuta la probabilità che Bill sia un ragioniere alla luce della somiglianza del suo profilo con lo stereotipo: il risultato, in questo esperimento, è un bias condiviso. Se ogni rispondente commette lo stesso errore, non vi è rumore. Ma la sostituzione non produce sempre una simile unanimità. Quando la domanda «Esiste il cambiamento climatico?» viene sostituita con «Mi fido di chi ne dichiara l'esistenza?», è chiaro che la risposta varierà da persona a persona, a seconda del milieu sociale, delle fonti di informazione abituali, dell'affiliazione politica e così via. Lo stesso bias psicologico crea una varietà di giudizi, quindi un rumore interpersonale.

La sostituzione può inoltre generare rumore occasionale. Se un individuo risponde a una domanda sulla propria soddisfazione personale in base all'umore del momento, inevitabilmente darà risposte diverse perfino nel corso della stessa giornata. A una mattina felice può seguire un pomeriggio angosciante, e i cambiamenti di umore temporanei possono condurre a indicazioni molto diverse sulla soddisfazione personale a seconda del momento in cui si riceve la chiamata dell'intervistatore. Come ricorderete, nel capitolo 7 abbiamo esaminato alcuni esempi di rumore occasionale riconducibili a bias psicologici.

Anche i pregiudizi producono sia bias sia rumore. Torniamo a un esempio menzionato nell'introduzione: la sconvolgente disparità nella percentuale di richiedenti asilo ammessi dai giudici. Quando un giudice ammette il 5% dei richiedenti e un altro, nel medesimo tribunale, accetta l'88% delle richieste, possiamo essere sicuri che vi siano bias nell'una e nell'altra direzione. Da un punto di vista più generale, le differenze individuali nei bias possono provocare un rumore sistemico enorme. Naturalmente è anche possibile che il sistema sia talmente affetto da bias

che le decisioni di tutti i giudici (o quasi) siano influenzate in maniera simile.

Infine, l'eccesso di coerenza può produrre bias o rumore, a seconda che la sequenza di informazioni e il significato loro attribuito siano identici per tutti i decisori (o quasi) oppure no. Poniamo, per esempio, un candidato fisicamente attraente che, in virtù del suo aspetto, faccia una prima impressione positiva alla maggior parte dei reclutatori: se l'aspetto fisico è irrilevante per la posizione a cui ambisce il candidato, l'alone positivo condurrà a un errore condiviso, ovvero a un bias.

D'altro canto, molte decisioni complesse richiedono informazioni che arrivano essenzialmente in ordine sparso. Pensate ai periti assicurativi del capitolo 2: l'ordine in cui accedono ai dati relativi a una richiesta di indennizzo cambia da un perito all'altro e da un caso all'altro, generando variazioni casuali nelle impressioni iniziali che per effetto dell'eccesso di coerenza produrranno distorsioni altrettanto casuali nei giudizi finali. Ciò che ne conseguirà sarà un rumore sistemico.

In breve, i bias psicologici sono meccanismi universali che spesso producono errori condivisi; ma quando vi sono grandi differenze individuali tra i bias (pregiudizi diversi), o quando l'effetto dei bias dipende dal contesto (fattori scatenanti diversi), ci sarà rumore.

Sia il bias sia il rumore portano all'errore, pertanto qualsiasi cosa riduca i bias psicologici migliorerà il giudizio. Torneremo al tema dell'eliminazione dei bias nella parte 5. Per il momento, proseguiamo nell'esplorazione del processo di giudizio.

A proposito di euristiche, bias e rumore

«Sappiamo di avere dei bias psicologici, ma dovremmo evitare di attribuire qualsiasi errore a "bias" non meglio specificati.»

«Quando sostituiamo il quesito a cui dovremmo rispondere con un altro più semplice, rischiamo di cadere in errore. Per esempio, giudicando la probabilità sulla base della somiglianza trascureremo il tasso di base.»

«I pregiudizi e altri bias che portano a saltare alle conclusioni inducono le persone a cercare di piegare i dati fattuali alla propria posizione iniziale.»

«Ci formiamo rapidamente un'impressione, e vi restiamo attaccati perfino quando subentrano informazioni contraddittorie. Questa tendenza viene definita eccesso di coerenza.»

«I bias psicologici possono diventare bias statistici, se condivisi da molti. Spesso, però, le persone presentano bias diversi: in questi casi, i bias psicologici creano rumore sistemico.»

[1] D. Kahneman, *Pensieri lenti e veloci*, cit.

[2] Qui occorre fare una precisazione. Gli psicologi che studiano i bias di giudizio non si accontentano di avere cinque partecipanti per gruppo come nella figura 12, per un valido motivo: poiché i giudizi sono affetti da rumore, raramente i risultati di ogni gruppo sperimentale si raccoglieranno in un cluster ben definito come quello rappresentato in figura. Le persone hanno una diversa suscettibilità a ogni bias e non trascurano *del tutto* le variabili rilevanti. Per esempio, con un gran numero di partecipanti, quasi certamente si riscontrerebbe che l'insensibilità alle dimensioni del campione è imperfetta: la probabilità media assegnata alla possibilità che Gambardi perda il lavoro è leggermente più alta su tre anni che su due. La descrizione dell'insensibilità alle dimensioni del campione è però comunque appropriata, perché la differenza tra le previsioni dei due gruppi è molto più ridotta di quanto dovrebbe essere.

[3] D. Kahneman *et al.* (a cura di), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, New York 1982, cap. 6; D. Kahneman, A. Tversky, *On the Psychology of Prediction*, in "Psychological Review", 80(1973), n. 4, pp. 237-251.

[4] Vedi, per esempio, S.N. Kaplan, B.A. Minton, *How Has CEO Turnover Changed?*, in "International Review of Finance", 12(2012), n. 1, pp. 57-87. Vedi anche D. Jenter, K. Lewellen, *Performance-Induced CEO Turnover*, in "Harvard Law School Forum on Corporate Governance", 2 settembre 2020, [corpgov.law.harvard.edu/2020/09/02/performance-induced-ceo-turnover].

[5] J.W. Rinzler, *The Making of Star Wars Return of the Jedi: The Definitive Story*, Del Rey, New York 2013, p. 64.

[6] C. Sunstein, *The World According to Star Wars*, HarperCollins, New York 2016.

[7] Ci riferiamo qui al semplice caso in cui un pregiudizio esiste già nel momento in cui si inizia a formulare un giudizio. In realtà, anche in assenza di tale pregiudizio, un bias verso una certa conclusione può svilupparsi via via che si accumulano le informazioni, in virtù della tendenza alla semplicità e alla coerenza. Quando emerge una conclusione provvisoria, il bias di conferma interviene sulla raccolta e sull'interpretazione delle nuove informazioni a suo favore.

[8] Questo tipo di ragionamento è stato definito *bias della credenza*. Vedi J.St.B.T. Evans, J.L. Barson, P. Pollard, *On the Conflict between Logic and Belief in Syllogistic Reasoning*, in "Memory & Cognition", 11(1983), n. 3, pp. 295-306.

[9] D. Ariely, G. Loewenstein, D. Prelec, *'Coherent Arbitrariness': Stable Demand Curves Without Stable Preferences*, in "Quarterly Journal of Economics", 118(2003), n. 1, pp. 73-105.

[10] A.D. Galinsky, T. Mussweiler, *First Offers as Anchors: The Role of Perspective-Taking and Negotiator Focus*, in "Journal of Personality and Social Psychology", 81(2001), n. 4, pp. 657-669.

[11] S.E. Asch, *Forming Impressions of Personality*, in "Journal of Abnormal and Social Psychology", 41(1946), n. 3, pp. 258-290, impiegò per la prima volta una serie di aggettivi in ordine diverso per illustrare questo fenomeno.

[12] S.K. Dallas *et al.*, *Don't Count Calorie Labeling Out: Calorie Counts on the Left Side of Menu Items Lead to Lower Calorie Food Choices*, in "Journal of Consumer Psychology", 29(2019), n. 1, pp. 60-69.

14

L’operazione di matching

Guardate il cielo: quante probabilità ci sono che tra due ore piova?

Forse non vi sarà difficile rispondere a questa domanda. Il vostro giudizio – per esempio, che sia “molto probabile” che piova presto – non ha richiesto alcuno sforzo: in qualche modo la vostra valutazione del cielo grigio si è tradotta in un giudizio di probabilità.

Quello che avete appena svolto è un esempio elementare di matching. Se abbiamo detto che il giudizio è un’operazione in cui si assegna un valore a un’impressione soggettiva (o a un aspetto di una certa impressione) all’interno di una scala, il matching è una parte fondamentale di tale processo. Quando ci viene chiesto quanto siamo felici da uno a dieci, o come valutiamo la nostra esperienza di acquisto da una a cinque stelle, compiamo un’operazione di matching: ci proponiamo di individuare un valore che corrisponda al nostro umore o alla nostra esperienza all’interno di una scala di giudizio.

Matching e coerenza

Torniamo al caso di Bill del precedente capitolo, in cui lo abbiamo descritto in questi termini: «Bill è un trentatreenne intelligente ma con poca fantasia, abitudinario e nel complesso spento. A scuola era bravo in matematica ma non brillava nelle materie umanistiche». Vi abbiamo chiesto di elaborare una stima della probabilità che Bill avesse varie

occupazioni e hobby, e abbiamo visto che avete risposto alla domanda sostituendo un giudizio di probabilità con uno di somiglianza: non vi siete davvero chiesti quali erano le probabilità che Bill fosse un ragioniere, ma quanto era simile allo stereotipo di quella professione. Torniamo allora a una domanda lasciata in sospeso: come siete arrivati a questo giudizio?

Non è difficile valutare fino a che punto la descrizione di Bill corrisponde agli stereotipi su professioni e hobby: è chiaro che Bill somiglia più a un ragioniere che a un tipico jazzista, e ancor meno a un surfista. L'esempio illustra la straordinaria versatilità del matching, particolarmente evidente nei giudizi sulle persone. Avreste potuto rispondere a qualsiasi tipo di domanda su Bill, per esempio: come vi sentireste se doveste ritrovarvi su un'isola deserta con lui? Probabilmente avrete una risposta intuitiva immediata a questa domanda sulla base delle poche informazioni fornite. Ma se vi dicessimo che Bill è un esploratore provetto che ha sviluppato tecniche di sopravvivenza fuori dal comune? Se questo vi sorprende, come presumiamo, è perché vi siete scontrati con una mancanza di coerenza.

L'intensità della sorpresa è data dall'incompatibilità di questa nuova informazione con l'immagine che vi eravate fatti di Bill in precedenza. Se il suo coraggio e le sue tecniche di sopravvivenza fossero state menzionate nella descrizione iniziale, sareste arrivati a un'immagine del tutto diversa di quest'uomo, magari quella di una persona che esprime la sua vitalità solo nei grandi spazi aperti. L'impressione generale di Bill sarebbe stata meno coerente, quindi più difficile da far corrispondere a una categoria professionale oppure a un hobby, ma non avreste reagito con la stessa sorpresa.

I segnali contraddittori rendono più difficile giungere a un senso di coerenza e arrivare a un giudizio che sembri corrispondere in modo soddisfacente alla realtà. Poiché i giudizi complessi sono caratterizzati dalla

presenza di segnali contrastanti, ci aspettiamo di ritrovarvi molto rumore. Un giudizio di questo tipo era richiesto nel caso Gambardi, in cui alcune indicazioni erano positive e altre negative. Torneremo sui giudizi complessi nel capitolo 16, mentre nella restante parte di questo capitolo ci concentreremo su giudizi relativamente semplici, e in particolare su quelli basati su *scale di intensità*.

Matching di intensità

Alcune delle scale su cui ci basiamo per esprimere giudizi sono di tipo qualitativo, come nel caso di professioni, hobby e diagnosi mediche. Si distinguono per il fatto che i loro valori non sono ordinati: il rosso non è né superiore né inferiore al blu.

Molti giudizi, tuttavia, vengono espressi sulla base di scale di intensità quantitativa. Le misurazioni fisiche di dimensione, peso, luminosità, temperatura o volume sonoro, le stime di costo, di valore, di probabilità o di frequenza, sono tutti esempi di giudizi che utilizzano scale di tipo quantitativo, così come quelli basati su scale più astratte come il grado di fiducia, di forza, di bellezza, di rabbia, di paura, di immoralità o di severità della pena.

Il tratto distintivo di tutte queste dimensioni quantitative è che è sempre possibile porre in ordine crescente due valori qualsiasi appartenenti alla stessa scala. Possiamo dire che la fustigazione è una punizione più severa di una tiratina d'orecchi, o che preferiamo *Amleto* ad *Aspettando Godot*, proprio come è possibile dire che il sole è più luminoso della luna, che un elefante pesa più di un criceto e che la temperatura media di Miami è più elevata di quella di Toronto.

Le persone hanno grandi capacità intuitive quando si tratta di trovare una corrispondenza tra l'intensità di dimensioni non correlate tra loro, sovrapponendo una scala di intensità all'altra.[1] Potreste correlare l'attuale livello di conflitto politico nel vostro paese con la temperatura estiva di una città che conoscete bene, oppure l'intensità dell'affetto che provate per vari cantanti all'altezza degli edifici della vostra città. (Se pensate che Bob Dylan sia il migliore, per esempio, potreste correlare il vostro entusiasmo per lui al palazzo più alto che vedete dalla finestra.) E se vi venisse chiesto di valutare un ristorante paragonandolo alla lunghezza di un romanzo invece di attribuirgli il solito valore su una scala da una a cinque stelle, questa richiesta vi sembrerebbe curiosa ma non impossibile. (Il vostro ristorante preferito potrebbe essere associato a *Guerra e pace*.) In ogni caso, per quanto strano, sarebbe chiaro cosa volete dire.

Nelle conversazioni di ogni giorno, l'intervallo di valori all'interno di una scala varia in funzione del contesto: se diciamo che una persona ha messo da parte una bella somma, questa frase avrà un certo significato se stiamo festeggiando il pensionamento di un consulente finanziario di successo, e un altro se vogliamo complimentarci con una giovane babysitter. Anche i concetti di *grande* e *piccolo* dipendono in tutto e per tutto dal sistema di riferimento: una frase come «Un grande topo è salito sulla proboscide del piccolo elefante», per esempio, non ci sembra priva di senso.

Il bias delle previsioni basate sul matching

Il seguente indovinello illustra allo stesso tempo il potere del matching e un errore di giudizio a esso sistematicamente associato.[2]

> Julie è in procinto di laurearsi. Sulla base dell'informazione fornita di seguito, indovinate qual è la media dei suoi voti (su una scala standard da 0,0 a 4,0):
>
> Julie legge speditamente dall'età di quattro anni.

Che media avrà?

Se avete familiarità con il sistema di calcolo della media dei voti americano, vi sarà subito venuto in mente un numero, probabilmente vicino a 3,7 o 3,8. Questa stima immediata della media dei voti di Julie illustra il processo di matching appena descritto.

Innanzitutto avete valutato la lettura precoce di Julie; in questo caso si trattava di una valutazione facile, perché Julie ha iniziato a leggere insolitamente presto, e la sua precocità ha fatto sì che la collocaste in una categoria all'interno di una certa scala. Se doveste descrivere la scala che avete impiegato, probabilmente direste che al punto più alto avete posto la categoria degli "straordinariamente precoci", e notereste che Julie non vi rientra (alcuni bambini iniziano a leggere prima dei due anni). Probabilmente Julie rientrerebbe in quella successiva, la fascia dei bambini "insolitamente ma non straordinariamente precoci".

In un secondo momento avete stabilito una corrispondenza tra un giudizio sulla media dei voti e la vostra valutazione di Julie: senza neanche rendervene conto, avrete cercato un valore della media dei voti in linea con la descrizione "insolita ma non straordinaria". Leggendo la storia di Julie, avete formulato di punto in bianco una *previsione basata sul matching*.

Eseguire i calcoli richiesti per svolgere questi compiti di valutazione e matching richiederebbe del tempo, ma nel sistema 1 dei pensieri veloci si arriva a un giudizio rapido senza alcuno sforzo. Il processo che vi ha portato alla valutazione della media dei voti di Julie richiede una complessa sequenza di eventi mentali non direttamente osservabili. La specificità del meccanismo mentale del matching è insolita in psicologia, ma le evidenze scientifiche non lasciano dubbi. Sulla base di molti esperimenti di questo tipo possiamo essere certi che le due domande riportate di seguito, se poste a diversi gruppi di persone, daranno esattamente gli stessi risultati:[3]

- ❑ In termini percentuali, quanti studenti del corso di Julie hanno cominciato a leggere prima di lei?
- ❑ In termini percentuali, quanti studenti del corso di Julie hanno una media dei voti più alta della sua?

La prima domanda è gestibile in autonomia: viene semplicemente chiesto di valutare le informazioni che sono state fornite su Julie. La seconda, che richiede una previsione lontana, certo è più difficile, ma d'istinto si è tentati di dare la stessa risposta già fornita alla domanda precedente.

Queste due domande su Julie sono analoghe a quelle che abbiamo definito disorientanti per tutti, parlando dell'illusione di validità. La prima implica una valutazione sull'"intensità" delle informazioni che avete ricevuto su Julie, mentre la seconda verte sull'intensità di una previsione. E anche dopo questa spiegazione, abbiamo il sospetto che risulti ancora difficile distinguere tra le due.

La previsione intuitiva della media dei voti di Julie chiama in causa il meccanismo psicologico descritto nel capitolo 13: la sostituzione di una domanda difficile con un'altra semplice. Il sistema 1 semplifica la richiesta di elaborare una previsione difficile, trasformandola in una più semplice: quanto ci colpisce il fatto che Julie abbia cominciato a leggere a quattro anni? Per passare direttamente dall'età, misurata in anni, alla media dei voti, misurata in punti, occorre operare un ulteriore matching.

La sostituzione, naturalmente, può avvenire soltanto se le informazioni disponibili sono pertinenti: se di Julie sapessimo solo che ha un talento per la corsa o che il ballo non è il suo forte, non avremmo alcuna informazione. Ma qualsiasi dato che possa essere interpretato come un'indicazione plausibile della sua intelligenza potrà costituire un accettabile sostituto.

Rimpiazzare un quesito con un altro non potrà che indurre in errore quando le vere risposte alle due domande sono diverse: per quanto plausibile possa sembrare, la sostituzione della media dei voti con l'età in

cui Julie ha iniziato a leggere è evidentemente assurda. Pensate infatti agli eventi che potrebbero essersi verificati nella vita della ragazza dopo i quattro anni: magari è stata coinvolta in un terribile incidente, o è rimasta traumatizzata dal divorzio dei suoi genitori; può avere incontrato sul suo percorso un insegnante carismatico ed esserne stata profondamente influenzata; forse è rimasta incinta giovanissima. Ognuno di questi eventi, così come molti altri ancora, potrebbe avere inciso sui suoi studi universitari.

La previsione basata sul matching si può giustificare solo nel caso in cui vi fosse una perfetta correlazione tra la capacità di lettura precoce e la media dei voti universitari, ma evidentemente non è così. D'altro canto, anche tralasciare del tutto l'informazione sull'età in cui Julie ha iniziato a leggere sarebbe un errore, perché questo dato in effetti è pertinente. La previsione ottimale deve collocarsi tra i due estremi della conoscenza assoluta e della totale ignoranza.

Quanto sappiamo di un caso quando non abbiamo nessuna informazione specifica, ma conosciamo solo la categoria in cui rientra? Per rispondere a questa domanda bisogna ricorrere a quella che abbiamo definito la "visione esterna" del caso. Se ci fosse stato chiesto di avanzare una previsione sulla media dei voti di Julie senza avere alcuna informazione su di lei, sicuramente la nostra previsione si sarebbe attestata su un valore medio come 3,2. Questa è una previsione basata sulla visione esterna. La stima migliore della media dei voti di Julie dovrà quindi essere superiore a 3,2 e inferiore a 3,8. La precisa collocazione dipenderà dal valore predittivo dell'informazione: più vi fidate del fatto che l'età in cui un bambino inizia a leggere sia un valido predittore della media dei voti, più alta sarà la vostra stima. Nel caso di Julie, si tratta di un'informazione piuttosto debole, di conseguenza la previsione più ragionevole sarà più vicina al valore medio.

Vi è un metodo tecnico ma piuttosto semplice per correggere l'errore nelle previsioni basate sul matching, di cui troverete una descrizione dettagliata nell'appendice C.

Sebbene conducano a risultati assurdi dal punto di vista statistico, è difficile resistere alla tentazione di formulare previsioni che sembrano adattarsi bene alla realtà fattuale. I direttori commerciali spesso presumono che il commesso che quest'anno ha dato i risultati migliori farà lo stesso in futuro; gli alti dirigenti a volte si imbattono in un candidato di grande talento e immaginano che, una volta assunto, scalerà i vertici dell'organizzazione; i produttori si aspettano sempre che il prossimo film di un regista che ha sbancato con la sua ultima uscita avrà altrettanto successo.

Previsioni di questo tipo sono destinate a portare a grosse delusioni; per contro, se vengono formulate in un momento particolarmente negativo, risulteranno con ogni probabilità troppo pessimistiche. Le previsioni intuitive correlate ai dati fattuali sono sempre troppo polarizzate, sia in positivo sia in negativo. (Tecnicamente, questi errori predittivi si definiscono *non regressivi*, perché non prendono in considerazione il fenomeno statistico chiamato *regressione verso la media*.)

Occorre notare, tuttavia, che la sostituzione e il matching non sono sempre alla base delle previsioni. Nell'ottica dei due sistemi, il sistema 1, quello dei pensieri intuitivi, propone soluzioni associative veloci ai problemi nel momento in cui si presentano, ma queste intuizioni devono essere accolte dal sistema 2 dei pensieri riflessivi prima di diventare credenze. Le previsioni basate sul matching talvolta vengono respinte a favore di risposte più complesse: per esempio, le persone sono più restie a correlare le loro previsioni a dati di realtà sfavorevoli. È ipotizzabile, dunque, che avreste avuto qualche esitazione a prevedere che Julie

avrebbe preso voti più bassi se avesse cominciato a leggere tardi. Questa asimmetria fra previsioni favorevoli e sfavorevoli svanisce quando si dispone di più informazioni.

Proponiamo la visione esterna come correttivo per previsioni intuitive di qualsiasi tipo. Facciamo un esempio: in precedenza, parlando delle prospettive future di Michael Gambardi, vi abbiamo suggerito di ancorare il vostro giudizio sulle sue probabilità di successo al relativo tasso di base (il tasso di successo degli amministratori delegati a due anni dall'insediamento). Nel caso di previsioni quantitative come quella sui voti di Julie, assumere la visione esterna significa ancorare la vostra previsione al risultato medio. Questo tipo di correttivo può essere trascurato solo per i problemi di facile risoluzione, quando le informazioni disponibili incoraggiano a esprimere previsioni con la massima sicurezza. Quando è richiesto un giudizio serio, la visione esterna deve essere tenuta in considerazione.

Il rumore nel matching: i limiti del giudizio assoluto

La nostra capacità limitata di distinguere tra più categorie all'interno di una scala di intensità limita a sua volta l'accuratezza dell'operazione di matching: parole come "grande" o "ricco" assegnano la stessa etichetta a un intervallo di valori relativi alle dimensioni della grandezza o della ricchezza. Ciò può costituire una notevole fonte di rumore.

Il consulente finanziario che va in pensione merita certamente l'etichetta di "ricco", ma *quanto* è ricco? Possiamo scegliere tra vari aggettivi: benestante, abbiente, agiato, facoltoso, ricchissimo e così via. Se vi venisse fornita una descrizione dettagliata della ricchezza di alcuni individui e doveste associare un aggettivo a ciascuno di loro, quante

categorie diverse formereste, senza ricorrere a confronti specifici tra i vari casi?

Il numero di categorie che siamo in grado di distinguere su una scala di intensità è indicato nel titolo dell'articolo *The Magical Number Seven, Plus or Minus Two* ("Il numero magico sette, più o meno due"), pubblicato nel 1956 e divenuto un classico della psicologia.[4] Superato questo limite, le persone iniziano a commettere errori come, per esempio, assegnare A a una categoria più alta di B, quando in un confronto testa a testa piazzerebbero B più in alto rispetto ad A.

Immaginate un insieme di linee di quattro lunghezze diverse tra i 2 e i 4 centimetri, in cui ogni linea supera la successiva di una stessa lunghezza. Le linee vi vengono mostrate una alla volta, e dovete assegnare loro un numero da 1 a 4, dove 1 indica la più corta e 4 la più lunga. Semplice, no? Ora supponiamo che vi vengano mostrate delle linee di cinque lunghezze diverse e dobbiate ripetere la stessa operazione assegnando loro un numero da 1 a 5. Ancora semplice. Quand'è che inizierete a commettere errori? Quando vi avvicinerete al numero magico di sette linee. Questa cifra, stranamente, ha molto poco a che fare con l'intervallo di lunghezza: se fossero lunghe dai 2 ai 6 centimetri piuttosto che dai 2 ai 4, comincereste comunque a sbagliare una volta superate le sette linee. Lo stesso risultato si ottiene con suoni ascoltati a volumi differenti o luci di diversa intensità. C'è un limite nella capacità delle persone di assegnare etichette differenziate a stimoli riguardanti una determinata dimensione, e si colloca in prossimità del numero sette.

Questo limite nella facoltà di operare distinzioni è importante, perché la nostra capacità di correlare dei valori appartenenti a diverse scale di intensità non potrà superare quella di assegnare dei valori all'interno di ciascuna scala. L'operazione di matching è uno strumento versatile del

sistema 1 del pensiero veloce, ed è alla base di molti giudizi intuitivi, ma è alquanto approssimativa.

Il numero magico non è un vincolo assoluto: le persone possono essere addestrate a operare distinzioni più sottili attraverso categorizzazioni gerarchiche. Per esempio, possiamo certamente distinguere diverse categorie di ricchezza tra i multimilionari, e i giudici possono distinguere diversi gradi di gravità in varie categorie di reati, a loro volta disposti in ordine di gravità. Per arrivare a un tale livello di sofisticazione, però, è necessario avvalersi di categorie preesistenti, tra cui sia possibile tracciare delle nette linee di confine. In altre parole, nell'assegnare etichette a una serie di linee non potete decidere di separare le più lunghe dalle più corte e di trattarle come due categorie diverse. Nella modalità del pensiero veloce, la categorizzazione non è sotto il nostro controllo volontario.

C'è un modo per venire a capo della limitata differenziazione interna delle scale aggettivali: invece di usare delle etichette, si può operare un confronto. Siamo molto più bravi a paragonare diversi casi, piuttosto che a piazzarli su una scala.

Pensate a cosa fareste se vi venisse chiesto di impiegare una scala da 1 a 20 per valutare un ampio insieme di ristoranti, oppure di cantanti. Una valutazione da 1 a 5 stelle sarebbe facilmente gestibile, ma non potrete garantire una perfetta affidabilità su una scala di venti punti. (Big Pizza si merita tre stelle, ma gli dareste undici o dodici punti?) La soluzione è semplice, anche se non veloce. Prima valutereste i ristoranti, o i cantanti, su una scala da 1 a 5 per suddividerli in cinque categorie; poi ordinereste ciascun elemento all'interno di ogni categoria, operazione che dovreste essere in grado di fare senza molte sovrapposizioni: probabilmente saprete se preferite Big Pizza a Burger King, o Taylor Swift a Bob Dylan, anche se li avete assegnati alla stessa categoria. Per semplificare, a questo punto

potreste identificare quattro livelli all'interno di ognuna delle cinque categorie e distinguere tra diversi livelli di disprezzo anche tra i cantanti che odiate di più.[5]

La *ratio* psicologica soggiacente a questo esercizio è chiarissima: i paragoni espliciti tra diversi oggetti di giudizio consentono distinzioni molto più sottili delle valutazioni singole degli stessi oggetti esaminati uno per volta. Lo stesso vale per i giudizi sulla lunghezza delle linee di cui parlavamo in precedenza: la nostra capacità di paragonare linee che ci vengono mostrate in immediata successione è di gran lunga superiore a quella di etichettare le diverse lunghezze, e saremo ancora più accurati quando le linee da confrontare saranno visibili in contemporanea.

I vantaggi dei giudizi comparativi si riscontrano in molti campi. Se avete una vaga idea di quanto siano ricche un certo numero di persone, fareste meglio a paragonare individui appartenenti allo stesso intervallo che a etichettarli uno per uno sulla base della loro ricchezza. Se dovete valutare degli elaborati, sarete più precisi se li ordinate dal migliore al peggiore, invece di leggerli e giudicarli singolarmente. I giudizi comparativi o relativi sono più accurati di quelli categorici o assoluti; come suggeriscono gli esempi citati, però, richiedono più tempo e più energie.

Valutare singoli oggetti su una scala dichiaratamente comparativa conserva alcuni dei vantaggi del giudizio comparativo. In alcuni contesti, primo fra tutti l'istruzione, in vista dell'ammissione o della promozione dei candidati spesso si richiede a chi li presenta di inserirli nella fascia del "migliore 5%" o del "migliore 20%" rispetto a una data popolazione, per esempio "il totale dei vostri studenti", oppure "i programmatori con pari livello di esperienza". Queste valutazioni andrebbero sempre prese *cum grano salis*, perché non c'è modo di assicurarsi che chi le formula stia usando correttamente la scala di valutazione. Solo in alcuni contesti si

potrà risalire al valutatore: prendiamo il caso di un analista finanziario che, nello stimare un certo numero di investimenti, assegni il 90% dei casi alla fascia del "migliore 20%"; una discrepanza così evidente sarà facile da individuare e correggere. L'uso di giudizi comparativi è uno dei rimedi al rumore che esamineremo nella parte 5.

Molti compiti di giudizio richiedono il matching di casi singoli con una categoria di una certa scala (per esempio, una scala di gradimento a sette punti) oppure l'impiego di un gruppo di aggettivi ordinati (come "improbabile" o "estremamente improbabile" per valutare la probabilità di un evento). Questo tipo di matching introduce rumore, perché è approssimativo: gli individui possono interpretare in maniera diversa le etichette anche quando concordano sulla sostanza del giudizio. Una procedura che impone esplicitamente dei giudizi comparativi tenderà a ridurre il rumore. Nel prossimo capitolo analizzeremo ulteriormente come l'impiego delle scale sbagliate possa, al contrario, aumentarlo.

A proposito di matching

«Siamo d'accordo che questo film è "bellissimo", ma sembra che a te sia piaciuto molto meno che a me. Stiamo usando lo stesso aggettivo, ma anche la stessa scala?»

«Pensavamo che la seconda stagione di questa serie sarebbe stata strepitosa come la prima. Abbiamo basato la nostra previsione sul matching, e ci siamo sbagliati.»

«È difficile essere coerenti quando si valutano degli elaborati. Non sarebbe meglio stabilire una graduatoria?»

[1] S.S. Stevens, *On the Operation Known as Judgment*, in "American Scientist", 54(1966), n. 4, pp. 385-401. In questo libro utilizziamo il termine "matching" con un significato più ampio di quello adottato da Stevens, che si riferisce alle sole scale relative, su cui torneremo nel capitolo 15.

[2] L'esempio è stato presentato per la prima volta in D. Kahneman, *Pensieri lenti e veloci*, cit.

[3] D. Kahneman, A. Tversky, *On the Psychology of Prediction*, in "Psychological Review", 80(1973), pp. 237-251.

[4] G.A. Miller, *The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information*, in "Psychological Review", 63(1956), n. 2, pp. 81-97.

[5] R.D. Goffin, J.M. Olson, *Is It All Relative? Comparative Judgments and the Possible Improvement of Self-Ratings and Ratings of Others*, in "Perspectives on Psychological Science", 6(2011), pp. 48-60.

15

Scale

Mettetevi nei panni di un giurato in un processo civile. Avete ascoltato il caso riassunto di seguito, su cui poi vi è stato chiesto di pronunciarvi esprimendo un giudizio.

> *Joan Glover vs. General Assistance*
> Joan Glover, una bambina di sei anni, ha subìto un lungo ricovero ospedaliero dopo aver ingerito un elevato numero di pasticche di Allerfree, un farmaco antiallergico da banco. Poiché l'overdose di farmaci ha indebolito il suo apparato respiratorio, per il resto della sua vita la bambina sarà più soggetta a patologie come asma ed enfisema. Il vasetto di Allerfree aveva un tappo di sicurezza salvabimbo inadeguato.
>
> Il produttore di Allerfree è General Assistance, una grande società di farmaci da banco con utili annuali che vanno dai cento ai duecento milioni di dollari. Un regolamento federale impone l'impiego di tappi di sicurezza in tutti i vasetti di medicinali, ma l'azienda ha sistematicamente disatteso la normativa utilizzando tappi con un tasso di malfunzionamenti molto più elevato rispetto agli altri presenti sul mercato. Un documento interno alla società riporta che «questo stupido regolamento federale, del tutto inutile, ci fa sprecare soldi», e dichiara che il rischio di incorrere in sanzioni è basso. Nel documento si aggiunge che, in ogni caso, «le pene per chi viola la legge sono bassissime: in sostanza ci verrà chiesto di migliorare la sicurezza dei nostri tappi in futuro». Pur avendo ricevuto un avvertimento da un funzionario della Food and Drug Administration americana, la compagnia ha deciso di non intraprendere alcuna azione correttiva.

Ora vi chiediamo di esprimere tre giudizi, riflettendo bene prima di rispondere.

Indignazione:
Quale delle seguenti definizioni esprime al meglio la vostra opinione in merito alle azioni dell'imputato? (Cerchiate il valore corrispondente.)

Del tutto accettabili		*Discutibili*		*Sconvolgenti*		*Assolutamente scandalose*
0	1	2	3	4	5	6

Intento punitivo:
Oltre alla richiesta di un risarcimento danni, quale pena andrebbe assegnata all'imputato? (Cerchiate il valore che si avvicina di più all'entità della pena che secondo voi sarebbe appropriata.)

Nessuna pena		*Pena lieve*		*Pena severa*		*Pena molto severa*
0	1	2	3	4	5	6

Danni: In aggiunta alla richiesta di un risarcimento, quali danni *punitivi* eventualmente l'imputato dovrebbe essere chiamato a pagare, come sanzione e come deterrente per l'imputato stesso e per altri, affinché non vengano compiute azioni simili in futuro? (Scrivete la vostra risposta qui sotto.)

$.....

La storia di Joan Glover è una versione leggermente ridotta di un caso impiegato in uno studio presentato da due di noi (Kahneman e Sunstein, insieme al nostro amico e collaboratore David Schkade) nel 1998.[1] Poiché nel presente capitolo ci soffermeremo su questo studio, abbiamo ritenuto utile sottoporvi uno dei casi trattati, che oggi riteniamo un esempio istruttivo di controllo del rumore, in quanto riprende molti dei temi di questo libro.

Il capitolo si concentrerà sulla *scala di risposta* in quanto fonte pervasiva di rumore: è possibile che le persone differiscano nei giudizi non perché sono in disaccordo nella sostanza, ma perché utilizzano una scala in modi

diversi. Nel valutare le prestazioni di un dipendente, potreste dire che su una scala da 0 a 6 siano da 4, che, secondo voi, è un voto piuttosto buono; qualcun altro potrebbe dire che, sulla stessa scala, le prestazioni del dipendente siano da 3, che, secondo lui, è un voto altrettanto buono. L'ambiguità nella formulazione delle scale è un problema diffuso, e sono state svolte molte ricerche sui problemi di comunicazione che sorgono da varie espressioni come: «oltre ogni ragionevole dubbio»,[2] «prove inconfutabili», «prestazioni eccezionali» e «improbabile eventualità».[3] I giudizi espressi in queste formule sono inevitabilmente affetti da rumore, perché vengono interpretati in maniera diversa da chi le pronuncia e da chi le ascolta.

Nello studio in cui si inseriva il caso di Joan Glover abbiamo osservato gli effetti di una scala ambigua in una situazione in cui poteva avere conseguenze molto gravi. Il tema dello studio era il rumore nei danni punitivi stabiliti da una giuria. Come potrete dedurre dalla terza domanda sul caso Glover, negli Stati Uniti (e in alcuni altri paesi) la legge consente ai giurati nelle cause civili di imporre danni punitivi a imputati le cui azioni sono risultate particolarmente oltraggiose. Tali danni si aggiungono ai risarcimenti, che sono pensati per permettere alla parte lesa di ripristinare la situazione preesistente al danno. Quando, come nel caso Glover, un prodotto ha provocato un danno e i querelanti hanno vinto la causa contro la società, questi ultimi si aggiudicheranno un risarcimento in denaro per coprire le spese mediche e le eventuali perdite salariali, ma potrebbero ricevere anche i danni punitivi, come avvertimento per l'imputato e per altre società affini. In questo caso il comportamento di General Assistance è stato ovviamente riprovevole, e ricade nel ventaglio di azioni per cui una giuria potrebbe imporre dei danni punitivi.

Un grosso problema relativo all'istituto dei danni punitivi riguarda la loro imprevedibilità: lo stesso illecito può essere punito con una richiesta di danni che vanno da importi modesti a cifre enormi. Avvalendoci della terminologia adottata in questo libro, diremmo che si tratta di un sistema affetto da rumore. Le richieste di danni punitivi vengono spesso negate e, anche quando sono accolte, di solito non aggiungono molto ai risarcimenti. Vi sono però notevoli eccezioni, e talvolta le cifre esorbitanti stabilite dai giurati appaiono sorprendenti e arbitrarie. Un esempio citato di frequente è la richiesta di danni punitivi pari a quattro milioni di dollari a una concessionaria di automobili per non aver rivelato al querelante che la sua nuova BMW era stata riverniciata.[4]

Nel nostro studio sui danni punitivi abbiamo chiesto a 899 partecipanti di valutare il caso di Joan Glover e nove altri casi simili in cui i querelanti avessero subìto un danno e fatto causa alla società ritenuta responsabile. Diversamente da voi, i partecipanti hanno risposto solo a una delle tre domande su indignazione, intento punitivo e danni pecuniari per tutti e dieci i casi; sono poi stati suddivisi in gruppi più piccoli, a ciascuno dei quali è stata assegnata una particolare versione di ogni caso. Le diverse versioni variavano nel danno subìto dal querelante e nel fatturato della società sotto accusa, per un totale di ventotto scenari. Il nostro obiettivo era duplice: comprovare una teoria sulla psicologia dei danni punitivi e indagare sul ruolo della scala monetaria (qui in dollari) come principale fonte di rumore in questo istituto giuridico.

L'ipotesi dell'indignazione

Il tema della determinazione di una giusta pena è da secoli oggetto di dibattito filosofico e giuridico. La nostra ipotesi era che una questione

ritenuta spinosa dai filosofi diventa in realtà semplicissima per la gente comune, che procede alla sostituzione di un interrogativo difficile con uno più facile. Quando vi si chiede quale pena andrebbe inferta a General Assistance, in realtà voi rispondete a una domanda più semplice, ovvero: «Quanto mi fa rabbia?». L'intensità della pena che assegnerete sarà poi correlata a quella della vostra rabbia.

Per provare la validità dell'ipotesi, abbiamo chiesto a diversi gruppi di partecipanti di rispondere o alla domanda sull'intento punitivo o a quella sull'indignazione, e abbiamo poi confrontato i punteggi medi delle risposte alle due domande nei ventotto scenari impiegati nello studio. A conferma della nostra ipotesi della sostituzione, la correlazione tra i punteggi medi dell'indignazione e dell'intento punitivo era pressoché perfetta: 0,98 (PC = 94%). Questa correlazione avalla l'ipotesi di partenza: il senso di indignazione è il principale determinante dell'intento punitivo.[5]

Ma non è l'unico. Avete notato qualcosa nella storia di Joan che ha attirato maggiormente la vostra attenzione quando avete valutato l'intento punitivo, rispetto a quando vi siete espressi riguardo all'indignazione? Se la risposta è sì, presumiamo si tratti del danno da lei subìto. Potete ritenere indegno un certo comportamento anche senza conoscerne le conseguenze; in questo caso, il comportamento di General Assistance senz'altro lo era. Al contrario, le intuizioni sull'intento punitivo hanno una componente retributiva, che, banalmente, coincide con la legge del taglione. Il desiderio di ritorsione spiega perché il tentato omicidio e l'omicidio siano trattati diversamente tanto dalla legge quanto dai giurati: un potenziale omicida che abbia la fortuna di mancare il suo bersaglio subirà una punizione meno severa.

Per scoprire se l'entità del danno procurato incide davvero sull'intento punitivo ma non sull'indignazione, abbiamo mostrato a diversi gruppi di

rispondenti due versioni del caso Glover e di altri: nella prima, quella presentata anche a voi, il danno era molto grave, mentre nella seconda, meno traumatica, Joan «ha dovuto trascorrere diversi giorni in ospedale e ora è spaventata dalla vista di pillole di qualsiasi tipo. Quando i genitori provano a somministrarle anche farmaci salutari come vitamine, aspirine o compresse per il raffreddore, piange in maniera incontrollata e dice di avere paura». Questa versione descrive un'esperienza drammatica per la bambina, ma un danno fisico molto più contenuto rispetto alle patologie a lungo termine descritte nel primo scenario. Come previsto, il punteggio medio dell'indignazione della prima e della seconda versione era quasi identico (4,24 per il danno grave, 4,19 per il danno lieve): l'indignazione si basa solo sul comportamento dell'imputato, non sulle relative conseguenze. Per contro, il punteggio medio dell'intento punitivo era pari a 4,93 per il danno grave e a 4,65 per il danno lieve, una differenza contenuta ma statisticamente attendibile. Il valore mediano del risarcimento economico richiesto era di due milioni di dollari nella prima versione e di un milione nella seconda. Risultati simili sono stati ottenuti per diversi altri casi.

Questi dati sottolineano una caratteristica fondamentale del processo di giudizio: l'effetto impercettibile del compito di giudizio nella ponderazione di diversi aspetti di un caso. I partecipanti che hanno attribuito un punteggio al proprio intento punitivo e alla propria indignazione non erano consapevoli che in quel momento stavano prendendo posizione sulla questione filosofica relativa alla giustizia retributiva, né sapevano di stare pesando i vari fattori coinvolti nel caso. Eppure hanno assegnato un peso prossimo allo zero al danno fisico mentre valutavano la propria indignazione, e un peso significativo allo stesso fattore quando si è trattato di determinare la pena. Ricordate che i partecipanti hanno letto una sola

versione della storia: la loro assegnazione di una pena più elevata al danno più grave non derivava da un confronto esplicito, ma era il risultato di un'operazione automatica di matching delle due condizioni. Le risposte dei partecipanti erano basate più sul pensiero veloce che su quello lento.

Rumore nelle scale

Il secondo obiettivo dello studio era scoprire perché i danni punitivi siano soggetti a rumore. L'ipotesi di partenza era che i giurati in genere concordano sul grado di severità della punizione da infliggere all'imputato, ma differiscono ampiamente tra loro nel modo in cui traducono il proprio intento punitivo in una cifra su una scala economica.

Per come è strutturato, lo studio ci permette di confrontare il livello di rumore nei giudizi sui medesimi casi all'interno di tre scale: indignazione, intento punitivo e risarcimento danni in dollari. Per misurare il rumore applichiamo il metodo già utilizzato per analizzare i risultati del controllo del rumore tra i giudici federali del capitolo 6, ovvero presumiamo che la media dei giudizi individuali di un caso possa essere considerata un valore giusto, non affetto da bias. (Partiamo da questo presupposto ai fini dell'analisi, ma precisiamo che potrebbe essere errato.) In un mondo ideale, tutti i giurati che impiegano una certa scala esprimerebbero giudizi unanimi su ogni caso; ogni deviazione dal giudizio medio va considerata come un errore, e simili errori sono alla base del rumore sistemico.

Nel capitolo 6 abbiamo anche osservato che il rumore sistemico può essere scomposto in rumore di livello e rumore strutturale. In questo caso, il primo è la variabilità nel grado generale di severità tra i diversi giurati, mentre il secondo è la variabilità nella risposta di un giurato nei singoli casi

rispetto alla sua stessa media. Pertanto, possiamo scomporre la varianza complessiva dei giudizi in tre elementi:

Varianza dei giudizi = Varianza delle giuste pene + (Rumore di livello)^2 + (Rumore strutturale)^2

Questa analisi, che scompone la varianza dei giudizi in tre termini, è stata condotta in forma separata per i tre giudizi di indignazione, intento punitivo e risarcimento in dollari.

I risultati sono illustrati nella figura 13.[6] La scala meno soggetta a rumore è quella dell'intento punitivo, dove al rumore sistemico è imputabile il 51% della varianza, quindi giudizi rumorosi e giuste pene si equivalgono. La scala dell'indignazione è decisamente più affetta da rumore, che qui arriva al 71%, mentre quella del risarcimento è di gran lunga la peggiore: addirittura il 94% della varianza dei giudizi è dovuto al rumore!

Le differenze sono impressionanti perché le tre scale, in termini di contenuto, sono quasi identiche. Abbiamo già visto come i valori giusti di indignazione e intento punitivo fossero quasi perfettamente correlati, come suggerito dall'ipotesi dell'indignazione. I punteggi dell'intento punitivo e del risarcimento rispondono esattamente alla stessa domanda – ovvero quanto andrebbe punita General Assistance – in unità di misura diverse. Come spieghiamo, allora, le grandi differenze illustrate dalla figura 13?

Probabilmente converremo sul fatto che la scala dell'indignazione non sia molto precisa. Certo, si può parlare di comportamento «del tutto accettabile», ma se esiste un limite alla rabbia che possono suscitare General Assistance o altri imputati, quel limite è piuttosto vago. Cosa vuol

dire che un comportamento è «assolutamente scandaloso»? La mancanza di chiarezza del valore più alto della scala rende inevitabile la presenza di un certo grado di rumore.

L'intento punitivo è più specifico: «pena severa» è una formula più precisa di «assolutamente scandaloso», perché una «pena molto severa» deve rientrare nel limite massimo prescritto dalla legge. Magari vorrete assegnare il massimo della pena al colpevole, ma non potete, per esempio, chiedere una condanna a morte per l'amministratore delegato di General Assistance e di tutto il direttivo. (O almeno lo speriamo.) La scala dell'intento punitivo è meno ambigua perché il suo valore massimo è specificato con maggior chiarezza; è inevitabile, quindi, che sia anche meno soggetta a rumore.

Indignazione e intento punitivo sono stati misurati su scale di valutazione simili, definite in modo più o meno chiaro attraverso indicazioni verbali, mentre la scala del risarcimento appartiene a una classe diversa, e molto più problematica.

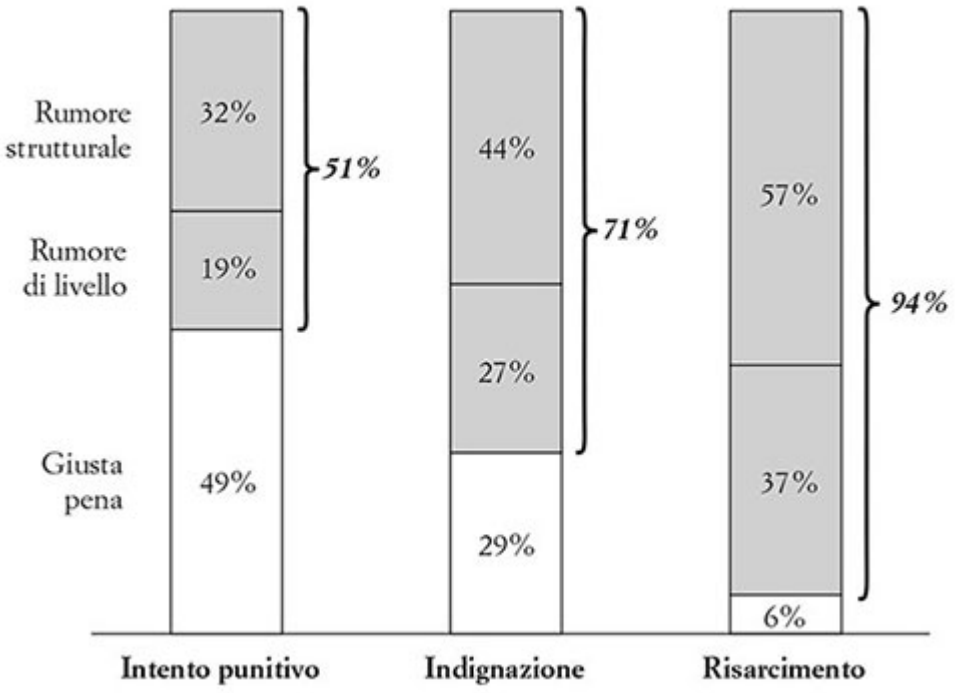


Figura 13. Componenti della varianza del giudizio

Dollari e ancoraggi

Il titolo del nostro articolo accademico ne riassume l'idea fondamentale: *Shared Outrage and Erratic Awards: The Psychology of Punitive Damages* ("Indignazione condivisa e risarcimenti imprevedibili: la psicologia dei danni punitivi"). Tra i giurati coinvolti nel nostro esperimento vi era un discreto livello di accordo nel punteggio assegnato all'intento punitivo, per lo più spiegabile sulla base dell'indignazione. Tuttavia, la misura monetaria, che simulava più da vicino il contesto processuale, era soggetta a un livello di rumore inaccettabile.

Il motivo non è certo un mistero. Se provaste a indicare una somma monetaria precisa per i danni punitivi del caso Glover, avreste di sicuro l'impressione che la vostra scelta di una cifra sia sostanzialmente arbitraria. Questa impressione di arbitrarietà ci dà un'informazione importante, in quanto ci dice che altri prenderanno una decisione altrettanto arbitraria e del tutto diversa, e che tali giudizi saranno soggetti a un elevato grado di rumore. Questo, in effetti, è un tratto caratteristico della classe di scale a cui appartiene quella del risarcimento monetario.

Il celeberrimo psicologo di Harvard Stanley S. Stevens scoprì che, in maniera del tutto inattesa, le persone hanno le stesse intuizioni forti sul *rapporto* tra l'intensità di molte esperienze e atteggiamenti soggettivi.[7] Sono in grado di regolare una lampada in modo che risulti "il doppio più luminosa" di un'altra, e concordano sul fatto che l'impatto emotivo di una condanna a dieci mesi di detenzione non è affatto dieci volte peggiore di quello di una condanna a un mese. Stevens definì le scale che si basano su tali intuizioni *scale a rapporti equivalenti.*

È possibile capire che le nostre intuizioni su questioni monetarie si esprimono come rapporti osservando la facilità con cui comprendiamo frasi come: «Sara ha avuto un aumento del 60%!», oppure «Quel riccone del nostro vicino ha perso metà della sua ricchezza da un giorno all'altro». La scala in dollari dei danni punitivi è una scala a rapporti per la misurazione dell'intento punitivo. Come altre scale di questo tipo, ha uno zero assoluto (zero dollari), ma non un limite superiore.

Stevens scoprì che una scala a rapporti (come quella in dollari) può essere vincolata utilizzando un unico ancoraggio intermedio (in gergo "modulo"). Nei suoi esperimenti, Stevens esponeva gli osservatori a una certa luce, chiedendo loro di «attribuire un valore di 10 (o 50, o 200) alla sua luminosità, e su quella base assegnare un valore di luminosità ad altre fonti di luce». In linea con l'ipotesi di ricerca, i numeri assegnati dagli osservatori a luci di diversa intensità erano proporzionali all'ancoraggio arbitrario che gli era stato chiesto di adottare: un osservatore ancorato al valore 200 esprimeva giudizi venti volte superiori a chi era ancorato al valore 10, e anche la deviazione standard dei giudizi dell'osservatore era proporzionale all'ancoraggio.

Nel capitolo 13 abbiamo descritto un divertente esempio di ancoraggio in cui la disponibilità delle persone a pagare per un oggetto era fortemente influenzata dalla domanda posta in precedenza, ovvero se avrebbero pagato (in dollari) una somma pari alle ultime due cifre del loro numero di previdenza sociale. Ancora più sorprendente era il fatto che l'ancoraggio iniziale influenzasse anche la loro disponibilità a spendere per vari altri oggetti. I partecipanti che venivano indotti ad accettare di pagare una lauta somma per un mouse senza fili erano disposti a sborsare una cifra proporzionalmente più alta anche per una tastiera senza fili. A quanto pare, le persone sono molto più sensibili al valore *relativo* di merci

comparabili che al loro valore assoluto. Gli autori dello studio chiamarono questo effetto persistente di un unico ancoraggio "arbitrarietà coerente".[8]

Per comprendere l'effetto di un ancoraggio arbitrario nel caso di Joan Glover, immaginate che il testo posto all'inizio di questo capitolo includesse la seguente informazione:

> In un caso simile che ha visto coinvolta un'altra società farmaceutica, la bambina vittima dell'incidente ha subìto un trauma psicologico moderato. Sono stati chiesti danni punitivi pari a un milione e mezzo di dollari.

Notate subito che il problema di assegnare una pena a General Assistance è diventato d'un tratto molto più semplice; anzi, forse avete anche già pensato a un importo pecuniario. Ora esiste un moltiplicatore (o "rapporto equivalente") del risarcimento in dollari corrispondente al contrasto tra il danno grave procurato a Joan e quello lieve subìto dall'altra bambina. Peraltro, il singolo ancoraggio che avete letto (un milione e mezzo di dollari) è sufficiente per vincolare l'intera scala monetaria della pena. Ora vi sarà facile stabilire i danni per casi di maggiore e minore entità rispetto ai due fin qui considerati.

Se servono degli ancoraggi per formulare giudizi su una scala a rapporti, cosa accade quando non ne vengono forniti? Stevens arrivò a una risposta: in assenza di un'indicazione da parte dello sperimentatore, le persone sono costrette a compiere una scelta arbitraria al primo impiego della scala; da quel momento in poi, regolano i loro giudizi di conseguenza, impiegando quella prima risposta come ancoraggio.

Forse vi sarete accorti che il compito di stabilire i danni punitivi nel caso Glover è un esempio di attribuzione di un valore su una scala senza un ancoraggio. Come gli osservatori privi di ancoraggio coinvolti nell'esperimento di Stevens, avete preso una decisione arbitraria su quale fosse la pena corretta per General Assistance. I partecipanti del nostro

studio sui danni punitivi si trovavano di fronte allo stesso problema: anche loro erano costretti a prendere una prima decisione arbitraria sul primo caso sottoposto alla loro attenzione. Diversamente da voi, però, loro non solo hanno preso una decisione arbitraria, ma hanno stabilito danni punitivi per altri nove casi; questi nove giudizi, però, non erano arbitrari, perché potevano essere coerenti all'ancoraggio del primo giudizio e poi tra loro.

I risultati di Stevens indicano che l'ancoraggio prodotto dai singoli dovrebbe avere un grande effetto sui valori assoluti dei loro successivi giudizi monetari, ma nessun effetto sulle posizioni relative dei dieci casi: in altre parole, da una somma iniziale elevata deriveranno altri giudizi tutti proporzionalmente elevati, senza incidere sulle loro dimensioni relative. Questo ragionamento porta a una conclusione inaspettata: per quanto possano apparire irrimediabilmente affetti da rumore, i giudizi monetari in realtà riflettono gli intenti punitivi dei giudicanti. Per scoprire tali intenti, non dobbiamo fare altro che sostituire i valori assoluti in dollari con i rispettivi punteggi.

Per verificare questa idea abbiamo ripetuto l'analisi del rumore dopo aver sostituito ogni risarcimento monetario con la relativa posizione sulla scala dei dieci giudizi espressi da un singolo individuo. Il risarcimento più elevato aveva il punteggio 1, il secondo 2 e così via. La trasformazione dell'ammontare dei risarcimenti in punteggi elimina ogni errore di livello dei giurati, perché la distribuzione dei punteggi da 1 a 10 è identica per tutti, se si escludono gli occasionali valori *ex aequo*. (Nel caso ve lo stiate chiedendo, c'erano più versioni del questionario perché a ogni individuo veniva chiesto di esprimersi su dieci scenari su ventotto. Abbiamo condotto analisi distinte per ogni gruppo di partecipanti che aveva risposto alle domande sugli stessi dieci scenari, e riportato la media.)

I risultati erano straordinari: il tasso di rumore nei giudizi calava dal 94% al 49% (figura 14). La trasformazione dei risarcimenti monetari in punteggi rivelava che i giurati erano in realtà sostanzialmente d'accordo sulla pena appropriata da assegnare nei diversi casi.[9] Anzi, i punteggi dei risarcimenti erano semmai leggermente *meno* affetti da rumore rispetto ai punteggi originari dell'intento punitivo.

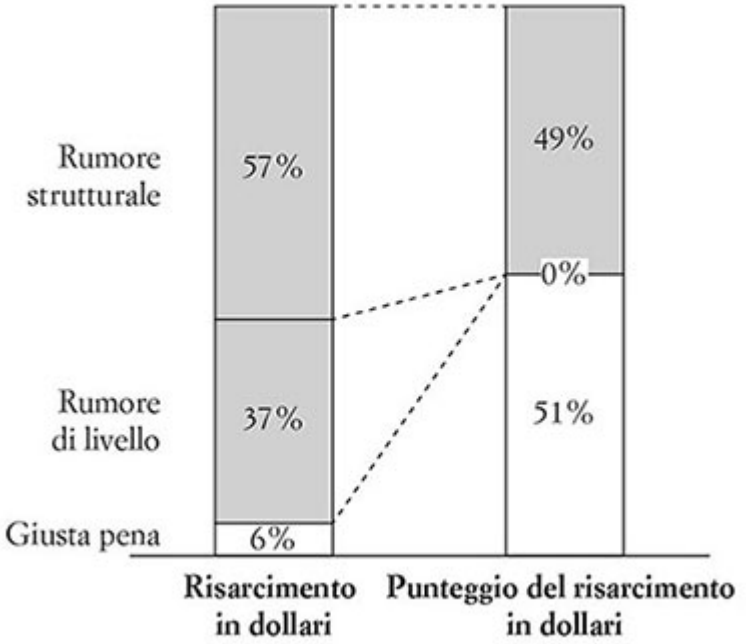


Figura 14. Rumore nel valore vs. rumore nel punteggio

Una spiacevole conclusione

I risultati sono coerenti con la teoria qui esposta: i risarcimenti in dollari stabiliti per tutti i casi erano ancorati al numero arbitrario scelto da ogni giurato per il primo caso da lui esaminato. Il punteggio relativo dei casi riflette gli atteggiamenti dei soggetti con discreta accuratezza, pertanto non è molto soggetto a rumore, ma i valori assoluti dei risarcimenti in

dollari sono sostanzialmente insensati, perché dipendono dal numero arbitrario stabilito per il primo caso.

Paradossalmente, il caso valutato dai giurati nei processi reali è il primo e l'unico che sono chiamati a giudicare: la prassi giuridica statunitense chiede alle giurie civili di stabilire un risarcimento in dollari per un solo caso, senza che possano usufruire di alcun ancoraggio di riferimento; anzi, la legge proibisce di fornire qualsiasi comunicazione ai giurati rispetto all'ammontare dei danni punitivi in altri casi. Il presupposto implicito della norma è che il senso di giustizia dei giurati li condurrà direttamente da un esame dell'illecito all'assegnazione della pena corretta, un assunto che non ha alcun senso sotto il profilo psicologico, perché postula una capacità che gli esseri umani non possiedono. Gli istituti giuridici dovrebbero riconoscere i limiti delle persone che amministrano la giustizia.

I danni punitivi sono un esempio estremo: raramente i giudizi professionali vengono espressi su scale così ambigue. Ma poiché l'ambiguità delle scale è un fatto comune, dallo studio sui danni punitivi si possono trarre due insegnamenti generali, applicabili in campo aziendale, accademico, sportivo, governativo eccetera. Il primo è che la scelta di una scala può fare una grande differenza nel livello di rumore dei giudizi, perché le scale ambigue sono soggette a rumore. Il secondo è che sostituire dei giudizi assoluti con dei giudizi relativi, ove possibile, probabilmente ridurrà il rumore.

A proposito di scale

«C'è molto rumore nei nostri giudizi. Forse perché interpretiamo diversamente la scala?»

«Perché non stabilire un ancoraggio che funga da punto di riferimento all'interno della scala?»

«Per ridurre il rumore, forse dovremmo sostituire i giudizi con dei punteggi?»

[1] D. Kahneman, D. Schkade, C. Sunstein, *Shared Outrage and Erratic Awards: The Psychology of Punitive Damages*, in "Journal of Risk and Uncertainty", 16(1998), pp. 49-86, [link.springer.com/article/10.1023/A:1007710408413]; Iid., *Assessing Punitive Damages (with Notes on Cognition and Valuation in Law)*, in "Yale Law Journal", 107(1998), n. 7, pp. 2071-2153. I costi della ricerca sono stati coperti da Exxon in un accordo una tantum, ma i ricercatori non sono stati pagati dalla società, la quale non aveva il controllo dei dati né una conoscenza anticipata dei risultati prima della loro pubblicazione su riviste accademiche.

[2] A. Keane, P. McKeown, *The Modern Law of Evidence*, Oxford University Press, New York 2014.

[3] A. Mauboussin, M.J. Mauboussin, *If You Say Something Is 'Likely,' How Likely Do People Think It Is?*, in "Harvard Business Review", 3 luglio 2018.

[4] *BMW v. Gore*, 517 U.S. 559(1996), [supreme.justia.com/cases/federal/us/517/559].

[5] Per un approfondimento sul ruolo dell'emozione nei giudizi morali, vedi J. Haidt, *The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment*, in "Psychological Review", 108(2001), n. 4, pp. 814-834; J. Greene, *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*, Penguin Press, New York 2014.

[6] Considerato l'elevato livello di rumore di questi punteggi, vi sorprenderà l'altissima correlazione (0,98) tra i giudizi sull'indignazione e sull'intento punitivo, che sostiene l'ipotesi dell'indignazione. La sorpresa svanisce se ricordate che la correlazione viene calcolata tra le *medie* dei giudizi. Su una media di cento giudizi, il rumore (la deviazione standard dei giudizi) si riduce di dieci volte. Il rumore, dunque, cessa di essere un fattore rilevante quando vengono aggregati molti giudizi. Vedi il capitolo 21.

[7] S.S. Stevens, *Psychophysics: Introduction to Its Perceptual, Neural and Social Prospects*, John Wiley & Sons, New York 1975.

[8] D. Ariely, G. Loewenstein, D. Prelec, *'Coherent Arbitrariness': Stable Demand Curves Without Stable Preferences*, in "Quarterly Journal of Economics", 118(2003), n. 1, pp. 73-106.

[9] Una trasformazione in punteggi implica una perdita di informazione, perché non vengono preservate le distanze tra i giudizi. Poniamo che vi siano soltanto tre casi e un giurato avanzi una richiesta di danni rispettivamente di 10 milioni, 2 milioni e 1 milione di dollari. È chiaro che il giurato intende esprimere una differenza di intento punitivo maggiore tra i primi due casi che tra il secondo e il terzo. Una volta convertita in punteggi, tuttavia, la differenza sarà la stessa, cioè un salto di una sola posizione. Questo problema potrebbe essere risolto convertendo i giudizi in punti standard.

16

Schemi

Ricordate Julie, la bambina precoce di cui avete cercato di indovinare la media finale dei voti nel capitolo 14? Eccone una descrizione più completa:

> Julie era figlia unica. Suo padre era un avvocato importante, sua madre un architetto. Quando Julie aveva circa tre anni, suo padre contrasse una malattia autoimmune che lo costrinse a lavorare da casa. Trascorse molto tempo con lei e, con tanta pazienza, le insegnò a leggere. A quattro anni leggeva già speditamente. Suo padre cercò anche di insegnarle l'aritmetica, ma lei la trovava difficile. Alle elementari Julie era una brava alunna, ma bisognosa di affetto e malvista dalla classe; passava molto tempo da sola e si appassionò al birdwatching, perché amava restare a osservare gli uccelli insieme al suo zio preferito.
>
> I suoi genitori divorziarono quando Julie aveva undici anni, e lei andò in crisi: a scuola prendeva voti bassi e aveva frequenti scatti emotivi. Alle superiori riportò ottimi risultati in materie come la biologia e la scrittura creativa, e i suoi voti eccellenti in fisica sorpresero tutti. Ma trascurava quasi tutte le altre materie, e concluse quel ciclo di studi con un voto non molto elevato.
>
> Julie non venne ammessa nelle scuole più prestigiose e finì per iscriversi a una buona scuola pubblica, dove scelse di specializzarsi in studi ambientali. Nei primi due anni di università continuò a ricadere in schemi di rapporti spesso complicati e cominciò a fumare regolarmente marijuana. Alla fine del secondo anno, però, sviluppò un forte desiderio di iscriversi alla facoltà di medicina e iniziò a impegnarsi molto di più nello studio.

Secondo voi con quale media finale si è laureata Julie?

Problemi: facili e difficili

Ovviamente questo problema (chiamiamolo Julie 2.0) è molto più difficile rispetto alla versione precedente. Nel caso Julie 1.0 sapevate semplicemente che da bambina aveva cominciato a leggere all'età di quattro anni. Con quest'unico indizio, attingendo al potente strumento del matching, vi è subito venuta in mente una stima intuitiva della sua media finale dei voti.

Potreste ancora avvalervi del matching, se aveste diversi segnali che puntassero nella stessa direzione. Per esempio, nella descrizione di Bill, il ragioniere appassionato di musica jazz, tutte le informazioni in vostro possesso («con poca fantasia», «bravo in matematica», «non brillava nelle materie umanistiche») andavano a comporre un quadro coerente e stereotipico. Analogamente, se la maggior parte delle esperienze di vita del caso Julie 2.0 fossero coerenti con una storia di precocità e risultati eccellenti (magari perfino con qualche dato indicante prestazioni non superiori alla media), non trovereste questo compito così difficile. Quando le informazioni disponibili compongono un quadro coerente, il sistema 1 del pensiero veloce non ha difficoltà a trovare un significato: i problemi di giudizio semplici si risolvono facilmente, e quasi tutti saranno d'accordo sulla soluzione.

Ma il caso Julie 2.0 è ben diverso: a complicare il problema qui è la presenza di più segnali discordanti. Vi sono indicazioni di capacità e motivazione, ma anche di debolezza caratteriale e risultati mediocri. Questa storia sembra senza capo né coda, e non è detto che si riesca a darle un senso, perché i vari elementi non si inseriscono in un'interpretazione coerente. Non che questa incoerenza la renda meno realistica o addirittura implausibile: spesso la vita è più complicata di come ci piace raccontarla.

Segnali multipli contrastanti creano l'ambiguità che caratterizza i problemi di giudizio complessi, e che spiega anche perché i problemi

difficili sono più soggetti a rumore di quelli semplici. La regola è facile: se qualcosa può essere visto in più modi, le persone la vedranno diversamente. Potranno selezionare elementi differenti per costruire il nucleo della propria narrazione, quindi arriveranno a molte possibili conclusioni. Se avete avuto difficoltà a dare un senso alla storia di Julie 2.0, potete stare certi che altri lettori costruiranno storie diverse che giustificheranno giudizi lontani dai vostri. È questo il tipo di variabilità che produce rumore strutturale.

Quand'è che ci si sente sicuri del proprio giudizio? È necessario che siano soddisfatte due condizioni: la storia che portate avanti deve essere complessivamente coerente, e non devono esserci valide alternative. La coerenza complessiva si raggiunge quando tutti i dettagli dell'interpretazione adottata si inseriscono bene nella storia e si rafforzano a vicenda. Naturalmente potete anche raggiungere l'obiettivo della coerenza – seppure in maniera meno elegante – trascurando o liquidando tutto ciò che non quadra. Lo stesso vale per le interpretazioni alternative: l'esperto che sia riuscito a "risolvere" un problema di giudizio è in grado di dire non solo perché la sua spiegazione è corretta, ma anche perché le altre non lo sono; tuttavia, è anche vero che si può raggiungere lo stesso grado di sicurezza con un'interpretazione scorretta, se le alternative non vengono considerate o vengono occultate di proposito.

Se si guarda alla sicurezza in se stessi in quest'ottica, si può dedurre che la fiducia soggettiva nel proprio giudizio non è affatto una garanzia di accuratezza. Peraltro, l'occultamento delle interpretazioni alternative, un processo percettivo ben documentato, potrebbe portare al fenomeno che abbiamo definito "illusione di accordo" (vedi capitolo 2): se le persone non immaginano alternative possibili alle proprie conclusioni, tenderanno a presumere che anche altri osservatori debbano arrivare alle stesse

conclusioni.[1] Naturalmente sono pochi quelli che hanno la fortuna di essere molto sicuri di ogni giudizio che formulano, e ciascuno di noi ha sperimentato l'incertezza, forse anche pochi minuti fa di fronte al caso di Julie 2.0. Non tutti siamo sempre sicurissimi dei nostri giudizi, ma quasi sempre siamo più sicuri di quanto dovremmo.[2]

Rumore strutturale: stabile o temporaneo

Abbiamo definito l'errore strutturale come un errore nel giudizio individuale su un certo caso che non è spiegabile come la somma dei singoli effetti del caso stesso e di chi lo giudica. Un esempio estremo potrebbe essere quello del giudice di norma clemente, che diventa insolitamente severo quando si tratta di condannare un certo tipo di imputato (per esempio, chi ha commesso un'infrazione stradale). Oppure quello di un investitore di solito accorto che abbandona le sue consuete cautele di fronte al progetto di una startup promettente. È evidente che la maggior parte degli errori strutturali non è così estrema: avremo un errore strutturale di media entità quando un giudice clemente lo è meno del solito di fronte ai recidivi, oppure ancora più del solito con le giovani donne.

Gli errori strutturali sono dovuti a una combinazione di fattori temporanei e permanenti. I primi sono quelli già individuati come fonti di errore occasionale, per esempio il buonumore di un giudice al momento di esprimere una sentenza oppure un evento infausto verificatosi di recente che gli occupa la mente. I fattori permanenti, invece, sono elementi più stabili: si pensi all'insolito entusiasmo di un datore di lavoro di fronte a persone che hanno frequentato una certa università, o all'inusuale propensione di un medico a raccomandare il ricovero a pazienti affetti da

polmonite. È possibile descrivere l'errore in un singolo giudizio con una semplice equazione:

Errore strutturale =
Errore strutturale stabile + Errore (occasionale) temporaneo

Poiché l'errore strutturale stabile e l'errore (occasionale) temporaneo sono indipendenti e irrelati tra loro, possiamo impiegare la precedente equazione anche per analizzare la loro varianza:

$$\textbf{(Errore strutturale)}^2 =$$
$$\textbf{(Errore strutturale stabile)}^2 + \textbf{(Errore occasionale)}^2$$

Come abbiamo fatto per altre componenti dell'errore e del rumore, possiamo rappresentare graficamente l'equazione come somma dei quadrati costruiti sui lati di un triangolo rettangolo (figura 15):

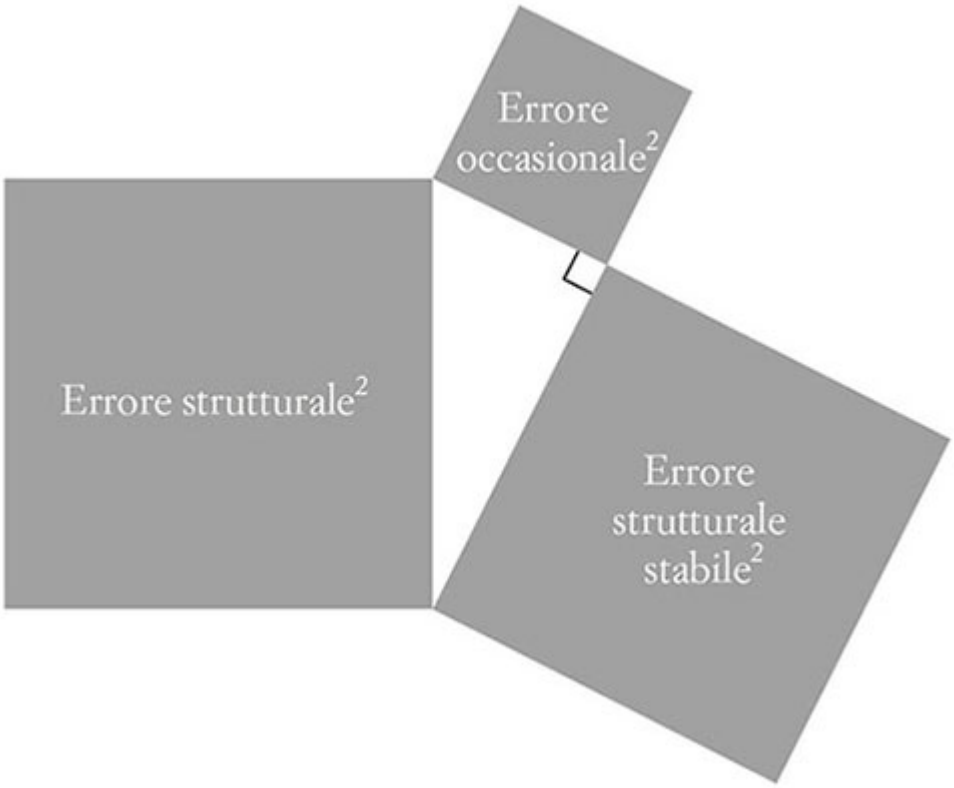


Figura 15. Scomposizione del rumore strutturale

Un caso semplice di rumore strutturale stabile è ravvisabile nei reclutatori che prevedono le prestazioni future dei dirigenti sulla base di un insieme di indici. Nel capitolo 9 abbiamo parlato di un "modello del giudice": il modello di un singolo reclutatore assegna un peso a ogni indice, che corrisponderà alla sua importanza nel giudizio di quel particolare reclutatore. Il peso varia a seconda dei reclutatori: per uno potrà contare di più la leadership, per un altro la capacità di comunicazione. Tali differenze creano una variabilità nella classificazione dei candidati da parte dei reclutatori, che è un tipico esempio di quello che chiamiamo errore strutturale stabile.

Le reazioni personali ai singoli casi possono anche creare schemi stabili ma molto specifici. Pensate a cosa vi ha indotto a prestare più attenzione a certi aspetti della storia di Julie piuttosto che ad altri. Certi dettagli del caso

potranno richiamare le vostre esperienze personali: magari in Julie c'è qualcosa che vi ricorda un vostro parente che era sempre sul punto di avere successo ma alla fine ha fallito per quelli che a vostro avviso erano profondi difetti caratteriali evidenti sin dall'adolescenza; oppure la sua storia potrebbe ricordarvi quella di un vostro amico che, dopo un'adolescenza tormentata, è riuscito a entrare alla facoltà di medicina e ora è uno specialista di successo. Le associazioni richiamate da Julie in persone diverse sono idiosincratiche e imprevedibili, ma probabilmente anche stabili: se aveste letto la sua storia la settimana scorsa, vi sarebbero venute in mente le stesse persone e avreste visto il racconto della sua vita nella stessa ottica personale.

Le differenze individuali nella qualità dei giudizi sono un'altra fonte di rumore strutturale. Immaginate un esperto di previsioni con poteri di chiaroveggenza di cui nessuno è consapevole (neanche lui stesso). Le sue accurate previsioni in molti casi si distaccheranno dalla media, e in assenza di informazioni sugli esiti finali tali deviazioni verranno considerate errori strutturali. Quando i giudizi non sono verificabili, perfino una maggiore accuratezza sembrerà rumore strutturale.

Questo tipo di rumore nasce anche dalle differenze sistematiche nella capacità di esprimere giudizi validi su dimensioni diverse di un caso. Consideriamo per esempio il processo di selezione nelle società sportive professionistiche, durante il quale gli allenatori potranno concentrarsi su varie abilità di gioco, i medici sulla predisposizione alle malattie, gli psicologi sulla motivazione e sulla capacità di recupero. Quando questi diversi specialisti valutano gli stessi giocatori, possiamo aspettarci un notevole grado di rumore strutturale. Analogamente, più professionisti che occupano lo stesso ruolo generico potranno essere più qualificati in certi aspetti del compito di giudizio che in altri. In questi casi il rumore

strutturale si può descrivere più come variabilità nelle conoscenze personali che come errore.

Quando i professionisti prendono decisioni per conto proprio, la variabilità delle loro qualifiche non è altro che rumore; se invece un direttivo ha la possibilità di creare gruppi che esprimano giudizi collettivi, la diversità delle qualifiche diventa un potenziale punto di forza, perché professionisti diversi tratteranno aspetti diversi del giudizio e si completeranno a vicenda. Discuteremo di questa opportunità, e di come coglierla, nel capitolo 21.

Nei precedenti capitoli abbiamo parlato delle due lotterie in cui si ritrovano coinvolti il cliente di una compagnia assicurativa o l'imputato assegnato a un certo giudice. Ora capiamo che la prima lotteria, in cui un professionista viene estratto a sorte in un gruppo di colleghi, non seleziona semplicemente il livello medio dei giudizi di quel soggetto (l'errore di livello), ma anche un insieme caleidoscopico e unico di valori, preferenze, credenze, ricordi, esperienze e associazioni che gli appartengono. Ogni volta che esprimiamo un giudizio, tutti noi partiamo dal nostro bagaglio di esperienze: ci portiamo dietro l'habitus mentale che ci siamo formati sul lavoro e il sapere che abbiamo acquisito dai nostri mentori, i successi che hanno rafforzato la nostra fiducia in noi stessi e gli errori che staremo attenti a non ripetere. Nel nostro cervello ci saranno le regole che ricordiamo, quelle che abbiamo dimenticato e quelle che abbiamo imparato a ignorare. Nessuno è identico a un altro in ognuno di questi aspetti; gli errori strutturali stabili di ognuno sono unici.

La seconda lotteria è quella che estrae a sorte il momento in cui esprimiamo i nostri giudizi, il nostro umore temporaneo e altre circostanze esterne che non dovrebbero condizionarci ma lo fanno. Questa è la lotteria che crea il rumore occasionale. Immaginate, per esempio, di aver letto un

articolo sul consumo di droga nei campus universitari appena prima di esaminare il caso di Julie. L'articolo menzionava la storia di uno studente brillante deciso a intraprendere studi di diritto con il massimo zelo, che non era riuscito a recuperare le lacune accumulate nei primi anni di università, quando faceva uso di droga. Poiché questa storia è fresca nella vostra mente, sarete portati a prestare più attenzione al consumo di marijuana di Julie nel valutare le sue possibilità di successo complessive. Tuttavia, è probabile che non ricordereste l'articolo se il caso di Julie vi venisse sottoposto tra un paio di settimane (e ovviamente non avreste saputo nulla dell'articolo se aveste letto la storia di Julie il giorno prima). L'effetto della lettura dell'articolo è temporaneo: è una fonte di rumore occasionale.

Come illustra questo esempio, non vi è una netta discontinuità tra il rumore strutturale stabile e la variante instabile che chiamiamo rumore occasionale. La differenza principale sta nella natura permanente o temporanea della sensibilità propria di una persona a certi aspetti del caso: quando a far scattare il rumore strutturale è un elemento radicato nelle nostre esperienze e nei nostri valori personali, possiamo aspettarci una struttura stabile che riflette la nostra unicità.

L'analogia della personalità

L'idea che il modo in cui le persone rispondono a particolari elementi o combinazioni di elementi sia unico non è così intuitiva. Per comprenderla, possiamo prendere in considerazione un'altra combinazione complessa di elementi che conosciamo molto bene: la personalità dei nostri conoscenti. Effettivamente, la circostanza di un giudice che si esprime su un caso dovrebbe essere considerata un'occorrenza particolare di un tema più

generale molto studiato nelle ricerche sulla personalità: come agisce una persona in una data situazione. Decenni di studi su questo tema più generale possono dirci qualcosa anche sul giudizio.

Da tempo gli psicologi cercano di capire e misurare le differenze individuali della personalità. Le persone si differenziano in molti modi; un vecchio studio condotto su un dizionario ha identificato ben diciottomila parole in grado di descrivere una persona.[3] Oggi il principale modello della personalità, chiamato Big Five, consta di cinque grandi raggruppamenti (estroversione, gradevolezza, coscienziosità, apertura mentale, nevroticismo), ciascuno dei quali copre una gamma di tratti distintivi. Un tratto di personalità è considerato un predittore di comportamenti reali: se qualcuno viene descritto come coscienzioso, ci aspettiamo di poter osservare un comportamento corrispondente (è puntuale, mantiene gli impegni e così via). E se Andrew ha un livello di aggressività più elevato di Brad, dovremmo osservare che, nella maggior parte delle situazioni, Andrew avrà un comportamento più aggressivo di Brad. In realtà, però, la validità dei tratti generali nella previsione di comportamenti specifici è piuttosto limitata: una correlazione di 0,30 (PC = 60%) sarebbe considerata alta.[4]

Il buonsenso ci dice che se il comportamento può essere motivato dalla personalità, è però anche fortemente influenzato dalle *situazioni*: in certi contesti nessuno è aggressivo, in altri lo sono tutti. Quando si tratterà di consolare un amico colpito da un lutto, né Andrew né Brad agiranno in maniera aggressiva, mentre a una partita di calcio entrambi mostreranno un certo grado di aggressività. Insomma, non ci sorprende che i comportamenti dipendano tanto dalla personalità quanto dalle situazioni.

A rendere ogni persona unica e interessante è che questo connubio di personalità e situazione non è una semplice operazione meccanica: le

situazioni che scatenano una maggiore o minore aggressività, per esempio, non sono le stesse per tutti. Anche se Andrew e Brad sono in media ugualmente aggressivi, non è detto che manifestino la stessa aggressività in ogni contesto. Magari Andrew lo è con i suoi pari ma non con i superiori, mentre il livello di aggressività di Brad non dipende dal grado gerarchico. Forse Brad è particolarmente incline all'aggressività quando viene criticato e insolitamente controllato quando subisce una minaccia fisica.[5]

Con ogni probabilità queste strutture di risposta caratteristiche resteranno stabili nel tempo, contribuendo in grande misura a quella che consideriamo la *personalità* di un individuo, pur non prestandosi a una descrizione per tratti generali. Andrew e Brad forse riceverebbero lo stesso punteggio in un test dell'aggressività, ma la loro struttura di risposta a determinati fattori scatenanti e circostanze è unica. Due persone che possiedono un certo tratto in uguale misura – per esempio, sono ugualmente ostinate o generose – dovrebbero essere descritte con due distribuzioni comportamentali che abbiano la stessa media ma non necessariamente la stessa struttura di risposta a situazioni diverse.

Ora vi sarà chiaro il parallelismo tra questi concetti legati alla personalità e il modello di giudizio presentato in precedenza. Le differenze di livello tra i giudici corrispondono alle differenze dei punteggi relativi ai tratti di personalità, i quali rappresentano una media dei comportamenti in molteplici situazioni. I casi sono analoghi alle situazioni: il giudizio di una persona su un particolare problema è prevedibile solo fino a un certo punto sulla base del livello medio di quella persona, così come gli specifici comportamenti sono prevedibili solo fino a un certo punto sulla base dei tratti di personalità. La classificazione degli individui in virtù dei loro giudizi varia in maniera sostanziale da caso a caso, perché le persone differiscono nelle proprie reazioni agli elementi e alle combinazioni di

elementi di ciascun caso. Il tratto distintivo di un individuo che esprime giudizi e decisioni è la struttura peculiare della sua sensibilità a certi elementi, e la conseguente struttura peculiare del suo giudizio sui casi.

L'unicità della personalità di solito viene elogiata, ma in questo libro ci interessano i giudizi professionali, in cui la variazione è problematica e il rumore è un errore. Il senso dell'analogia è che il rumore strutturale nei giudizi non è casuale, anche se difficilmente si riesce a spiegare, e anche se gli individui che esprimono giudizi peculiari non sono in grado di spiegarli.

A proposito del rumore strutturale

«Sembri sicuro delle tue conclusioni, ma questo non è un problema semplice: ci sono segnali che puntano in direzioni diverse. E se avessi tralasciato le interpretazioni alternative dei fatti?»

«Abbiamo esaminato lo stesso candidato, e di solito siamo altrettanto esigenti nei colloqui, eppure siamo arrivati a due giudizi completamente diversi. Da dove nasce questo rumore strutturale?»

«L'unicità della nostra personalità è ciò che ci rende innovativi e creativi, interessanti e coinvolgenti. Quando parliamo di giudizi, però, questa unicità non è un bene.»

tanti altri libri cercando dasolo

[1] R. Blake, N.K. Logothetis, *Visual competition*, in "Nature Reviews Neuroscience", 3(2002), pp. 13-21; M.A. Gernsbacher, M.E. Faust, *The Mechanism of Suppression: A Component of General Comprehension Skill*, in "Journal of Experimental Psychology: Learning, Memory, and Cognition", 17(1991), pp. 245-262; M.C. Stites, K.D. Federmeier, *Subsequent to Suppression: Downstream Comprehension Consequences of Noun/Verb Ambiguity in Natural Reading*, in "Journal of Experimental Psychology: Learning, Memory, and Cognition", 41(2015), pp. 1497-1515.

[2] D.A. Moore, D. Schatz, *The three faces of overconfidence*, in "Social and Personality Psychology Compass", 11(2017), n. 8, articolo e12331.

[3] Lo studio di Allport e Odbert (1936) sul lessico inglese relativo alla personalità è citato in O.P. John, S. Strivastava, *The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives*, in L. Pervin, O.P. John (a cura di), *Handbook of Personality: Theory and Research*, 2ª ed., Guilford, New York 1999.

[4] I.W. Eisenberg *et al.*, *Uncovering the structure of self-regulation through data-driven ontology discovery*, in "Nature Communications", 10(2019), p. 2319.

[5] W. Mischel, *Toward an integrative science of the person*, in "Annual Review of Psychology", 55(2004), pp. 1-22.

17

Le fonti di rumore

A questo punto ci auguriamo che converrete con noi che dove c'è giudizio, c'è rumore. E ci auguriamo anche che, nel vostro caso, non sia più di quanto pensiate. Questo mantra sul rumore ci ha dato la spinta per avviare il nostro progetto, ma negli anni la nostra riflessione sul tema ha subìto un'evoluzione. Riprendiamo ora le principali lezioni che abbiamo imparato sulle componenti del rumore, sulla loro rispettiva importanza nel quadro generale e sul ruolo del rumore nello studio del giudizio.

Le componenti del rumore

La figura 16 mostra una rappresentazione grafica complessiva delle tre equazioni presentate nei capitoli 5, 6 e 16, ovvero tre successive scomposizioni dell'errore:

- l'errore nelle sue componenti del bias e del rumore sistemico;
- il rumore sistemico nelle sue componenti del rumore di livello e del rumore strutturale;
- il rumore strutturale nelle sue componenti dell'errore strutturale stabile e del rumore occasionale.

Ecco dunque come l'errore quadratico medio può essere scomposto nei quadrati del bias e delle tre componenti del rumore trattate in precedenza:[1]

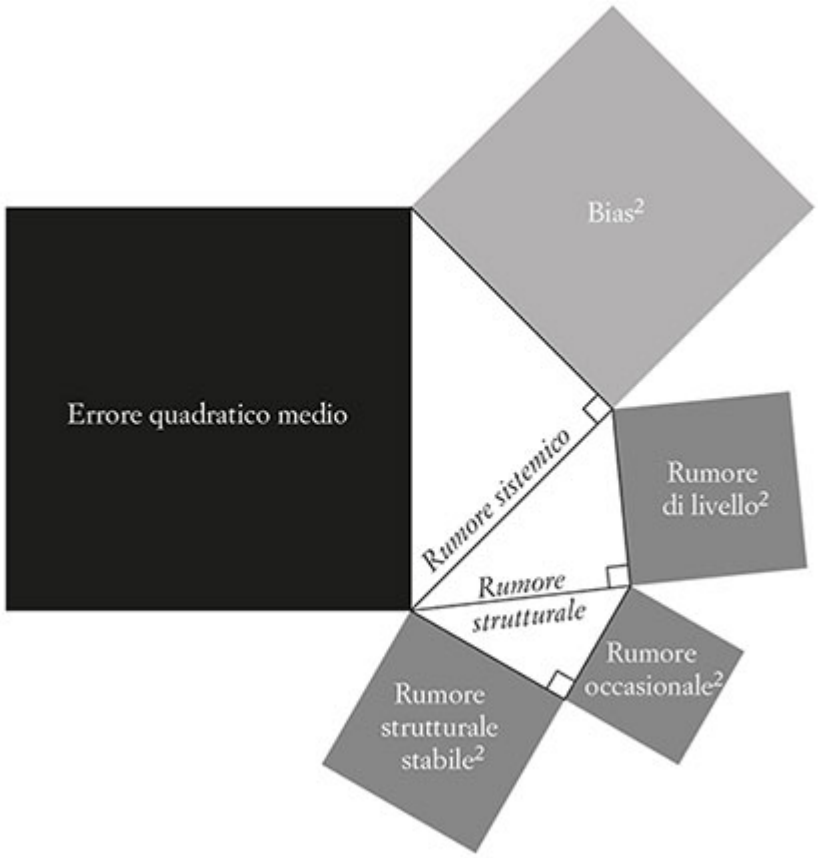


Figura 16. Errore, bias e componenti del rumore

All'inizio della nostra ricerca ci siamo concentrati sul rispettivo peso del bias e del rumore nell'errore totale, e siamo arrivati a concludere che il rumore è una componente dell'errore spesso superiore al bias, che quindi valeva la pena di analizzare nel dettaglio.

Le nostre riflessioni iniziali sugli elementi costitutivi del rumore sono state guidate dalla strutturazione di complessi controlli del rumore, in cui a più persone veniva chiesto di esprimere giudizi individuali su più casi. Lo studio sui giudici federali e quello sui danni punitivi ne sono due esempi. I dati raccolti in questi studi hanno condotto a solide stime del rumore di livello. D'altro canto, poiché ogni partecipante giudicava ogni caso una volta sola, non c'era modo di stabilire se l'errore residuo, che abbiamo definito errore strutturale, fosse temporaneo o stabile. Nell'analisi

statistica tradizionale l'errore residuo viene comunemente classificato come termine di errore e trattato come casuale. In sostanza, secondo l'interpretazione classica il rumore strutturale consterebbe esclusivamente di rumore occasionale.

Questa interpretazione convenzionale del rumore strutturale come errore casuale ha posto per molto tempo un freno alle nostre riflessioni. Sembrava naturale concentrarsi sul rumore di livello, ovvero le differenze sistematiche tra giudici severi e clementi, o tra esperti che avanzano previsioni ottimistiche e pessimistiche. Eravamo inoltre incuriositi dalle evidenze in merito all'influenza sui giudizi delle circostanze irrilevanti e temporanee che producono il rumore occasionale.

Pian piano i dati ci hanno indotto a concludere che i giudizi soggetti a rumore di persone diverse sono in gran parte determinati da qualcosa che non è né un bias generale dell'individuo, né un elemento temporaneo o casuale: le persistenti reazioni personali di particolari individui a una molteplicità di fattori, che determinano le loro reazioni a specifici casi. Abbiamo così deciso che i nostri assunti di partenza sulla natura temporanea del rumore strutturale andavano accantonati.

Anche se volevamo stare attenti a non trarre eccessive generalizzazioni da quello che è pur sempre un insieme limitato di esempi, gli studi che abbiamo elaborato, nel complesso, indicano che il rumore strutturale stabile è in realtà più significativo delle altre componenti del rumore sistemico. Poiché raramente si ha un quadro complessivo delle componenti dell'errore all'interno del medesimo studio, occorre una sorta di triangolazione per formulare questa conclusione provvisoria. Ecco allora quel che sappiamo e quel che non sappiamo.

Stabilire le dimensioni delle componenti

Innanzitutto abbiamo diverse stime del peso relativo del rumore di livello e del rumore strutturale, e, nel complesso, sembra che il secondo sia più rilevante del primo. Nella compagnia assicurativa del capitolo 2, per esempio, le differenze tra i sottoscrittori nella media dei premi fissati ammontavano solo al 20% del rumore sistemico totale; il restante 80% era rappresentato dal rumore strutturale. Tra i giudici federali del capitolo 6, il rumore di livello (le differenze nella severità media) rappresentava poco meno della metà del rumore sistemico totale, mentre il rumore strutturale era la componente maggiore. Nell'esperimento sui danni punitivi l'entità complessiva del rumore sistemico variava notevolmente a seconda della scala impiegata (intento punitivo, indignazione o risarcimento in dollari), ma la porzione di rumore strutturale al suo interno era grosso modo costante: ammontava al 63%, al 62% e al 61% del rumore sistemico totale in tutte e tre le scale adottate nello studio. Altri studi che esamineremo nella parte 5, in particolare quelli sulle decisioni legate al personale, sono in linea con questa conclusione provvisoria.

Il fatto che in questi studi generalmente il rumore di livello non sia la componente maggiore del rumore sistemico è già un dato importante, perché il rumore di livello è l'unica forma di rumore che le organizzazioni (talvolta) riescono a monitorare senza condurre un apposito controllo. Quando i casi vengono assegnati in maniera più o meno casuale ai singoli professionisti, le differenze nel livello medio delle loro decisioni attestano la presenza del rumore di livello. Gli studi sugli uffici dei brevetti, per esempio, mostrano ampie differenze nella propensione media degli esaminatori ad assegnare brevetti, con conseguenti effetti sull'incidenza dei contenziosi brevettuali.[2] Analogamente, la diversa propensione di ogni responsabile dei servizi sociali a dare in affido i minori ha conseguenze a lungo termine sul benessere di questi ultimi.[3] Tali osservazioni si basano

unicamente su una stima del rumore di livello; ma se il rumore strutturale è superiore al rumore di livello, allora questi dati già sconvolgenti sottostimano l'ampiezza del problema del rumore almeno della metà. (Esistono eccezioni a questa regola provvisoria: la scandalosa variabilità nelle decisioni dei giudici che gestiscono le richieste di asilo è quasi certamente dovuta più al rumore di livello che al rumore strutturale, che sospettiamo sia comunque rilevante.)[4]

Il prossimo passo consisterà nell'analizzare il rumore strutturale separandolo nelle sue due componenti. Vi sono buoni motivi per presumere che il rumore strutturale stabile, più che quello occasionale, sia la componente dominante. Il controllo effettuato sulle condanne dei giudici federali ne è un ottimo esempio. Partiamo dalla possibilità estrema che tutto il rumore strutturale sia temporaneo. Se così fosse, le condanne sarebbero instabili e incoerenti nel tempo, in una misura che riteniamo implausibile: dovremmo aspettarci che la differenza media nei giudizi *sullo stesso caso* espressi *dallo stesso giudice* in diverse occasioni sia di circa 2,8 anni.[5] La variabilità delle condanne medie tra i vari giudici è già sconcertante, ma la stessa variabilità nelle condanne di un singolo giudice in diverse occasioni sarebbe grottesca. Sembra più ragionevole concludere che i giudici differiscono nelle loro reazioni a diversi imputati e a diversi reati, e che tali differenze siano decisamente personali ma stabili.

Per quantificare in maniera più precisa quanto rumore strutturale sia stabile e quanto occasionale, servono studi in cui gli stessi giudici esprimono due valutazioni indipendenti su uno stesso caso, ma come abbiamo osservato, in genere ottenere due giudizi indipendenti in questo tipo di studi è impossibile, perché è difficile garantire che il secondo giudizio su un caso sia davvero indipendente dal primo. Specialmente

quando si tratta di un giudizio complesso, vi è un'alta probabilità che l'individuo riconosca il problema e replichi quello originario.

Un gruppo di ricerca di Princeton guidato da Alexander Todorov ha elaborato tecniche sperimentali molto ingegnose per risolvere questo problema.[6] I ricercatori hanno reclutato come partecipanti alcuni operatori di Amazon Mechanical Turk, un sito in cui gli iscritti forniscono servizi a breve termine per conto di programmatori – per esempio rispondere a questionari –, e vengono pagati sulla base del tempo dedicato a espletare la richiesta. In un esperimento hanno sottoposto ai partecipanti delle immagini di volti (generati da un programma informatico, ma in tutto e per tutto indistinguibili da quelli di persone reali), chiedendo loro di valutarli sulla base di vari attributi come la gradevolezza e l'affidabilità; a distanza di una settimana, gli stessi volti sono stati poi riproposti agli stessi rispondenti.

In questo esperimento sarebbe lecito aspettarsi un consenso inferiore rispetto a giudizi professionali come quelli dei giudici. Tutti potrebbero essere d'accordo sul fatto che certe persone sono molto attraenti e altre molto sgradevoli, ma entro un intervallo significativo ci aspettiamo che le reazioni ai volti siano per lo più idiosincratiche. In effetti, non c'è stato un grande accordo tra gli osservatori: nei punteggi sull'affidabilità, per esempio, le differenze tra le immagini erano in grado di spiegare solo il 18% della varianza dei giudizi. Il restante 82% dipendeva dal rumore.

È altrettanto lecito aspettarsi una minore stabilità in queste risposte, perché la qualità dei giudizi espressi da partecipanti pagati per rispondere a delle domande online spesso è sostanzialmente inferiore rispetto a quella riscontrata in contesti professionali. Ciò nonostante, la componente maggiore di rumore era il rumore strutturale stabile, mentre la seconda per dimensione era il rumore di livello, ovvero le differenze tra gli

osservatori nei punteggi medi sull'affidabilità. Il rumore occasionale, per quanto comunque sostanzioso, era la componente minore.

I ricercatori giunsero alle stesse conclusioni quando chiesero ai partecipanti di esprimere altri giudizi, per esempio sulle loro preferenze tra più auto o alimenti, o su aspetti più simili a quelli che definiamo giudizi professionali. In una replica dello studio sui danni punitivi analizzato nel capitolo 15, i partecipanti espressero una valutazione del proprio intento punitivo in dieci casi di lesione personale, in due occasioni diverse a distanza di una settimana; anche qui, il rumore strutturale stabile era la componente più elevata. In tutti questi studi solitamente gli individui non sono d'accordo gli uni con gli altri, ma restano piuttosto stabili nei propri giudizi. Questa «coerenza priva di consenso», per usare le parole dei ricercatori, fornisce una chiara prova della presenza di rumore strutturale stabile.

Il dato più emblematico sul ruolo delle strutture stabili ci giunge dal grande studio sui giudici chiamati a decidere sul rilascio su cauzione menzionato nel capitolo 10.[7] In una parte di questo straordinario lavoro di ricerca, gli autori hanno creato un modello statistico che simulava fino a che punto ogni giudice impiegasse le informazioni disponibili per decidere se concedere o no la libertà provvisoria. Hanno elaborato modelli personalizzati di 173 giudici e poi impiegato queste simulazioni per prendere decisioni in merito a circa 141 833 casi, generando 173 decisioni per ciascun caso, per un totale di oltre ventiquattro milioni di decisioni.[8] Su nostra richiesta gli autori hanno gentilmente condotto un'analisi particolare, separando la varianza dei giudizi in tre componenti: la "vera" varianza delle decisioni medie per ciascun caso, il rumore di livello dato dalle differenze tra i giudici nella propensione a concedere la libertà provvisoria, e il rumore strutturale residuo.

Quest'analisi è pertinente al nostro discorso perché il rumore strutturale, secondo la misurazione condotta in questo studio, è interamente stabile. La variabilità casuale del rumore occasionale non è rappresentata, perché si tratta di un'analisi di *modelli* che prevedono la decisione di un giudice, e di conseguenza vengono incluse solo le regole individuali di previsione comprovatamente stabili.

Le conclusioni non lasciavano dubbi: questo rumore strutturale stabile era quasi quattro volte superiore al rumore di livello (il primo ammontava al 26%, il secondo al 7% della varianza totale).[9] Le strutture di giudizio stabili, idiosincratiche e individuali che si riuscivano a identificare erano di gran lunga superiori alle differenze nel grado trasversale di severità.

Tutti questi dati sono in linea con la ricerca sul rumore occasionale esaminata nel capitolo 7: se l'esistenza del rumore occasionale è sorprendente e perfino inquietante, non vi sono indicazioni che la variabilità intrapersonale sia superiore alle differenze interpersonali. La componente più importante del rumore sistemico è quella che inizialmente avevamo trascurato: il rumore strutturale stabile, ovvero la variabilità tra i giudici nei propri giudizi su casi particolari.

Data la relativa scarsità di ricerche in questo campo, le nostre conclusioni sono provvisorie, ma riflettono comunque un cambiamento nella nostra riflessione sul rumore e su come contrastarlo. In linea di principio il rumore di livello – ovvero le semplici differenze trasversali tra i giudici – dovrebbe essere un problema relativamente facile da misurare e affrontare. Se vi sono esaminatori insolitamente "duri", assistenti sociali "cauti" o responsabili dei prestiti "avversi al rischio", le organizzazioni che li assumono potrebbero cercare di uniformare il livello medio dei loro giudizi. Le università, per esempio, affrontano questo problema chiedendo

ai docenti di attenersi a una distribuzione dei voti predeterminata all'interno di ogni classe.

Purtroppo ora ci rendiamo conto che concentrarsi sul rumore di livello porta a ignorare una gran parte di ciò che costituisce le differenze individuali. Il rumore è per lo più il prodotto non di differenze di livello, ma di interazioni: come si pongono giudici diversi di fronte a particolari imputati, insegnanti diversi di fronte a particolari studenti, assistenti sociali diversi di fronte a particolari famiglie, leader diversi di fronte a particolari visioni del futuro. Il rumore è più che altro un sottoprodotto della nostra specificità, della nostra "personalità di giudizio". Ridurre il rumore di livello resta un obiettivo importante, ma conseguire esclusivamente questo risultato lascerebbe il problema del rumore sistemico in gran parte irrisolto.

Spiegare l'errore

Abbiamo scoperto che c'è molto da dire riguardo al rumore, ma questo tema è quasi del tutto assente nella consapevolezza generale e nei discorsi sul giudizio e sull'errore. Per quanto siano evidenti la sua presenza e i molteplici meccanismi da cui è prodotto, il rumore viene citato raramente come fattore fondamentale del giudizio. Com'è possibile? Perché non chiamiamo mai in causa il rumore per spiegare i giudizi sbagliati, mentre condanniamo di continuo i bias? Perché è così insolito prendere in considerazione il rumore come fonte di errore, malgrado la sua onnipresenza?

La risposta a questa serie di interrogativi è che, benché la media degli errori (il bias) e la variabilità di questi ultimi (il rumore) rivestano un ruolo equivalente nell'equazione di errore, ci rapportiamo a loro in maniera

profondamente diversa, e il nostro modo abituale di dare un senso al mondo che ci circonda rende praticamente impossibile riconoscere il ruolo del rumore.

Nelle pagine precedenti abbiamo osservato che ci è facile dare un senso agli eventi col senno di poi, anche se non avremmo potuto prevederli prima che avvenissero. Nella valle della normalità, gli eventi non ci sorprendono e si spiegano facilmente.

Lo stesso vale per i giudizi. Come altri eventi, i giudizi e le decisioni ricadono quasi sempre nella valle della normalità: di solito non ci sorprendono. Per fare un esempio, i giudizi che producono risultati soddisfacenti sono normali, e raramente vengono messi in discussione. Quando il calciatore scelto per tirare un rigore segna un gol, quando l'intervento del cardiochirurgo riesce bene o una startup ha successo, presumiamo che i motivi che hanno spinto i decisori a compiere le loro scelte fossero giusti. Dopotutto, tali si sono dimostrati. Come qualsiasi altra delle tante storie che non ci sorprendono, una storia di successo si spiega da sé una volta che il risultato è noto.

Sentiamo, invece, la necessità di spiegare gli esiti anomali, sia quelli negativi sia, talvolta, quelli sorprendentemente positivi, come quando un assurdo azzardo societario dà ottimi risultati. Le spiegazioni che fanno appello all'errore o al fiuto individuale sono molto più in voga di quanto meriterebbero: le grandi scommesse del passato vengono lette come colpi di genio o di follia una volta che si sa come sono andate a finire. Questa forte tendenza a dare la colpa o il merito a un agente per atti e risultati attribuibili alla fortuna o a circostanze oggettive costituisce un noto bias psicologico definito *errore fondamentale di attribuzione*. Un altro bias, il senno di poi, distorce i giudizi affinché esiti impossibili da prevedere appaiano facilmente pronosticabili a posteriori.

Non è difficile trovare spiegazioni per gli errori di giudizio: dopotutto, motivare un giudizio è più facile che risalire alle cause degli eventi. Possiamo sempre appellarci alle motivazioni che hanno spinto qualcuno a prendere una decisione e, se questo non basta, possiamo dare la colpa alla sua incompetenza. Negli ultimi decenni, inoltre, si ricorre comunemente a un'altra spiegazione: il bias psicologico.

Un nutrito corpus di ricerche nel campo della psicologia e dell'economia comportamentale ha individuato un lungo elenco di bias psicologici: la fallacia della pianificazione, l'eccessiva fiducia in se stessi, l'avversione alla perdita, l'effetto dotazione, il pregiudizio dello status quo, l'eccessiva svalutazione del futuro ("bias del presente") e molti altri ancora, compresi, naturalmente, i bias a favore o a sfavore di varie categorie di persone. Sono ben note le condizioni in cui è probabile che ciascuno di questi bias influenzi i giudizi e le decisioni, e sono stati elaborati diversi strumenti che permetterebbero a un osservatore di riconoscere sul momento l'affiorare del bias in un processo decisionale.

Un bias psicologico rappresenta una spiegazione causale legittima di un errore di giudizio solo nel caso si possa prevedere prima o individuare sul momento; se identificato solo dopo l'evento può comunque costituire una spiegazione utile, per quanto provvisoria, se permette di avanzare previsioni sul futuro. Per esempio, la sorprendente bocciatura di una candidata forte per una certa posizione può condurre a formulare un'ipotesi più generale sul bias di genere, che le nomine future della stessa commissione confermeranno o confuteranno. Consideriamo, invece, una spiegazione causale applicabile a un solo evento: «Se in quel particolare caso non ce l'hanno fatta, è perché saranno stati troppo sicuri di sé». Questa affermazione è assolutamente stupida, ma dà a chi la pronuncia la pia illusione di aver capito tutto. Il docente di economia aziendale Phil

Rosenzweig sostiene, non a torto, che nelle discussioni sui risultati commerciali si avanzino spesso spiegazioni insensate che chiamano in causa i bias.[10] Questa tendenza conferma quanto sia diffusa la necessità di trovare nessi causali per dare un senso all'esperienza.

Il rumore è statistico

Come abbiamo osservato nel capitolo 12, normalmente pensiamo in termini causali: tendiamo a concentrarci sul particolare per seguire e creare storie coerenti su casi individuali, in cui spesso i fallimenti sono attribuiti a errori, e gli errori a bias. La facilità con cui si possono spiegare i giudizi sbagliati non lascia spazio al rumore nel nostro esame degli errori.

L'invisibilità del rumore è una diretta conseguenza del pensiero causale. Il rumore è un dato intrinsecamente statistico: diventa visibile solo quando pensiamo in termini statistici a un insieme di giudizi simili. A quel punto, però, diventa difficile non notarlo: è la variabilità rilevata nelle statistiche retrospettive sulle sentenze di condanna e sui premi assicurativi; è l'intervallo di possibilità di cui tenere conto nell'esprimere previsioni sui risultati futuri; è la dispersione dei tiri sul bersaglio. In un'ottica causale, il rumore non è da nessuna parte; in un'ottica statistica, è ovunque.

Purtroppo, non è facile assumere una prospettiva statistica. Non serve alcuno sforzo per attribuire delle cause agli eventi che osserviamo, mentre pensare in termini statistici è qualcosa che si impara, e anche allora richiede uno sforzo. Le cause sono naturali, le statistiche sono difficili.

Pertanto, vi è un marcato squilibrio nel nostro modo di vedere il bias e il rumore come fonti di errore. Se avete un'infarinatura di psicologia, probabilmente ricorderete il fenomeno per cui una figura vivace e piena di dettagli spicca su uno sfondo indistinto; la nostra attenzione si fissa sulla

figura anche quando è piccola rispetto allo sfondo. La relazione figura-sfondo è un'utile metafora delle nostre intuizioni sul bias e sul rumore: il bias è la figura che si impone alla nostra attenzione, il rumore è ciò che sta dietro, a cui non badiamo neanche. Così facendo, abbiamo scarsa consapevolezza di una grande pecca nel nostro giudizio.

A proposito delle fonti di rumore

«Notiamo facilmente le differenze nel livello medio dei giudizi, ma quanto è esteso il rumore strutturale che non vediamo?»

«Sei convinto che questo giudizio sia dovuto a un bias, ma diresti lo stesso se l'esito fosse stato diverso? E sapresti dire se è stato viziato da rumore?»

«Giustamente, puntiamo a ridurre i bias. Cerchiamo anche di ridurre il rumore, però.»

[1] Se non esiste una regola generale sulla scomposizione di bias e rumore, le proporzioni riportate in questa figura sono grosso modo rappresentative di alcuni degli esempi, reali o fittizi, da noi censiti. Nello specifico, in questa figura il bias e il rumore sono uguali (come lo erano nelle previsioni di vendita del caso GoodSell). Il quadrato del rumore di livello è pari al 37% del quadrato del rumore sistemico (come nello studio sui danni punitivi). Il quadrato del rumore occasionale qui illustrato è pari a circa il 35% del quadrato del rumore strutturale.

[2] Vedi i riferimenti dati nell'introduzione. M.A. Lemley, B. Sampat, *Examiner Characteristics and Patent Office Outcomes*, in "Review of Economics and Statistics", 94(2012), n. 3, pp. 817-827. Vedi anche I. Cockburn, S. Kortum, S. Stern, *Are All Patent Examiners Equal? The Impact of Examiner Characteristics*, documento 8980, giugno 2002, [www.nber.org/papers/w8980]; M.D. Frakes, M.F. Wasserman, *Is the Time Allocated to Review Patent Applications Inducing Examiners to Grant Invalid Patents? Evidence from Microlevel Application Data*, in "Review of Economics and Statistics", 99(2017), n. 3, pp. 550-563.

[3] J.J. Doyle Jr., *Child Protection and Child Outcomes: Measuring the Effects of Foster Care*, in "American Economic Review", 95(2007), n. 5, pp. 1583-1610.

[4] A.I. Schoenholtz, J. Ramji-Nogales, P.G. Schrag, *Refugee Roulette: Disparities in Asylum Adjudication*, cit.

[5] Questo valore è una stima tratta dai calcoli esposti nel capitolo 6, dove la varianza delle interazioni rappresenta il 23% della varianza totale. Partendo dall'assunto che le condanne abbiano una distribuzione normale, la differenza assoluta media tra due osservazioni selezionate in maniera casuale è pari a una deviazione standard di 1,128.

[6] J.E. Martinez, B. Labbree, S. Uddenberg, A. Todorov, *Meaningful 'noise': Comparative judgments contain stable idiosyncratic contributions* (manoscritto inedito).

[7] J. Kleinberg *et al.*, *Human Decisions and Machine Predictions*, cit.

[8] Il modello creava per ciascun giudice sia una gerarchia dei 141 833 casi sia una soglia oltre cui sarebbe stato concesso il rilascio su cauzione. Il rumore di livello riflette la variabilità di queste soglie, mentre il rumore strutturale riflette la variabilità della gerarchizzazione dei casi.

[9] G. Stoddard, J. Ludwig, S. Mullainathan, scambio di email con gli autori, giugno-luglio 2020.

[10] P. Rosenzweig, *Left Brain, Right Stuff: How Leaders Make Winning Decisions*, Public Affairs, New York 2014.

QUINTA PARTE

Migliorare i giudizi

Come può un'organizzazione migliorare i giudizi espressi dai suoi professionisti? E, in particolare, come può ridurne il rumore? Se doveste rispondere a queste domande, come le affrontereste?

Il primo passo necessario è fare in modo che l'organizzazione riconosca l'importanza del problema. A questo scopo, raccomandiamo di condurre un controllo del rumore (vedi appendice A per una descrizione dettagliata), ovvero un'analisi in cui più individui sono chiamati a giudicare gli stessi problemi, e la variabilità di questi giudizi esprime il livello di rumore presente. Vi saranno casi in cui questa variabilità potrà essere attribuita all'incompetenza: alcuni sanno di cosa parlano, altri no. In presenza di tali lacune (che siano generali o ristrette a certi tipi di casi), naturalmente diventa prioritario intervenire per colmarle. Tuttavia, come abbiamo visto, potrà esserci un alto grado di rumore perfino nei giudizi di professionisti competenti e preparati.

Se il livello di rumore sistemico è tale da richiedere un intervento, sostituire il giudizio con regole e algoritmi è una possibilità che sarebbe il caso di considerare, in quanto eliminerà del tutto il rumore. Ma le regole hanno i loro problemi (come vedremo nella parte 6), e anche i più entusiastici sostenitori dell'intelligenza artificiale concordano sul fatto che gli algoritmi non sono e non saranno, nell'immediato futuro, un sostituto universale del giudizio umano. Il compito di migliorare il giudizio, quanto mai urgente, costituisce il tema di questa parte del libro.

Un modo sensato per farlo, naturalmente, è quello di selezionare i migliori soggetti giudicanti possibili. Al tiro a segno alcuni tiratori hanno una mira particolarmente buona, e lo stesso vale per qualsiasi compito che richieda un giudizio professionale: i più qualificati saranno meno affetti tanto da rumore quanto da bias. Come trovare i decisori migliori, a volte, è ovvio: se devi risolvere un problema di scacchi chiedi a un grande maestro, non certo agli autori di questo libro. Ma, per la maggior parte dei problemi, i tratti dei soggetti più bravi a giudicare sono difficili da distinguere; ne parleremo nel capitolo 18.

Esamineremo poi gli approcci possibili per ridurre gli errori di giudizio. Tanto il bias statistico quanto il rumore implicano bias psicologici; come vedremo nel capitolo 19, sono stati effettuati molti tentativi per contrastare questi ultimi, con alcuni evidenti fallimenti e alcuni evidenti successi. Passeremo brevemente in rassegna le strategie di eliminazione dei bias e indicheremo un approccio promettente che, a quanto ci risulta, non è ancora stato indagato in maniera sistematica: chiedere a un *osservatore decisionale* designato di cercare segni diagnostici che possano indicare, in tempo reale, se un lavoro di gruppo sia affetto da uno o più dei classici bias. L'appendice B fornisce un esempio di checklist che un osservatore decisionale potrebbe utilizzare per l'individuazione dei bias.

Ci sposteremo poi sul nucleo principale di questa parte del libro: la lotta contro il rumore, e introdurremo il tema dell'*igiene decisionale*, ovvero l'approccio da noi consigliato per ridurre il rumore nei giudizi. Presenteremo dei casi di studio tratti da cinque aree diverse, in ciascuna delle quali esamineremo la diffusione del rumore e alcune terribili storie a essa connesse. Analizzeremo poi alcune azioni volte a ridurlo, non sempre andate a buon fine. In ogni campo, naturalmente, sono stati impiegati più

approcci, ma, per facilitare l'esposizione, ogni capitolo si soffermerà su un'unica strategia di igiene decisionale.

Nel capitolo 20 partiremo dal caso della scienza forense, che illustrerà l'importanza del *sequenziamento delle informazioni*. La ricerca di coerenza induce a formarsi delle prime impressioni basate sui dati limitati di cui si dispone e poi a convalidare il pregiudizio che ne risulta; diventa fondamentale, quindi, non essere esposti a informazioni irrilevanti nella prima fase del processo di giudizio.

Nel capitolo 21 torneremo sul caso delle previsioni, che metterà in luce il valore di una delle più importanti strategie di riduzione del rumore: l'*aggregazione di giudizi multipli indipendenti*. Questo principio, basato su quella che abbiamo definito la "saggezza della folla", si basa sulla media di giudizi multipli indipendenti, che garantisce la riduzione del rumore. Oltre alla media semplice, vi sono altri metodi per aggregare i giudizi, che verranno presentati attraverso l'esempio delle previsioni.

Il capitolo 22 offrirà una disamina dell'errore in medicina e delle azioni intraprese per ridurlo. Si rileverà l'importanza e la generale applicabilità di una strategia di riduzione del rumore a cui si è già accennato nell'esempio delle condanne penali: le *linee guida di giudizio*, che possono configurarsi come un potente meccanismo di riduzione del rumore, in quanto abbattono in maniera diretta la variabilità tra i decisori nei giudizi finali.

Nel capitolo 23 torneremo su una sfida costante della vita aziendale: le valutazioni delle prestazioni. I tentativi di ridurre il rumore in questo campo evidenziano quanto sia importante adottare una *scala condivisa fondata su una visione esterna*. Questa è una strategia di igiene decisionale fondamentale, per un semplice motivo: i giudizi implicano la trasposizione di un'impressione su una scala, e se soggetti diversi adottano scale diverse, si creerà rumore.

Il capitolo 24 affronterà il tema correlato ma distinto della selezione del personale, su cui sono state svolte molte ricerche da un secolo a questa parte. Si rimarcherà il valore di una strategia di igiene decisionale essenziale: la *strutturazione di giudizi complessi*. Per "strutturazione" qui intendiamo la scomposizione di un giudizio nelle sue componenti, la gestione del processo di raccolta dati per arrivare a input indipendenti e il differimento della discussione olistica e del giudizio finale finché non siano stati raccolti tutti questi input.

Sulla base degli insegnamenti tratti dalla selezione del personale, proporremo, nel capitolo 25, un approccio generale alla valutazione delle opzioni chiamato *protocollo a valutazioni intermedie* o MAP (dall'inglese *mediating assessments protocol*). Il MAP parte dalla premessa che «le opzioni sono come i candidati», per descrivere in maniera schematica come un processo decisionale strutturato, insieme alle altre strategie di igiene decisionale già menzionate, possa essere introdotto all'interno di un tipico processo che coinvolga decisioni singole e ricorrenti.

Una precisazione generale prima di addentrarci nella trattazione: sarebbe utilissimo poter specificare, e perfino quantificare, i probabili benefici di ogni strategia di igiene decisionale in vari contesti, e altrettanto utile sarebbe sapere quali sono le più vantaggiose e come confrontarle. Quando si riesce a controllare il flusso di informazioni, fino a che punto si riduce il rumore? Se l'obiettivo è ridurre il rumore, nella pratica quanti giudizi andrebbero aggregati? La strutturazione dei giudizi può essere preziosa, ma in quale misura lo è, di preciso, in diversi contesti?

Poiché il tema del rumore ha ricevuto scarsa attenzione, queste domande restano aperte per ricerche future. Ai fini pratici, i vantaggi dell'una o dell'altra strategia dipenderanno dal particolare contesto in cui le si impiega. Prendiamo l'adozione di linee guida: talvolta produrranno

enormi benefici (come vedremo in alcune diagnosi mediche); in altri contesti, tuttavia, i suoi vantaggi potranno essere ridotti, forse perché sin dall'inizio non c'è molto rumore oppure perché anche le migliori linee guida non riducono di molto l'errore. In ogni contesto, chi è chiamato a decidere dovrebbe aspirare ad arrivare a una comprensione più precisa dei possibili benefici di ciascuna strategia di igiene decisionale, e dei relativi costi, di cui parleremo nella parte 6.

18

Giudici migliori per giudizi migliori

Finora abbiamo parlato di giudici umani senza operare alcuna distinzione tra loro, eppure è ovvio che, in ogni compito che richieda un giudizio, alcuni faranno meglio di altri. Perfino un aggregato di giudizi che esprima la saggezza della folla sarà più accurato se quella folla è composta da persone più capaci.[1] È importante allora capire come identificare giudici migliori.

Tre sono gli elementi che contano: i giudizi sono meno affetti da rumore e da bias quando chi li esprime è molto preparato, è più intelligente e ha il corretto stile cognitivo. Detto in altri termini, la bontà di un giudizio dipende da ciò che sai, da quanto ragioni bene e da come ragioni. I buoni giudici sono tendenzialmente brillanti e competenti, ma hanno anche una grande apertura mentale e sono disposti a imparare dalle nuove informazioni.

Esperti veri ed esperti di rispetto

È quasi tautologico dire che le capacità dei giudici incidono sulla qualità dei loro giudizi. Per esempio, è molto più probabile che un radiologo qualificato diagnostichi correttamente una polmonite, e le previsioni di alcuni "superprevisori" riguardo a eventi globali superano immancabilmente quelle dei loro pari meno dotati. Se si mette insieme un gruppo di avvocati con una certa esperienza nello stesso settore,

probabilmente faranno previsioni simili, e corrette, sull'esito delle tipiche controversie legali che finiscono in tribunale. Chi è molto qualificato è meno soggetto a rumore, e anche meno incline ai bias.

Parliamo di persone davvero esperte del problema in questione, la cui superiorità è verificabile grazie ai dati disponibili relativi agli esiti dei loro giudizi. In linea teorica, possiamo scegliere un medico, un esperto di previsioni o un avvocato sulla base dei risultati che hanno avuto in passato. (Per ovvi motivi, nella pratica questo approccio può essere complicato: vi sconsigliamo di chiedere al vostro medico di famiglia di sottoporsi a un esame delle competenze.)

Come abbiamo osservato, però, molti giudizi non sono verificabili. Entro certi limiti, non possiamo facilmente sapere o definire una volta per tutte quale sia il valore reale a cui mirano i giudizi: la sottoscrizione dei premi assicurativi e le condanne penali ricadono in questa categoria, così come l'assaggio del vino, la valutazione degli elaborati accademici, le recensioni di film e libri, e innumerevoli altre forme di giudizio. Eppure alcuni professionisti in questi campi vengono definiti "esperti", e la fiducia che riponiamo nei loro giudizi è del tutto basata sul rispetto tributatogli dai loro pari. Possiamo definirli *esperti di rispetto*.

Questo termine non vuole essere irriguardoso. Il fatto che alcuni esperti non siano soggetti a una valutazione dell'accuratezza dei loro giudizi non è una critica, ma un dato di fatto in molti campi. Molti professori, studiosi e consulenti aziendali sono esperti di rispetto, e la loro credibilità dipende dall'alta considerazione che ne hanno i loro studenti, i loro pari o i loro clienti. In tutti questi campi, e in molti altri ancora, i giudizi di un professionista possono essere confrontati solo con quelli dei suoi pari.

In assenza di valori reali per definire chi abbia ragione e chi no, spesso diamo valore all'opinione degli esperti di rispetto anche quando sono in

disaccordo tra loro. Immaginiamo, per esempio, una tavola rotonda in cui vari analisti politici abbiano punti di vista nettamente diversi sull'origine di una crisi diplomatica e sui suoi sviluppi futuri. (Questo disaccordo non è affatto insolito, e anzi il dibattito non sarebbe molto interessante se fossero tutti d'accordo.) Ciascuno degli analisti ritiene che vi sia un'opinione corretta, e che la propria sia quella che vi si avvicina di più. Ascoltandoli, forse diversi di loro vi sembreranno ugualmente brillanti, e le rispettive argomentazioni ugualmente convincenti. A quel punto non potete sapere chi ha ragione (e forse non lo saprete neanche dopo, se le analisi non sono formulate come previsioni chiaramente verificabili), ma sapete che almeno alcuni di loro avranno torto, visto che sono in disaccordo gli uni con gli altri. Nonostante ciò, rispettate la loro competenza.

Consideriamo ancora un altro gruppo di esperti, che non ha nulla a che fare con il campo delle previsioni. Tre filosofi morali, tutti molto qualificati, si riuniscono in una stanza; il primo è un seguace di Immanuel Kant, il secondo di Jeremy Bentham, il terzo di Aristotele. I tre sono in forte disaccordo sul concetto di obbligo morale: il dibattito potrebbe vertere su se e quando sia legittimo mentire, sui diritti degli animali, sulla finalità delle sanzioni penali. Voi ascoltate con attenzione, ammirando la chiarezza e l'acume dei loro ragionamenti, e vi sentite più vicini all'uno o all'altro, ma li rispettate tutti e tre.

Perché? E in generale, perché qualcuno che è di per sé rispettato per la qualità dei propri giudizi decide di fidarsi di qualcun altro in quanto esperto, quando non vi è alcun dato che possa definirlo tale in termini oggettivi? Quand'è che qualcuno diventa un esperto di rispetto?

In parte ciò dipende dall'esistenza di norme condivise, o di una dottrina professionale: spesso gli esperti ottengono una qualifica da una comunità di professionisti, e vengono formati e affiancati all'interno delle loro

organizzazioni. I medici che completano l'internato e i giovani avvocati che imparano dai loro soci più esperti non si limitano ad apprendere i ferri del mestiere, ma vengono istruiti all'impiego di certi metodi e all'osservanza di certe norme.

Queste norme condivise danno ai professionisti un'indicazione degli input da considerare e di come elaborare e giustificare i propri giudizi finali. Nella compagnia assicurativa, per esempio, i periti liquidatori non avevano alcuna difficoltà a convenire sui singoli punti da inserire in una lista dei controlli da effettuare su una richiesta di risarcimento, né a giustificare il motivo per cui andassero inseriti.

Tale accordo, naturalmente, non ha impedito agli stessi periti di operare le valutazioni più disparate, perché la dottrina non specifica come procedere nel dettaglio. Non è una ricetta da seguire in maniera meccanica: al contrario, lascia spazio all'interpretazione. Gli esperti elaborano pur sempre dei giudizi, non dei calcoli, per questo il rumore è inevitabile. Anche dei professionisti con la medesima formazione che concordano sulla dottrina da applicare si discosteranno nella sua applicazione.

Oltre alla conoscenza di norme condivise, occorre anche avere esperienza. Esistono giovani prodigi degli scacchi, della musica classica o del lancio del giavellotto, perché i risultati convalidano il livello della loro prestazione; ma ai sottoscrittori, agli esaminatori di impronte digitali o ai giudici di solito occorre qualche anno di esperienza per essere credibili. Non esistono giovani prodigi della sottoscrizione assicurativa.

Un'altra caratteristica degli esperti di rispetto è la loro capacità di elaborare ed esporre i propri giudizi con grande sicurezza, e noi tutti tendiamo a fidarci di più di persone che si mostrano sicure di sé che di chi manifesta i propri dubbi. L'euristica della fiducia indica che, in un gruppo, chi ha fiducia in se stesso ha più peso degli altri, anche se tale fiducia è

immotivata.[2] Gli esperti di rispetto brillano nella costruzione di storie coerenti: la loro esperienza li aiuta a riconoscere degli schemi, a ragionare per analogia con i casi precedenti, e a formarsi e ad avallare rapidamente delle ipotesi. Insomma, riescono facilmente a integrare i fatti empirici in una narrazione coerente che ispira fiducia.

Intelligenza

Formazione, esperienza e fiducia in se stessi ci spingono a fidarci degli esperti di rispetto, ma non offrono alcuna garanzia sulla qualità dei loro giudizi. Come facciamo a sapere quali esperti daranno giudizi corretti?

Validi motivi inducono a credere che l'intelligenza sia associabile a giudizi migliori. L'intelligenza è correlata a buone prestazioni praticamente in tutti i campi; a parità di altri fattori, è associata non solo a più alti risultati accademici, ma anche a più alte prestazioni professionali.[3]

Vi è un fitto dibattito, non sempre lucido, su come vadano misurate l'intelligenza o la capacità mentale generale (definita anche GMA, dall'inglese *general mental ability*, termine che oggi si preferisce a quoziente d'intelligenza o QI). Esistono luoghi comuni molto radicati sulla natura innata dell'intelligenza,[4] ma in realtà i test misurano le capacità che si sono sviluppate, le quali in parte derivano da tratti caratteriali e in parte sono influenzate dall'ambiente, comprese le opportunità formative. Molte persone esprimono inoltre la propria preoccupazione per l'impatto negativo che una selezione basata sulla GMA potrebbe avere su certi gruppi sociali e sulla legittimità dell'impiego di test simili allo scopo della scelta dei candidati.

Occorre distinguere le preoccupazioni sull'uso di questi test dalla realtà del loro valore predittivo. Da quando, più di un secolo fa, l'esercito

americano ha cominciato a impiegare test sulla capacità mentale, migliaia di studi hanno misurato il nesso tra i punteggi ottenuti e le prestazioni successive. Da questa mole di ricerche emerge un messaggio chiarissimo: come è stato osservato, «la GMA riesce a predire tanto il livello occupazionale raggiunto quanto le prestazioni all'interno dell'occupazione scelta meglio di qualsiasi altra capacità, tratto o disposizione, e anche meglio dell'esperienza professionale».[5] Naturalmente contano anche altre capacità cognitive (su questo torneremo dopo) e molti tratti della personalità, compresa la coscienziosità e la *grinta*, ovvero la perseveranza e la passione nel perseguimento di obiettivi a lungo termine.[6] E vi sono anche varie forme di intelligenza che i testi della GMA non misurano, come quella pratica e la creatività. Gli psicologi e i neuroscienziati operano una distinzione tra intelligenza cristallizzata, ovvero la capacità di risolvere problemi a partire da un bagaglio di conoscenze sul mondo (comprese le operazioni aritmetiche), e intelligenza fluida, ovvero la capacità di risolvere problemi del tutto nuovi.[7]

Eppure, per quanto grossolana e limitata possa essere, la GMA, misurata da test standardizzati contenenti domande su problemi verbali, quantitativi e spaziali, resta di gran lunga il miglior predittore singolo di esiti di un certo livello. Come aggiunge l'articolo menzionato sopra, il potere predittivo della GMA è «superiore a quello della maggior parte degli strumenti elaborati dalla ricerca psicologica».[8] L'associazione tra capacità mentale generale e successo professionale aumenta, logicamente, con la complessità della professione: l'intelligenza conta di più per un ingegnere missilistico che per chi deve eseguire compiti più semplici. In lavori di elevata complessità, la correlazione osservata tra punteggi dei testi standardizzati e prestazioni lavorative si aggira intorno allo 0,50 (PC = 67%).[9]

Come abbiamo visto, una correlazione di 0,50 indica un valore predittivo molto forte per gli standard delle scienze sociali.[10]

Soprattutto nei dibattiti sui giudizi professionali qualificati, spesso alla rilevanza della misurazione dell'intelligenza viene sollevata l'importante obiezione che a esprimere questi giudizi probabilmente saranno tutti individui con un'elevata GMA. Medici, giudici o sottoscrittori esperti sono molto più istruiti della popolazione generale, ed è altamente probabile che riportino punteggi più alti in qualsiasi misurazione delle capacità cognitive. Si potrebbe presumere che l'elevata GMA li accomuni tutti, cioè che sia solo la chiave di accesso al gruppo dei grandi talenti, non l'origine delle differenze tra i vari talenti all'interno del gruppo.

Questa credenza, per quanto diffusa, è inesatta. Senz'altro la gamma dei valori di capacità mentale generale riscontrabili in una data professione è più ampia alla base che al vertice della piramide occupazionale: vi sono individui con elevata GMA nei livelli professionali più bassi, ma quasi nessuno con GMA inferiore alla media nelle file degli avvocati, dei chimici o degli ingegneri.[11] Da questo punto di vista, quindi, un'elevata capacità mentale si configura come una condizione necessaria per avere accesso a professioni prestigiose.

Tuttavia, questo dato non deve distoglierci dalle differenze nei risultati conseguiti *all'interno* di questi gruppi. Anche nella fascia dell'1% di popolazione con capacità cognitive più elevate (valutate all'età di tredici anni), esiti eccezionali sono fortemente correlati con la capacità mentale generale.[12] Rispetto a chi occupa il quartile più basso di questo 1%, chi è nel quartile più alto ha una probabilità doppia o tripla di conseguire un dottorato di ricerca, pubblicare un libro o ottenere un brevetto. In altre parole, la differenza in termini di GMA conta non solo tra il

novantanovesimo percentile e l'ottantesimo o il cinquantesimo, ma anche – e tanto! – tra un percentile di 99,88 e uno di 99,13.

Un'altra sorprendente illustrazione del legame tra capacità ed esiti arriva da uno studio del 2013 condotto sugli amministratori delegati delle cinquecento società che si sono piazzate nella celebre classifica "Fortune 500", che raggruppa le cinquecento aziende statunitensi con il maggior fatturato, e sui quattrocentoventiquattro miliardari d'America (la fascia degli americani più ricchi, pari allo 0,0001% della popolazione totale).[13] Com'era prevedibile, lo studio ha riscontrato che questi gruppi iperelitari sono composti da persone appartenenti alla fascia più dotata sul piano intellettuale. Ha anche evidenziato, però, che *all'interno* di questi gruppi elevati livelli di istruzione e competenza sono correlati a elevate retribuzioni (per gli amministratori delegati) e maggiore ricchezza (per i miliardari). Per inciso, personaggi famosi che sono diventati miliardari dopo aver lasciato l'università, come Steve Jobs, Bill Gates e Mark Zuckerberg, sono l'eccezione che conferma la regola: se un terzo degli americani adulti ha conseguito una laurea di primo livello, tra i miliardari la percentuale sale all'88%.

La conclusione è evidente: la GMA contribuisce notevolmente alla qualità delle prestazioni in occupazioni che richiedono giudizi, anche all'interno di un gruppo di individui molto qualificati. L'idea che vi sia una soglia oltre la quale cessa di avere un peso non è sostenuta dall'evidenza empirica. Questa conclusione, a sua volta, suggerisce chiaramente che, se i giudizi professionali non sono verificabili ma si presume che mirino al centro di un bersaglio invisibile, è più probabile che i giudizi che più vi si avvicineranno saranno quelli delle persone con un'alta GMA. Se bisogna scegliere qualcuno per esprimere un giudizio, insomma, ha più senso scegliere chi ha la capacità mentale più elevata.

Ma quest'argomentazione logica ha un grosso limite: poiché non è possibile somministrare a tutti dei test standardizzati, per individuare le persone con un'alta GMA ci si dovrà affidare a delle ipotesi. Inoltre, una capacità mentale generale elevata migliora le prestazioni su molti fronti, compresa l'abilità di convincere gli altri di avere ragione: le persone con un'alta GMA sono più inclini degli altri a esprimere giudizi migliori e a essere dei veri esperti, ma anche a fare colpo sui loro pari, guadagnarsi la fiducia degli altri e diventare esperti di rispetto in assenza di riscontri reali. Gli astrologi medievali saranno stati tra le persone con la più alta capacità mentale della loro epoca.

Può avere senso riporre la propria fiducia in persone che sembrano intelligenti e che sono in grado di articolare giustificazioni convincenti per i loro giudizi, ma questa strategia è insufficiente, e a volte perfino controproducente. Esistono altri modi per identificare i veri esperti? Le persone che esprimono i giudizi migliori si possono riconoscere da altri tratti?

Stile cognitivo

A prescindere dalle capacità mentali, le persone differiscono nello *stile cognitivo*, ovvero nell'approccio ai compiti di giudizio. Sono stati sviluppati molti strumenti per identificare gli stili cognitivi, molti dei quali sono correlati alla GMA (e gli uni agli altri), ma misurano elementi diversi.

Una di queste misure è il *test di riflessione cognitiva* (o CRT, dall'inglese *cognitive reflection test*), reso celebre da un famoso quesito: «Una mazza e una palla da baseball costano in tutto 1,10 dollari. La mazza costa un dollaro in più della palla. Quanto costa la palla?». Per misurare la riflessione cognitiva sono state proposte diverse altre domande, tra cui: «Se in una

corsa superi la persona che è al secondo posto, a che posto arrivi?».[14] Le domande di questo test puntano a misurare quanto una persona sia incline a scartare la prima risposta (sbagliata) che viene in mente («Dieci centesimi» nella prima domanda, e «Primo posto» nell'altra). Punteggi bassi nel test di riflessione cognitiva si associano a molti giudizi e credenze diffusi, come credere nei fantasmi, nell'astrologia e nella percezione extrasensoriale;[15] predicono se le persone abboccheranno a fake news palesemente infondate,[16] e sono perfino in rapporto con l'uso che le persone fanno dei propri smartphone.[17]

Il CRT viene ritenuto da molti uno strumento per misurare un concetto più ampio: la propensione a adottare processi di pensiero riflessivi anziché impulsivi.[18] In pratica, certe persone amano addentrarsi in attente riflessioni, mentre altre, di fronte agli stessi problemi, tendono a credere ai loro primi impulsi. Nella nostra terminologia, il CRT può essere considerato una misura della propensione ad affidarsi ai pensieri lenti del sistema 2 più che ai pensieri veloci del sistema 1.

Sono stati sviluppati anche altri test di autovalutazione per misurare questa propensione (tutti, naturalmente, interconnessi tra loro). Nella scala del bisogno di cognizione, per esempio, viene chiesto alle persone quanto prendono gusto ad arrovellarsi sui problemi.[19] Per ottenere un punteggio elevato, bisogna dirsi d'accordo con affermazioni come: «Tendo a stabilire degli obiettivi che possono essere raggiunti solo compiendo un notevole sforzo mentale», e in disaccordo con altre del tipo: «Pensare non mi diverte». Le persone con un elevato bisogno di cognizione tendono a essere meno soggette ai bias cognitivi.[20] Sono state riscontrate anche associazioni piuttosto bizzarre: chi evita le recensioni dei film in cui viene rivelata la trama, probabilmente ha un elevato bisogno di cognizione; chi

ha un basso bisogno di cognizione, al contrario, preferisce avere qualche anticipazione.[21]

Poiché questa scala si basa sull'autovalutazione, e poiché la risposta socialmente desiderabile è piuttosto ovvia, è legittimo avere qualche dubbio: chi sta cercando di fare una bella impressione difficilmente si dirà d'accordo con l'affermazione «Pensare non mi diverte». Per questo, altri test cercano di misurare le capacità invece di ricorrere ad autodescrizioni.

Un esempio è la scala ADMC (*Adult Decision Making Competence*),[22] che misura la propensione di una persona a compiere tipici errori di giudizio come l'eccesso di fiducia in se stessi o l'incoerenza nella percezione del rischio. Un altro è la scala Halpern Critical Thinking Assessment,[23] che si concentra sulle capacità di pensiero critico, tra cui la propensione al pensiero razionale e tutta una serie di abilità che possono essere apprese. In questo tipo di valutazione vengono poste domande come: «Immagina che un amico ti chieda un consiglio su quale dieta scegliere, tra una che promette di fargli perdere dieci chili e una che promette di fargliene perdere quindici. Quali informazioni ti occorrerebbero per poter scegliere tra le due diete?». Se rispondete, per esempio, che vorreste sapere quante persone hanno avuto la perdita di peso promessa e se hanno mantenuto il peso raggiunto per almeno un anno, otterrete un punteggio alto nell'applicazione del pensiero critico. Chi raggiunge un buon risultato in una di queste due scale sembra più portato a esprimere giudizi migliori nella vita reale: subisce meno eventi negativi dovuti a scelte sbagliate, come dover pagare una maggiorazione per aver restituito in ritardo un film a noleggio o affrontare una gravidanza indesiderata.

Appare ragionevole presumere che tutte queste misurazioni dello stile e delle competenze cognitive, come molte altre, in generale riescano a predire i giudizi. La loro rilevanza, tuttavia, sembra variare a seconda del

compito. Quando Uriel Haran, Ilana Ritov e Barbara Mellers cercarono di identificare gli stili cognitivi in grado di predire la capacità previsionale, scoprirono che il bisogno di cognizione non consentiva di prevedere chi si sarebbe sforzato di più per attingere a ulteriori informazioni.[24] Non riscontrarono neanche che il bisogno di cognizione fosse sistematicamente associato a prestazioni più elevate.

L'unica misura dello stile cognitivo o della personalità che ritennero affidabile per la previsione della performance di previsione era un'altra scala, sviluppata dal professore di psicologia Jonathan Baron per misurare la cosiddetta "apertura mentale attiva", ovvero l'attitudine a ricercare attivamente informazioni in grado di contraddire le proprie ipotesi preesistenti,[25] come le opinioni contrarie di altri e l'attenta ponderazione di nuovi dati che contraddicono vecchie credenze. Chi ha un pensiero attivamente aperto si riconosce in enunciati come il seguente: «Non precludersi la possibilità di farsi convincere da un argomento contrario è indice di un buon carattere», mentre non è d'accordo con le frasi «Cambiare idea è un indice di debolezza» e «L'intuito è la guida migliore quando si tratta di prendere una decisione».

In altre parole, se i punteggi relativi alla riflessione cognitiva e al bisogno di cognizione misurano la propensione a impegnarsi in pensieri lenti e attenti, l'apertura mentale attiva va oltre: è l'umiltà di chi è sempre consapevole che il proprio pensiero sia in via di definizione, e desidera correggersi. Nel capitolo 21 vedremo che questo stile di pensiero caratterizza i migliori previsori, che cambiano spesso idea e rivedono le loro credenze alla luce delle nuove informazioni. Vi sono dati interessanti che dimostrano come l'apertura mentale attiva sia una capacità che si può insegnare.[26]

Qui non intendiamo trarre conclusioni inderogabili su come selezionare individui che esprimeranno giudizi corretti in un dato ambito, ma da questa breve rassegna emergono due princìpi generali. Innanzitutto, è saggio riconoscere la differenza tra gli ambiti in cui le competenze possono essere confermate da un confronto con valori reali (come nelle previsioni meteorologiche) e gli ambiti di pertinenza degli esperti di rispetto. Un analista politico potrà anche sembrare eloquente e persuasivo, e un grande maestro di scacchi potrà apparire esitante e incapace di spiegare il ragionamento che sta dietro ad alcune delle sue mosse, eppure probabilmente dovremmo essere più scettici sul giudizio professionale del primo che su quello del secondo.

Per di più, alcuni giudici saranno migliori dei loro pari ugualmente qualificati e competenti, e in tal caso è probabile che siano meno soggetti a bias e a rumore. Tra i molti fattori in grado di spiegare queste differenze, spiccano l'intelligenza e lo stile cognitivo. Anche se nessuna misura o scala presa singolarmente potrà predire la qualità di giudizio in maniera impeccabile, è utile scegliere persone che cercano attivamente nuove informazioni in grado di contraddire le loro credenze precedenti, integrano con metodo tali informazioni nella loro prospettiva corrente e sono disposte a cambiare idea di conseguenza, e perfino desiderose di farlo.

Non è detto che la personalità di chi dimostra un'eccellente capacità di giudizio ricada nel classico stereotipo del leader risoluto. Spesso molti tendono a fidarsi e ad apprezzare i leader fermi e sicuri, che sembrano sapere subito per natura ciò che è giusto fare. Queste personalità ispirano fiducia, ma i dati indicano che, se l'obiettivo è quello di ridurre l'errore, è meglio che i leader (e non solo) restino aperti alle controargomentazioni, nella consapevolezza che potrebbero sbagliarsi. Se arriveranno a essere risoluti sarà solo alla fine di un processo, non all'inizio.

A proposito di giudici migliori

«Sei un esperto, ma i tuoi giudizi sono verificabili o sei un esperto di rispetto?»

«Dobbiamo scegliere tra due opinioni, ma non sappiamo niente sulle competenze di coloro che le hanno formulate, né sulle loro esperienze pregresse. Ci converrà seguire le indicazioni della persona più intelligente delle due.»

«L'intelligenza, tuttavia, è solo una componente: altrettanto importante è come si pensa. Forse dovremmo scegliere la persona più riflessiva e con la mente più aperta, non quella più brillante.»

[1] A.E. Mannes *et al.*, *The Wisdom of Select Crowds*, in "Journal of Personality and Social Psychology, 107(2014), n. 2, pp. 276-299; J. Dana *et al.*, *The Composition of Optimally Wise Crowds*, in "Decision Analysis", 12(2015), n. 3, pp. 130-143.

[2] B.D. Pulford, A.M. Colmna, E.K. Buabang, E.M. Krockow, *The Persuasive Power of Knowledge: Testing the Confidence Heuristic*, in "Journal of Experimental Psychology: General", 147(2018), n. 10, pp. 1431-1444.

[3] N.R. Kuncel, S.A. Hezlett, *Fact and Fiction in Cognitive Ability Testing for Admissions and Hiring Decisions*, in "Current Directions in Psychological Science", 19(2010), n. 6, pp. 339-345.

[4] *Ibid.*

[5] F.L. Schmidt, J. Hunter, *General Mental Ability in the World of Work: Occupational Attainment and Job Performance*, in "Journal of Personality and Social Psychology», 86(2004), n. 1, p. 162.

[6] A.L. Duckworth *et al.*, *Who Does Well in Life? Conscientious Adults Excel in Both Objective and Subjective Success*, in "Frontiers in Psychology", 3(2012). Riguardo alla grinta, vedi A.L. Duckworth *et al.*, *Grit: Perseverance and Passion for Long-Term Goals*, in "Journal of Personality and Social Psychology", 92(2007), n. 6, pp. 1087-1101.

[7] R.E. Nisbett *et al.*, *Intelligence: New Findings and Theoretical Developments*, in "American Psychologist", 67(2012), n. 2, pp. 130-159.

[8] F.L. Schmidt, J. Hunter, *Occupational Attainment*, cit., p. 162.

[9] N.R. Kuncel, S.A. Hezlett, *Fact and Fiction*, cit.

[10] Queste correlazioni derivano da meta-analisi che correggono, nelle correlazioni osservate, gli errori di misurazione relativi al criterio e alla restrizione dell'intervallo. Vi è un dibattito tra i ricercatori sull'eventualità che queste correzioni esagerino il valore predittivo della GMA. Tuttavia, poiché questi dibattiti metodologici riguardano anche altri predittori, in genere gli esperti concordano sul fatto che la GMA (insieme alle prove pratiche o *work samples*; vedi capitolo 24) sia il miglior predittore del successo professionale di cui a oggi disponiamo. Vedi N.R. Kuncel, S.A. Hezlett, *Fact and Fiction*, cit.

[11] F.L. Schmidt, J. Hunter, *Occupational Attainment*, cit., p. 162.

[12] D. Lubinski, *Exceptional Cognitive Ability: The Phenotype*, in "Behavior Genetics", 39(2009), n. 4, pp. 350-358.

[13] J. Wai, *Investigating America's Elite: Cognitive Ability, Education, and Sex Differences*, in "Intelligence", 41(2013), n. 4, pp. 203-211.

[14] K.S. Thomson, D.M. Oppenheimer, *Investigating an Alternate Form of the Cognitive Reflection Test*, in "Judgment and Decision Making", 11(2016), n. 1, pp. 99-113.

[15] G. Pennycook *et al.*, *Everyday Consequences of Analytic Thinking*, in "Current Directions in Psychological Science", 24(2015), n. 6, pp. 425-432.

[16] G. Pennycook, D.G. Rand, *Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning*, in "Cognition", 188(2018), pp. 39-50.

[17] N. Barr *et al.*, *The Brain in Your Pocket: Evidence That Smartphones Are Used to Supplant Thinking*, in "Computers in Human Behavior", 48(2015), pp. 473-480.

[18] N. Patel, S. Glenn Baker, L.D. Scherer, *Evaluating the Cognitive Reflection Test as a Measure of Intuition/Reflection, Numeracy, and Insight Problem Solving, and the Implications for Understanding Real-World Judgments and Beliefs*, in "Journal of Experimental Psychology: General", 148(2019), n. 12, pp. 2129-2153.

[19] J.T. Cacioppo, R.E. Petty, *The Need for Cognition*, in "Journal of Personality and Social Psychology", 42(1982), n. 1, pp. 116-131.

[20] S.M. Smith, I.P. Levin, *Need for Cognition and Choice Framing Effects*, in "Journal of Behavioral Decision Making", 9(1996), n. 4, pp. 283-290.

[21] J.E. Rosenbaum, B.K. Johnson, *Who's Afraid of Spoilers? Need for Cognition, Need for Affect, and Narrative Selection and Enjoyment*, in "Psychology of Popular Media Culture", 5(2016), n. 3, pp. 273-289.

[22] W. Bruine De Bruin *et al.*, *Individual Differences in Adult Decision-Making Competence*, in "Journal of Personality and Social Psychology", 92(2007), n. 5, pp. 938-956.

[23] H.A. Butler, *Halpern Critical Thinking Assessment Predicts Real-World Outcomes of Critical Thinking*, in "Applied Cognitive Psychology", 26(2012), n. 5, pp. 721-729.

[24] U. Haran, I. Ritov, B. Mellers, *The Role of Actively Open-Minded Thinking in Information Acquisition, Accuracy, and Calibration*, in "Judgment and Decision Making", 8(2013), n. 3, pp. 188-201.

[25] *Ibid.*

[26] J. Baron, *Why Teach Thinking? An Essay*, in "Applied Psychology: An International Review", 42(1993), pp. 191-214; Id., *The Teaching of Thinking: Thinking and Deciding*, 2ª ed., Cambridge University Press, New York 1994, pp. 127-148.

19

Eliminazione dei bias e igiene decisionale

Molti ricercatori e organizzazioni si sono posti l'obiettivo di eliminare i bias dai giudizi. Questo capitolo prende in esame i loro risultati principali.[1] Distingueremo diversi tipi di interventi di eliminazione dei bias e ci soffermeremo su uno in particolare che merita di essere indagato. Passeremo poi alla riduzione del rumore e introdurremo l'idea di igiene decisionale.

Eliminare i bias ex post ed ex ante

Per caratterizzare i due principali approcci all'eliminazione dei bias è utile tornare all'analogia della misurazione. Poniamo che sappiate che la vostra bilancia aggiunge, in media, duecento grammi al vostro peso: è tarata male (dunque è affetta da bias), ma non per questo è inutilizzabile. Potete seguire due strade per risolvere il problema: correggere ogni lettura scorretta del peso sottraendo duecento grammi, anche se rischia di diventare noioso (e potreste dimenticarvi di farlo), oppure regolare lo strumento e migliorarne l'accuratezza una volta per tutte.

Questi due approcci all'eliminazione dei bias nella misurazione sono in stretta analogia con gli interventi volti a eliminarli nei giudizi: operano o *ex post*, correggendo i giudizi dopo che sono stati espressi, oppure *ex ante*, intervenendo prima che vengano formulati.

L'eliminazione dei bias *ex post*, o "correttiva", viene spesso eseguita in maniera intuitiva. Poniamo che stiate coordinando un gruppo responsabile di un progetto e che i suoi componenti prevedano di completarlo in tre mesi. Potrete aggiungere un periodo cuscinetto di un mese o più per correggere un bias (la fallacia della pianificazione) che ritenete possa affliggere il giudizio del gruppo.

Questo tipo di correzione talvolta viene intrapreso in forma più sistematica. Nel Regno Unito il Dipartimento del Tesoro ha pubblicato un "libro verde" che costituisce una guida alla valutazione di programmi e progetti.[2] Il testo invita a tenere conto dei bias ottimistici applicando adeguamenti percentuali alle stime del costo e della durata di un progetto, che in teoria dovrebbero essere basati sui livelli storici di bias ottimistico di una determinata organizzazione; nel caso non fossero disponibili dati storici, il libro verde raccomanda di applicare percentuali di adeguamento generiche per ogni tipo di progetto.

Gli interventi *ex ante*, o "preventivi", si suddividono a loro volta in due grandi categorie. I più validi mirano a modificare l'ambiente in cui hanno luogo il giudizio o la decisione; tali modifiche, chiamate anche *nudges*, "spinte gentili", puntano a ridurre l'effetto dei bias o anche a sfruttare i bias per arrivare a una decisione migliore.[3] Ne è un esempio l'iscrizione automatica ai regimi previdenziali: pensata per superare l'inerzia, la procrastinazione o il bias ottimistico, garantisce che una parte della busta paga dei dipendenti venga trattenuta per la pensione, salvo diversa indicazione esplicita, e si è dimostrata estremamente efficace nel far crescere i tassi di partecipazione. Questo programma può essere affiancato da fondi pensionistici integrativi, con cui i dipendenti possono decidere di destinare alla pensione una certa percentuale dei loro stipendi futuri. L'iscrizione automatica può essere adottata in vari campi, per esempio per

l'impiego di energia pulita, la distribuzione di pasti scolastici gratuiti ai bambini poveri o vari altri piani assistenziali.

Altre spinte gentili incidono su aspetti diversi dell'architettura decisionale. Possono fare in modo che la decisione giusta sia anche la più semplice, per esempio riducendo la burocrazia necessaria per accedere alle cure per problemi mentali; oppure possono rendere evidenti certe caratteristiche di un prodotto o di un'attività, per esempio esplicitandone i costi nascosti. I negozi di alimentari e i siti web si potrebbero facilmente progettare in modo da spingere le persone a superare i propri bias: se i cibi sani venissero esposti negli scaffali più visibili, è probabile che più persone li acquisterebbero.

Un'altra strategia *ex ante* prevede l'offerta di una formazione ai decisori perché riescano a riconoscere e superare i propri bias. Alcuni di questi interventi, chiamati *boosting*, puntano a migliorare le capacità delle persone coinvolte, per esempio dando loro dei rudimenti di statistica.[4]

Istruire le persone in modo che superino i propri bias è un'impresa pregevole, ma anche più complicata di quanto non sembri. Naturalmente l'istruzione è utile: chi ha seguito corsi avanzati di statistica, per esempio, è meno incline a commettere errori nel ragionamento statistico.[5] Ma insegnare a evitare i bias è difficile. Decenni di ricerche hanno dimostrato che i professionisti che hanno imparato a farlo nella propria area di competenza spesso faticano ad applicare queste tecniche in altri campi. I meteorologi, per esempio, hanno imparato a evitare l'eccesso di fiducia nelle proprie previsioni: quando annunciano il 70% di probabilità di pioggia, piove, grosso modo, nel 70% dei casi; eppure possono peccare di eccessiva fiducia proprio come gli altri rispondendo a domande di cultura generale.[6] La difficoltà nell'imparare a superare i bias sta nel riconoscere

che un nuovo problema è simile a uno già visto altrove, e che un bias individuato in un certo campo probabilmente si ripresenterà in un altro.

Ricercatori e docenti hanno avuto un certo successo nell'uso di metodi di insegnamento non tradizionali per facilitare il riconoscimento dei bias. In uno studio condotto da Carey Morewedge e dai suoi colleghi della Boston University sono stati impiegati video informativi e "giochi seri". I partecipanti hanno imparato a riconoscere gli errori causati dal bias di conferma, dall'ancoraggio e da altri bias psicologici, e, dopo ogni gioco, hanno ricevuto un riscontro sugli errori commessi e hanno imparato a non ripeterli. I giochi (e, in misura minore, i video) hanno ridotto il numero di errori in un test svolto subito dopo e poi a distanza di otto settimane con domande simili.[7] In un altro studio, Anne-Laure Sellier e i suoi colleghi hanno scoperto che gli studenti di Business Administration coinvolti in un videogioco informativo per imparare a superare il bias di conferma hanno applicato quell'insegnamento alla risoluzione di un caso aziendale in un altro corso, anche se non era stato detto loro che i due esercizi erano collegati.[8]

Un limite dell'eliminazione dei bias

Che correggano i bias *ex post* o ne evitino gli effetti attraverso le *nudges* o il *boosting*, quasi tutti gli approcci per l'eliminazione dei bias hanno un elemento in comune: puntano a un preciso bias che presumono sia presente. Questo assunto, spesso ragionevole, si dimostra talvolta errato.

Riprendiamo l'esempio della programmazione di un progetto. È ragionevole presumere che l'eccesso di fiducia riguardi in generale tutti i gruppi di progetto, ma non si può essere certi che questo sia l'unico bias (né il principale) di un gruppo particolare. È possibile, poniamo, che il

capogruppo abbia avuto una brutta esperienza con un progetto simile e abbia imparato a essere particolarmente accorto nella programmazione: in questo caso, il gruppo presenterà l'errore opposto rispetto a quello che si presumeva di dover correggere. O ancora, il gruppo ha sviluppato le proprie previsioni in analogia con un altro progetto simile e ha assunto come ancoraggio il tempo occorso per completare quel progetto, o magari, prevedendo che avreste aggiunto un periodo cuscinetto alla sua stima, vi ha anticipati dandovi un'indicazione ancora più ottimistica della scadenza che in realtà prevede.

Oppure considerate una decisione di investimento: è certamente possibile che si riscontri un eccesso di fiducia circa le prospettive d'investimento, ma un altro potente bias, l'avversione alla perdita, ha l'effetto opposto, rendendo i decisori restii a rischiare di perdere il proprio esborso iniziale. O ancora, considerate una società che stanzi delle risorse in vari progetti: i decisori potrebbero essere tanto ottimisti sugli effetti delle nuove iniziative (ancora un eccesso di fiducia) quanto troppo timorosi di sottrarre risorse alle unità preesistenti (un problema causato dal "pregiudizio dello status quo", che, come indica il nome, è la tendenza a lasciare tutto com'è).

Come illustrano questi esempi, è difficile sapere esattamente quali bias psicologici subentrino in un giudizio. In qualsiasi situazione di una certa complessità potrebbero essercene in atto molteplici, che concorrono ad accrescere l'errore in una certa direzione oppure si controbilanciano, con conseguenze imprevedibili.

Di conseguenza, l'eliminazione dei bias *ex post* o *ex ante* – che, rispettivamente, corregge o evita bias psicologici specifici – è utile solo in determinate situazioni, ovvero quando la direzione generale dell'errore è nota e si manifesta chiaramente come bias statistico. È probabile che i tipi

di decisioni in cui si prevedono forti bias traggano beneficio da questi interventi: per esempio, la fallacia della pianificazione è un dato sufficientemente comprovato da legittimare interventi di eliminazione dei bias nelle pianificazioni troppo ottimistiche.

Il problema è che in molte situazioni non si conosce in anticipo la direzione più probabile che prenderà l'errore, come avviene in tutti i casi in cui l'effetto dei bias psicologici varia da un giudice all'altro, risultando sostanzialmente imprevedibile e producendo rumore sistemico. Per ridurre l'errore in simili circostanze, occorre estendere il campo di ricerca per cercare di individuare più di un bias psicologico per volta.

L'osservatore decisionale

Consigliamo di intraprendere questa ricerca dei bias né prima né dopo che sia stata presa una decisione, ma contestualmente. Certo, è raro che una persona sia consapevole dei propri bias nel momento in cui ne viene sviata, e questa mancanza di consapevolezza costituisce di per sé il noto *bias del punto cieco*: spesso le persone riconoscono i bias più facilmente negli altri che in loro stesse.[9] Noi riteniamo che si possano formare degli osservatori per individuare, in tempo reale, i segnali diagnostici indicanti che uno o più bias stanno incidendo sulle decisioni o sulle indicazioni di qualcuno.

Per illustrare il funzionamento di questo processo, immaginate un gruppo che cerca di arrivare a un giudizio complesso e gravido di conseguenze: per esempio, un governo che deve decidere come reagire a una pandemia o a una crisi, una riunione in cui dei medici stanno vagliando il miglior trattamento possibile per un paziente con sintomi complessi, un consiglio di amministrazione che deve decidere su un'importante mossa strategica. Ora immaginate un *osservatore decisionale* che sorvegli uno dei

suddetti gruppi e impieghi una checklist per diagnosticare se qualche bias lo stia deviando dal miglior giudizio possibile.

Quello dell'osservatore decisionale non è un compito facile, e senz'altro in certe organizzazioni non è neanche realistico: individuare i bias non serve a niente se i decisori ultimi non si impegnano per contrastarli, anzi, devono essere proprio loro ad avviare il processo di osservazione decisionale e a favorire il lavoro dell'osservatore. Vi sconsigliamo di autoinvestirvi di questo ruolo: non vi fareste degli amici né influenzereste gli altri.

Esperimenti informali indicano, però, che con questo approccio si possono fare dei veri progressi, o quantomeno è una strategia utile se vi sono le giuste condizioni, specialmente quando i leader di un'organizzazione o di una squadra si impegnano seriamente nell'impresa e quando gli osservatori decisionali vengono scelti bene, e non sono a loro volta soggetti a seri bias.

Gli osservatori di cui parliamo ricadono in tre categorie. In alcune organizzazioni questo ruolo può essere assunto da un supervisore: invece di monitorare solo i contenuti delle proposte avanzate da un gruppo di progetto, il supervisore potrebbe anche prestare attenzione al *processo* con cui vengono sviluppate e alle dinamiche di gruppo; in questo modo sarà pronto a individuare i bias che potrebbero aver viziato lo sviluppo della proposta.[10] Altre organizzazioni potrebbero assegnare a un membro di ogni gruppo di lavoro il ruolo di guardiano del processo decisionale, una sorta di "commissario anti-bias" che ricordi in tempo reale ai colleghi i bias che potrebbero sviarli. Lo svantaggio di questo approccio è che l'osservatore decisionale si pone all'interno del gruppo come una sorta di avvocato del diavolo, e potrebbe perdere presto capitale politico. Infine, altre organizzazioni potrebbero affidarsi a un facilitatore esterno, che ha il

vantaggio della prospettiva neutrale (e i relativi svantaggi in termini di conoscenza dell'ambiente e di costi).

Per essere efficaci, gli osservatori decisionali necessitano di una formazione e di strumenti adeguati, uno dei quali è una checklist dei bias che stanno cercando di scovare. I motivi per adottare una checklist sono evidenti: è noto che portino a decisioni migliori in contesti ad alto rischio, e sono particolarmente adatte quando si tratta di evitare di ripetere errori già commessi in passato.[11]

Facciamo un esempio. Negli Stati Uniti le agenzie federali devono stendere un'analisi di impatto della regolamentazione formale prima di emettere nuove norme o procedure che richiedano ingenti spese, come quelle per la depurazione dell'aria o dell'acqua, la riduzione delle morti sul lavoro, l'aumento della sicurezza alimentare, la risposta a una crisi della sanità pubblica, la riduzione delle emissioni di gas serra o il potenziamento della sicurezza nazionale. Un documento tecnico dal nome poco attraente (*OMB Circular A-4*), della lunghezza di circa cinquanta pagine, stabilisce i requisiti dell'analisi, chiaramente pensati per neutralizzare i bias: le agenzie devono spiegare perché la regolamentazione è necessaria, considerare alternative più o meno stringenti, valutare costi e benefici, presentare le informazioni in maniera imparziale ed effettuare attualizzazioni appropriate. In molte agenzie, però, i funzionari governativi non si sono attenuti ai requisiti di questo complesso documento (e forse non l'hanno neanche letto). Pertanto, gli uffici federali hanno steso una semplice checklist di una pagina e mezzo per ridurre il rischio che le agenzie ignorino o non rispettino i requisiti fondamentali.[12]

Per indicare come stendere una possibile checklist per l'individuazione dei bias, ne abbiamo acclusa una nell'appendice B.[13] Questa checklist generica è solo un esempio: ogni osservatore decisionale vorrà senz'altro

elaborarne una costruita *ad hoc* sulla base delle esigenze dell'organizzazione, per aumentarne la pertinenza e facilitarne l'adozione.[14] È importante ricordare che la checklist non è una lista esaustiva di tutti i bias che potrebbero condizionare una decisione, ma intende evidenziare i più frequenti e rovinosi.

L'osservazione decisionale, accompagnata da una checklist appropriata, può contribuire a limitare gli effetti dei bias. Anche se abbiamo visto dei risultati incoraggianti in alcune applicazioni informali su scala ridotta, non ci risulta che sia stata effettuata alcuna indagine sistematica sugli effetti di questo approccio o sui pro e i contro delle varie modalità di impiego possibili. Speriamo di stimolare nuove sperimentazioni, da parte di professionisti quanto di ricercatori, di questa pratica di eliminazione dei bias in tempo reale che coinvolge la figura dell'osservatore decisionale.

Come ridurre il rumore: l'igiene decisionale

I bias sono errori che spesso riusciamo a notare e anche a spiegare. Vanno in una certa direzione, e per questo una spinta gentile può limitarne gli effetti dannosi o uno sforzo per migliorare i giudizi può servire a contrastarne alcuni in particolare. Spesso, inoltre, sono anche visibili, motivo per cui un osservatore può sperare di diagnosticarli nel momento stesso in cui viene presa una decisione.

Il rumore, d'altro canto, è un errore imprevedibile e non facilmente visibile o spiegabile, perciò spesso lo trascuriamo, anche quando causa grossi danni. Per questo motivo il rapporto tra le strategie per ridurre il rumore e quelle volte a eliminare i bias è lo stesso che intercorre tra le misure di igiene e un trattamento medico: l'obiettivo è prevenire una gamma imprecisata di potenziali errori prima del loro verificarsi.

Definiamo questo approccio alla riduzione del rumore *igiene decisionale*. Quando ci si lava le mani, non si sa di preciso quali germi si stanno evitando, si sa solo che farlo è una buona prevenzione contro vari germi (specialmente, ma non solo, durante una pandemia). Analogamente, seguire i princìpi di igiene decisionale significa adottare tecniche che riducono il rumore senza neanche sapere quali errori si stia contribuendo a evitare.

L'analogia con il lavaggio delle mani è voluta. Le misure di igiene possono essere tediose, e i loro vantaggi non sono direttamente visibili: potreste non arrivare mai a sapere in quali problemi vi hanno evitato di cadere. Per contro, quando sorge un problema non sempre è possibile ricondurlo a una specifica inosservanza, e questo spiega perché lavarsi le mani è una misura difficile da far rispettare, anche tra i professionisti sanitari che sanno bene quanto sia importante.

Al pari di questa e altre forme di prevenzione, l'igiene decisionale è preziosa ma ingrata. Correggere un bias ben identificato può quantomeno dare l'impressione tangibile di aver raggiunto un obiettivo, ma ciò non avviene con le procedure per la riduzione del rumore. Statisticamente, si eviteranno molti errori, ma non saprete mai quali; il rumore è un nemico invisibile, e prevenire l'attacco di un nemico invisibile non porterà che a una vittoria invisibile.

Visti i danni che può causare il rumore, tuttavia, vale comunque la pena combattere. Nei prossimi capitoli verranno presentate varie strategie di igiene decisionale impiegate in molteplici campi, come le scienze forensi, le previsioni, la medicina e le risorse umane. Nel capitolo 25 passeremo in rassegna queste strategie e mostreremo come si possano combinare in un approccio integrato finalizzato alla riduzione del rumore.

A proposito di eliminazione del rumore e igiene decisionale

«Sapete quali bias state contrastando di preciso, e in quale direzione stanno deviando i vostri risultati? Se la risposta è no, probabilmente sono in atto diversi bias, ed è difficile prevedere quale sarà quello dominante.»

«Prima di affrontare questa decisione, designiamo un osservatore decisionale.»

«In questo processo abbiamo mantenuto una buona igiene decisionale; probabilmente avremo preso la migliore decisione possibile.»

[1] Per un ottimo esame dei vari studi, vedi J.B. Soll *et al.*, *A User's Guide to Debiasing*, in G. Keren, G. Wu (a cura di), *The Wiley Blackwell Handbook of Judgment and Decision Making*, vol. 2, John Wiley & Sons, New York 2015, p. 684.

[2] HM Treasury, *The Green Book: Central Government Guidance on Appraisal and Evaluation*, UK Crown, London 2018, [https://bit.ly/3qzt2s4].

[3] R.H. Thaler, C.R. Sunstein, *Nudge. La spinta gentile. La nuova strategia per migliorare le nostre decisioni su denaro, salute, felicità*, Feltrinelli, Milano 2014.

[4] R. Hertwig, T. Grune-Yanoff, *Nudging and Boosting: Steering or Empowering Good Decisions*, in "Perspectives on Psychological Science", 12(2017), n. 6.

[5] G.T. Fong *et al.*, *The Effects of Statistical Training on Thinking About Everyday Problems*, in "Cognitive Psychology", 18(1986), n. 3, pp. 253-292.

[6] W.A. Wagenaar, G.B. Keren, *Does the Expert Know? The Reliability of Predictions and Confidence Ratings of Experts*, in E. Hollnagel, G. Mancini, D.D. Woods (a cura di), *Intelligent Decision Support in Process Environments*, Springer, Berlin 1986, pp. 87-103.

[7] C.K. Morewedge *et al.*, *Debiasing Decisions: Improved Decision Making with a Single Training Intervention*, in "Policy Insights from the Behavioral and Brain Sciences", 2(2015), n. 1, pp. 129-140.

[8] A.L. Sellier *et al.*, *Debiasing Training Transfers to Improve Decision Making in the Field*, in "Psychological Science", 30(2019), n. 9, pp. 1371-1379.

[9] E. Pronin *et al.*, *The Bias Blind Spot: Perceptions of Bias in Self Versus Others*, in "Personality and Social Psychology Bulletin", 28(2002), n. 3, pp. 369-381.

[10] D. Kahneman, D. Lovallo, O. Sibony, *Before You Make That Big Decision...*, in "Harvard Business Review", 89(2011), n. 6, pp. 50-60.

[11] A. Gawande, *Checklist Manifesto: How to Get Things Right*, Metropolitan Books, New York 2010.

[12] Office of Information and Regulatory Affairs, *Agency Checklist: Regulatory Impact Analysis*, [https://bit.ly/3y8mByN].

[13] Questa checklist è in parte adattata da D. Kahneman *et al.*, *Before You Make That Big Decision...*, cit.

[14] Vedi A. Gawande, *Checklist Manifesto*, cit.

20

Sequenziare le informazioni nella scienza forense

Nel marzo 2004 una serie di bombe piazzate su diversi treni regionali provocò centonovantadue morti e più di duemila feriti a Madrid. Un'impronta digitale trovata su un sacchetto di plastica presso la scena del crimine fu trasmessa dall'Interpol alle forze dell'ordine di tutto il mondo; di lì a qualche giorno, i laboratori forensi del Federal Bureau of Investigation (l'FBI) la attribuirono a Brandon Mayfield, un cittadino americano residente in Oregon.

Sembrava un sospettato plausibile: un ex ufficiale dell'esercito statunitense che aveva sposato una donna egiziana, si era convertito all'Islam e in veste di avvocato aveva rappresentato alcuni uomini accusati e poi condannati per aver cercato di unirsi ai talebani in Afghanistan. L'FBI lo teneva già d'occhio.

Mayfield fu posto sotto sorveglianza, la sua abitazione venne perquisita e controllata attraverso delle microspie e le sue telefonate furono intercettate. Non riuscendo a ottenere informazioni materiali tramite questi controlli, l'FBI lo arrestò, ma l'uomo non ricevette mai un'accusa formale. Non lasciava il paese da dieci anni. Mentre era in custodia cautelare, gli investigatori spagnoli, che avevano già informato l'FBI di quella che consideravano una mancata corrispondenza tra le impronte digitali di Mayfield e quella presente sul sacchetto di plastica, individuarono un altro sospetto.

L'uomo fu rilasciato dopo due settimane, e tempo dopo il governo americano si scusò, pagò una riparazione di due milioni di dollari e ordinò un'indagine approfondita sulle cause di quell'errore. Il risultato fu il seguente: «Si è trattato di un errore umano, non di un problema metodologico o tecnologico».[1]

Per fortuna errori di questo tipo capitano raramente, ma possono comunque insegnarci qualcosa. Com'è possibile che i migliori esperti di impronte digitali degli Stati Uniti abbiano erroneamente attribuito un'impronta a un uomo che non si era mai avvicinato alla scena del crimine? Per scoprirlo, dobbiamo prima capire come funziona l'analisi delle impronte digitali e in che rapporto si trova con altri tipi di giudizio professionale. Scopriremo così che l'analisi forense delle impronte, che tendiamo a considerare una scienza esatta, in realtà è soggetta ai bias psicologici degli esaminatori, che possono creare più rumore, e quindi più errori, di quanto immaginiamo. Vedremo poi quali misure ha adottato la comunità delle scienze forensi per affrontare questo problema, implementando una strategia di igiene decisionale adatta a tutti gli ambienti: un controllo scrupoloso del flusso di informazioni alla base dei giudizi.

Impronte digitali

Le impronte digitali sono le tracce lasciate dai dermatoglifi dell'ultima falange delle dita sulle superfici che tocchiamo. Anche se esistono esempi dell'utilizzo delle impronte digitali come segni identificativi nell'antichità, l'atto di nascita dell'analisi moderna risale alla fine dell'Ottocento, con la pubblicazione da parte di Henry Faulds, un medico scozzese, del primo

articolo scientifico in cui si raccomandava il loro impiego come tecnica di identificazione.

Nei decenni successivi, le impronte digitali si sono affermate come codice identificativo del casellario giudiziale, andando pian piano a sostituire le tecniche di misurazione antropometrica sviluppate da Alphonse Bertillon, un agente di polizia francese. Lo stesso Bertillon nel 1912 codificò un sistema formale per il confronto delle impronte, e Sir Francis Galton, che abbiamo già incontrato come teorico della saggezza della folla, aveva già sviluppato un sistema simile in Inghilterra. (Ma non c'è da sorprendersi che questi padri fondatori non vengano tanto celebrati: Galton riteneva che le impronte digitali fossero uno strumento utile per la classificazione degli individui su base razziale, mentre Bertillon, probabilmente per i suoi pregiudizi antisemiti, fornì una testimonianza – viziata – come esperto in materia che risultò decisiva nei processi del 1894 e del 1899 contro Alfred Dreyfus.)

Ben presto gli agenti di polizia scoprirono che le impronte digitali non servivano solo come segni identificativi dei recidivi: nel 1892 Juan Vucetich, un poliziotto argentino, fu il primo a confrontare un'impronta latente lasciata sulla scena del crimine con quella del pollice di un indiziato. Da allora la pratica di raccogliere *impronte latenti* (quelle lasciate da qualcuno sulla scena del crimine) e confrontarle con *impronte archiviate* (quelle raccolte in condizioni controllate da individui noti) costituisce l'applicazione più importante dell'analisi delle impronte digitali e fornisce il tipo di prove forensi più ampiamente utilizzato.

Se vi è capitato di imbattervi in un lettore elettronico di impronte (come quelli utilizzati dai servizi di immigrazione in molti paesi), probabilmente penserete che il confronto delle impronte digitali sia un fatto meccanico, semplice e automatico. Ma confrontare un'impronta latente raccolta sulla

scena del crimine con una archiviata è un esercizio molto più delicato del confronto tra due impronte pulite. Quando premete saldamente le dita su un lettore costruito appositamente per questo scopo, producete un'immagine chiara e standardizzata. Al contrario, le impronte latenti sono spesso parziali, poco chiare, sbavate o comunque distorte, e non forniscono la stessa quantità e qualità di informazioni di un'impronta raccolta in un ambiente controllato e dedicato; spesso, inoltre, si sovrappongono ad altre impronte, dello stesso individuo o di altri, e contengono tracce di polvere e di altri reperti presenti sulla superficie. Per decidere se corrispondono alle impronte archiviate di un sospetto occorre un giudizio esperto, ed è qui che subentrano gli esaminatori umani.

Una volta ottenuta un'impronta latente, di solito gli esaminatori seguono un processo chiamato ACE-V (sigla per l'inglese *analysis, comparison, evaluation and verification*, ovvero "analisi, confronto, valutazione e verifica"). Innanzitutto devono analizzarla per stabilire se sia sufficientemente valida per il confronto. In caso affermativo, la confrontano con un'impronta archiviata e formulano una valutazione, che può portare a un'*identificazione* (le impronte provengono dalla stessa persona), a un'*esclusione* (le impronte non provengono dalla stessa persona) o a una decisione inconcludente. L'eventuale identificazione fa scattare il quarto passaggio: la verifica da parte di un altro esaminatore.

Per decenni l'affidabilità di questa procedura è rimasta indiscussa. Se le testimonianze oculari si sono rivelate pericolosamente inaffidabili e perfino le confessioni possono essere false, le impronte digitali, almeno fino all'introduzione delle analisi del DNA, erano considerate la prova più credibile: basti pensare che fino al 2002 non erano mai state messe in discussione nei tribunali americani. All'epoca, il sito web dell'FBI, per esempio, era molto chiaro: «Le impronte digitali costituiscono un *infallibile*

strumento di identificazione personale».[2] Nei rarissimi casi in cui si verificavano degli errori, questi venivano attribuiti all'incompetenza o a intenti fraudolenti.

Le impronte digitali sono state considerate per molto tempo prove inconfutabili anche perché è difficile smentirle: spesso, infatti, non si conosce il "valore reale" di un campione di impronte, cioè la verità empirica su chi abbia davvero commesso il reato. Nel caso di Mayfield e in pochi altri simili, l'errore era particolarmente eclatante, ma, in generale, se un sospetto impugna le conclusioni dell'esperto, la prova delle impronte digitali verrà considerata più affidabile.

Abbiamo visto come non conoscere il valore reale non sia un fatto insolito né un ostacolo alla misurazione del rumore. Possiamo quindi chiederci: quanto rumore è presente nelle analisi delle impronte digitali? O, per essere più precisi, visto che gli esperti di impronte, a differenza dei giudici e dei sottoscrittori, non esprimono un valore numerico ma un giudizio categorico, quanto spesso si trovano in disaccordo e perché? Il primo a cercare di rispondere a questa domanda fu Itiel Dror, un neuroscienziato cognitivista dello University College London, che di fatto svolse una serie di controlli del rumore in un campo in cui non si pensava che il rumore costituisse un problema.

Rumore occasionale nelle analisi delle impronte digitali

Potrà sembrare strano che uno scienziato cognitivista – uno psicologo – metta in discussione il lavoro degli esaminatori di impronte digitali; dopotutto, come abbiamo imparato da serie televisive come *CSI – Scena del crimine* e altre successive, si tratta di professionisti seri, provvisti di guanti di lattice e microscopi, con una passione per il lessico scientifico. Ma Dror

si rese conto che l'esame delle impronte digitali era chiaramente una questione di giudizio, e da buon neuroscienziato cognitivista sapeva bene che dove c'è giudizio, c'è rumore.

Per verificare la sua ipotesi, Dror si concentrò dapprima sul rumore occasionale – ovvero la variabilità tra i giudizi degli *stessi* esperti che ritornano due volte sulla *stessa* evidenza – sostenendo che «se gli esperti non sono affidabili, nel senso che non sono coerenti con le loro precedenti affermazioni, le basi dei loro giudizi e della loro professionalità andranno messe in discussione».[3]

Le impronte digitali costituiscono un perfetto banco di prova per il controllo del rumore occasionale, perché, al contrario dei casi affidati ai medici o ai giudici, non si ricordano facilmente. A ogni modo, occorre prevedere un intervallo di tempo adeguato per assicurarsi che gli esaminatori abbiano dimenticato le impronte. (Negli studi di Dror alcuni esperti coraggiosi e dalla grande apertura mentale accettarono di partecipare, *in qualsiasi momento nell'arco di cinque anni*, a una serie di ricerche senza esserne preventivamente informati.) Inoltre l'esperimento dovrà essere svolto durante il lavoro di routine degli esperti, in modo che non sappiano di essere sotto osservazione. Se, in tali circostanze, il giudizio degli esaminatori cambierà da un test all'altro, saremo in presenza di rumore occasionale.

Il bias di conferma forense

In due dei suoi studi originari, Dror introdusse una modifica importante: quando vedevano le impronte per la seconda volta, alcuni esaminatori ricevevano informazioni aggiuntive condizionanti, quindi portatrici di bias, sul relativo caso. Per esempio, agli esaminatori che in precedenza avevano

individuato una corrispondenza nelle impronte, questa volta veniva detto che il sospettato aveva un alibi, o che le prove balistiche indicavano che non era stato lui. Ad altri che dopo il primo esame avevano concluso che il sospettato fosse innocente o che le impronte fossero inconcludenti, la seconda volta si diceva che l'investigatore riteneva che il sospetto fosse colpevole, o che quest'ultimo avesse confessato il reato. Dror definì questo esperimento un test della *biasability* degli esperti, ovvero della loro suscettibilità ai bias, perché le informazioni contestuali fornite attivavano un bias psicologico (per la precisione, un bias di conferma) in una data direzione.

Effettivamente si scoprì che gli esaminatori erano tutt'altro che immuni ai condizionamenti. Quando lo stesso esperto, una volta entrato in possesso di informazioni condizionanti, analizzava le stesse impronte digitali considerate in precedenza, il suo giudizio cambiava. Nel primo studio, quattro esperti su cinque modificarono la loro decisione precedente a fronte di informazioni contestuali importanti che portavano a escludere il sospettato.[4] Nel secondo, sei esperti esaminarono quattro coppie di impronte, e le informazioni condizionanti portarono a un ripensamento in quattro delle ventiquattro decisioni;[5] è vero, la maggior parte delle valutazioni era rimasta invariata, ma in decisioni di questo tipo un cambiamento di una su sei non è poco, anzi. In seguito altri ricercatori sono giunti alle stesse conclusioni.

Come era prevedibile, gli esaminatori erano più inclini a cambiare idea quando si trovavano di fronte a una decisione difficile, quando le informazioni condizionanti avevano una certa forza, e quando il passaggio avveniva da una decisione definitiva a una inconcludente. È comunque preoccupante che «degli esaminatori esperti di impronte digitali abbiano

preso le loro decisioni sulla base del contesto più che delle informazioni reali contenute nelle impronte».[6]

L'effetto delle informazioni condizionanti non è circoscritto alle conclusioni degli esaminatori (identificazione, esclusione o inconcludenza), ma incide su *ciò che percepiscono*, oltre che su *come* interpretano tale percezione. In un altro studio, Dror e i suoi colleghi dimostrarono che gli esaminatori posti in un contesto affetto da bias non vedono le stesse cose – in senso letterale – di chi non viene esposto a informazioni condizionanti:[7] quando l'impronta latente è accostata alla stampa di un'impronta archiviata target, gli esaminatori osservano un numero decisamente inferiore di dettagli (chiamati *minutiae*) rispetto a quanti ne individuano se gli viene mostrata la sola impronta latente. Uno studio indipendente successivo ha confermato tale conclusione, aggiungendo che «non è chiaro come questo accada».[8]

Per indicare l'impatto delle informazioni condizionanti Dror ha coniato il termine *bias di conferma forense*. Questo bias è stato poi documentato in altre tecniche forensi quali l'analisi delle tracce ematiche, l'indagine sugli incendi dolosi, l'analisi dei resti scheletrici e la patologia forense. Perfino le analisi del DNA, ampiamente ritenute il nuovo standard di riferimento delle scienze forensi, possono essere soggette al bias di conferma, almeno quando gli esperti devono valutare miscele di DNA complesse.[9]

La suscettibilità degli esperti forensi al bias di conferma non è un semplice problema teorico, perché sul piano pratico non vengono attuate misure precauzionali sistematiche di nessun tipo per garantire che gli esaminatori non siano esposti a informazioni condizionanti; al contrario, spesso ricevono informazioni del genere direttamente nelle lettere di presentazione che accompagnano le prove da analizzare.[10] Sovente,

inoltre, gli esperti sono in diretto contatto con la polizia, con il pubblico ministero e con altri esaminatori.

Il bias di conferma solleva poi un altro problema. Un importante meccanismo di tutela contro gli errori previsto dalla procedura ACE-V è la verifica indipendente di un altro esperto prima che venga confermata l'identificazione, ma spesso accade che vengano verificate in forma indipendente solo le identificazioni stesse. Di conseguenza, vi è un forte rischio di incorrere nel bias di conferma, poiché l'esaminatore che effettua la verifica sa già che il primo esame ha dato come esito un'identificazione.[11] Pertanto tale verifica non offre il vantaggio che di norma deriva dall'aggregazione di giudizi indipendenti, perché, di fatto, *non* è indipendente.

Nel caso di Mayfield sembra che si sia generata una cascata di bias di conferma, in cui non due ma tre esperti dell'FBI hanno concorso all'identificazione erronea. Come osservato dalla successiva indagine sull'errore, sembra che il primo esaminatore sia rimasto colpito «dalla forza della correlazione» ottenuta dal motore di ricerca della banca dati di impronte digitali.[12] Anche se, a quanto pare, non era a conoscenza dei dati anagrafici di Mayfield, i risultati forniti dal sistema computerizzato che aveva svolto la ricerca, «associati alla pressione intrinseca data dall'alto profilo del caso», furono sufficienti per produrre l'iniziale bias di conferma. Una volta effettuata un'identificazione erronea da parte del primo esaminatore, proseguiva la relazione, «gli esami successivi erano inquinati»: poiché l'analisi iniziale era stata effettuata da un supervisore molto autorevole, «per gli altri esperti dell'agenzia divenne sempre più difficile contestarla». Quel primo errore era stato replicato e amplificato, fino ad arrivare a una certezza quasi assoluta sulla colpevolezza di Mayfield. Peraltro, anche un esperto indipendente molto stimato nominato

dal tribunale per esaminare le prove su richiesta della difesa confermò l'identificazione dell'FBI.[13]

Lo stesso fenomeno può verificarsi in tutte le discipline forensi, tra le quali l'identificazione delle impronte latenti è ritenuta una delle più oggettive, e anche nello scambio di informazioni tra una disciplina e l'altra. Se gli esaminatori delle impronte digitali possono essere soggetti a bias, a maggior ragione lo saranno gli esperti operanti in altri campi; se per esempio un esperto di incendi dolosi viene a sapere che è stata trovata una corrispondenza nelle impronte, ciò potrà introdurre un bias anche nel suo giudizio; e se un odontoiatra forense sa che l'analisi del DNA ha identificato un sospetto, è probabile che sarà meno incline a indicare che l'impronta di un morso non corrisponda a quella del sospetto. Questi esempi ampliano lo spettro dei bias a cascata: proprio come nelle decisioni di gruppo descritte nel capitolo 8, un errore iniziale indotto da un bias di conferma diventa l'informazione condizionante che influenza un secondo esperto, il cui giudizio influenzerà un terzo e così via.[14]

Una volta stabilito che le informazioni condizionanti creano variabilità, Dror e i suoi colleghi hanno svelato ulteriori evidenze di errori occasionali. Anche quando gli esperti di impronte digitali non ricevono informazioni condizionanti, talvolta cambiano idea su un campione di impronte già esaminato in precedenza.[15] Come si potrà intuire, ripensamenti del genere sono meno frequenti, ma si verificano comunque. Un più ampio studio del 2012 commissionato dall'FBI è giunto alle medesime conclusioni chiedendo a settantadue esaminatori di riconsiderare venticinque coppie di impronte già valutate sette mesi prima.[16] Attingendo a un ampio campione di esperti altamente qualificati, lo studio ha confermato che gli esaminatori di impronte digitali talvolta sono soggetti a rumore occasionale: circa una decisione su dieci è stata cambiata, e la maggior parte dei cambiamenti è

partita da o è risultata in un esito inconcludente, mentre nessun ripensamento ha portato a una falsa identificazione. Il dato più allarmante dello studio è che alcune identificazioni che avevano portato a una condanna avrebbero potuto essere ritenute inconcludenti in un altro esame. Quando gli stessi esaminatori analizzano le stesse impronte digitali, anche se il contesto non è progettato per indurli in errore ma al contrario è mantenuto il più possibile invariato, vi sono incongruenze nelle loro decisioni.

Un po' di rumore, ma quanto errore?

I risultati degli studi sollevano la questione della possibilità di incorrere in errori giudiziari. Non possiamo non interrogarci sull'affidabilità degli esperti che testimoniano in tribunale: la validità richiede affidabilità, perché è difficile essere in linea con la realtà se non si è in linea con se stessi.

Quanti errori, di preciso, vengono causati dalle pecche della scienza forense? Da una rassegna dei casi di trecentocinquanta persone ingiustamente condannate alla pena carceraria rimesse in libertà grazie all'Innocence Project, un'associazione no-profit statunitense istituita per opporsi alle detenzioni ingiuste, è emerso che l'applicazione scorretta della scienza forense toccava circa il 45% dei casi.[17] Per quanto questo dato statistico possa apparire sconfortante, la questione fondamentale per giudici e giurati è un'altra: per capire quanto devono fidarsi dell'esaminatore che si presenta al banco dei testimoni, devono sapere qual è la probabilità che gli scienziati forensi, compresi gli esaminatori di impronte digitali, compiano tragici errori.

Le risposte più concrete a questa domanda si ritrovano in un rapporto del President's Council of Advisors on Science and Technology (PCAST), un gruppo di consulenti composto dai più qualificati scienziati e ingegneri degli Stati Uniti, che nel 2016 compì un'indagine approfondita sull'impiego della scienza forense nei tribunali penali.[18] La relazione riassume le evidenze disponibili sulla validità dell'analisi delle impronte digitali e, in particolare, sulla probabilità che si verifichino identificazioni erronee (falsi positivi) come nel caso di Mayfield.

I dati a favore di questa tecnica si sono rivelati sorprendentemente esigui e, come sottolinea il PCAST, è «desolante» che solo di recente si sia cominciato a studiare l'argomento. I più credibili sono tratti dall'unico grande studio pubblicato sull'accuratezza dell'identificazione mediante impronte digitali, condotto da scienziati dell'FBI nel 2011.[19] Lo studio ha coinvolto 169 esaminatori, ciascuno dei quali ha confrontato circa cento coppie di impronte latenti e archiviate; sono state riscontrate pochissime identificazioni erronee: il tasso di falsi positivi era di uno su seicento.

Un simile tasso di errore è basso ma, come si dice nella relazione, è «*molto più alto* di quanto il grande pubblico (e, per estensione, la maggior parte dei giurati) sarebbe portato a credere in virtù delle persistenti rivendicazioni sull'accuratezza dell'analisi delle impronte digitali».[20] Per giunta, questo studio non conteneva informazioni contestuali condizionanti, e gli esaminatori che vi hanno partecipato sapevano di essere coinvolti in un test, il che potrebbe aver portato a una sottostima degli errori che si verificano nei casi reali. Una ricerca successiva condotta in Florida arrivò a un numero di falsi positivi molto più alto.[21] Le varie conclusioni presenti nella letteratura scientifica indicano che è necessario effettuare ulteriori ricerche sull'accuratezza delle decisioni degli esaminatori di impronte digitali e sul processo con cui si arriva a formularle.

Un dato rassicurante che sembra trasversale a tutti gli studi, tuttavia, è che gli errori degli esaminatori vanno nella direzione dell'eccesso di cautela. Anche se le loro analisi non sono perfettamente accurate, gli esperti sono consapevoli delle conseguenze dei loro giudizi, e tengono conto del costo sproporzionato di possibili errori: data l'altissima credibilità delle impronte digitali, infatti, un'identificazione erronea può avere effetti tragici. Altri tipi di errori hanno un impatto minore. Gli esperti dell'FBI osservano, per esempio, che «nella maggior parte dei casi, un'esclusione ha le stesse implicazioni operative di un risultato inconcludente».[22] Detto in altri termini, il fatto che venga trovata un'impronta digitale sull'arma del delitto è sufficiente per una condanna, ma non basta l'assenza di quell'impronta per scagionare un sospetto.

In linea con quanto si è detto sulla cautela degli esaminatori, i dati indicano che gli esperti ci pensano due volte (o anche di più) prima di propendere per l'identificazione. Nello studio dell'FBI sull'accuratezza delle identificazioni, meno di un terzo delle coppie "appaiate", ovvero quelle in cui l'impronta latente e quella archiviata appartengono alla stessa persona, ha condotto a un'identificazione (che in questo caso sarebbe stata corretta). Inoltre gli esaminatori sono molto più parchi nelle identificazioni di falsi positivi che nelle esclusioni di falsi negativi:[23] insomma, sono soggetti a bias, ma non nella stessa misura in entrambe le direzioni. Come ha rilevato Dror, «è più facile indurre gli esperti forensi a una conclusione non compromettente di inconcludenza piuttosto che a quella irrevocabile di identificazione».[24]

Gli esaminatori vengono formati in modo da considerare l'identificazione erronea come un peccato mortale da evitare a ogni costo, e va riconosciuto che agiscono in accordo con questo principio. Possiamo solo sperare che,

grazie a un tale livello di attenzione, le identificazioni erronee come quella del caso Mayfield e di pochi altri di alto profilo restino rare eccezioni.

Ascoltare il rumore

La riflessione sulla presenza del rumore nella scienza forense non dovrebbe essere vista come una critica agli scienziati stessi. È soltanto la conseguenza di un'osservazione già ripetuta più volte: dove c'è giudizio, c'è rumore, e più di quanto non si pensi. Un compito come l'analisi delle impronte digitali sembra oggettivo, tanto che molti di noi non tenderebbero a considerarlo neanche una forma di giudizio, eppure dà adito a incongruenze, disaccordi e, di tanto in tanto, errori. Per quanto basso possa essere il tasso di errore nell'identificazione mediante impronte digitali, non è inesistente, e, come evidenziato dal PCAST, le giurie dovrebbero esserne consapevoli.

Il primo passo per ridurre il rumore, ovviamente, sta nel riconoscere la possibilità che si verifichi. Questa ammissione non è così naturale nella comunità degli esaminatori delle impronte, molti dei quali erano inizialmente scettici sul controllo del rumore di Dror. L'idea che un esaminatore potesse essere inconsapevolmente influenzato dalle informazioni sul caso irritò molti esperti: in risposta allo studio, il presidente della Fingerprint Society scrisse che «un esaminatore di impronte digitali che [...] viene influenzato in una direzione o nell'altra nel suo processo decisionale [...] è talmente immaturo che dovrebbe andare a lavorare a Disneyland».[25] Il direttore di un grande laboratorio forense osservò che avere accesso alle informazioni sul caso – cioè proprio quelle che potrebbero deviare il giudizio dell'esaminatore – «è una fonte di soddisfazione personale [per gli esperti], che li porta ad apprezzare il

proprio lavoro *senza alterarne il giudizio*».[26] Perfino l'FBI, nella sua indagine interna sul caso Mayfield, ha osservato che «gli esaminatori delle impronte latenti di norma conducono verifiche in cui sono a conoscenza dei risultati dei precedenti esaminatori, *eppure tali risultati non ne influenzano le conclusioni*».[27] Queste affermazioni in sostanza negano l'esistenza del bias di conferma.

Anche quando sono consapevoli del rischio, gli scienziati forensi non sono immuni al bias del punto cieco, ovvero la tendenza a riconoscere la presenza del bias negli altri ma non in se stessi. In un sondaggio condotto su quattrocento professionisti delle scienze forensi provenienti da ventuno paesi,[28] il 71% si è detto d'accordo che «il bias cognitivo è un motivo di preoccupazione nelle scienze forensi in generale», ma solo il 26% riteneva che «i propri giudizi ne fossero influenzati». In pratica, circa la metà di questi professionisti crede che i giudizi dei colleghi siano affetti da rumore, ma i suoi ne siano immuni. Il rumore può essere un problema invisibile anche per chi, di professione, legge l'invisibile.

Sequenziamento delle informazioni

Grazie alla perseveranza di Dror e dei suoi colleghi, pian piano l'atteggiamento generale sta cambiando, e sono sempre più numerosi i laboratori forensi che iniziano a prendere nuove misure per ridurre l'errore nelle proprie analisi. Il rapporto del PCAST, per esempio, elogiava il laboratorio dell'FBI per aver rivisto le proprie procedure in modo da ridurre al minimo il rischio di incorrere nel bias di conferma.

I passi metodologici da compiere sono relativamente semplici, e prendono le mosse da una strategia di igiene decisionale applicabile in molti campi: il *sequenziamento delle informazioni per limitare la formazione di*

intuizioni premature. In ogni giudizio, alcune informazioni sono rilevanti, altre no. Non è sempre un bene avere molte informazioni, soprattutto se vi è la possibilità che introducano dei bias, portando il giudice verso un'intuizione prematura.

In questo spirito, le nuove procedure adottate nei laboratori forensi puntano a proteggere l'indipendenza dei giudizi degli esaminatori dando loro solo le informazioni necessarie, quando servono: in sostanza, il laboratorio li tiene il più possibile all'oscuro dei casi, rivelando le informazioni in maniera graduale. A questo scopo, Dror e i suoi colleghi hanno messo a punto un approccio chiamato *smascheramento sequenziale lineare*.[29]

Dror aggiunge un'altra raccomandazione all'interno della stessa strategia di igiene decisionale: gli esaminatori dovrebbero tenere traccia dei loro giudizi in ogni fase, per esempio documentando la propria analisi delle impronte digitali latenti *prima* di osservare le impronte archiviate per stabilire se corrispondono. Questa sequenza di fasi contribuisce a evitare il rischio che trovino solo quello che cercano. Dovrebbero inoltre annotare i propri giudizi sulle evidenze scientifiche prima di avere accesso alle informazioni contestuali che rischiano di deviare i loro giudizi. Se cambiano idea dopo aver attinto alle informazioni contestuali, simili scostamenti, e le loro motivazioni, andrebbero a loro volta documentati. Tale obbligo limiterebbe il rischio che una prima intuizione distorca l'intero processo.

La stessa logica informa una terza raccomandazione, di grande importanza in termini di igiene decisionale: quando viene chiesto a un esaminatore di verificare l'identificazione effettuata da un collega, il secondo esperto non dovrebbe essere a conoscenza del giudizio del primo.

La presenza del rumore nelle scienze forensi naturalmente preoccupa per le sue conseguenze potenzialmente fatali, ma è anche rivelatrice. Il fatto che siamo rimasti a lungo ignari della possibilità di commettere errori nell'identificazione delle impronte digitali mostra come la nostra fiducia nel giudizio degli esperti talvolta possa essere esagerata, e come un controllo possa rivelare un livello di rumore inaspettato. La capacità di attenuare queste manchevolezze attraverso processi relativamente semplici dovrebbe essere di incoraggiamento per tutti coloro che tengono a migliorare la qualità delle proprie decisioni.

La principale strategia di igiene decisionale illustrata in questo esempio – il sequenziamento delle informazioni – ha un'ampia applicabilità come misura cautelativa contro il rumore occasionale, che, come abbiamo visto, può essere scatenato da innumerevoli fattori, come l'umore e perfino la temperatura esterna. È impossibile riuscire a controllarli tutti, ma si può provare a preservare i giudizi almeno dai più ovvi. Tutti sappiamo, per esempio, che i giudizi possono essere alterati da ansia, paura e altre emozioni, e forse vi sarete accorti che, se è possibile, vale la pena rivedere le proprie idee in momenti diversi, quando a far scattare il rumore occasionale saranno stimoli altrettanto diversi.

Meno scontata è la possibilità che il giudizio venga alterato da un'altra fonte di rumore occasionale: le informazioni – anche quando si tratta di informazioni corrette. Come dimostra l'esempio degli esaminatori di impronte digitali, appena si è a conoscenza di ciò che pensano gli altri, il bias di conferma può indurre a formarsi troppo presto un'impressione generale e a trascurare le informazioni contraddittorie. Questa tendenza è ben sintetizzata dal titolo di due film di Hitchcock: nell'esprimere un giudizio dovremmo puntare a mantenere *L'ombra del dubbio*, non a essere *L'uomo che sapeva troppo*.

A proposito del sequenziamento delle informazioni

«Dove c'è giudizio, c'è rumore, perfino nell'analisi delle impronte digitali.»

«Abbiamo altre informazioni su questo caso, ma non riveliamole tutte agli esperti prima che esprimano un giudizio, per non influenzarli. Anzi, diciamo loro solo ciò che devono assolutamente sapere.»

«La seconda opinione non è indipendente, se chi la esprime conosce la prima. Ancora meno lo sarà la terza: può essere in atto un effetto a cascata.»

«Per contrastare il rumore, bisogna prima ammetterne l'esistenza.»

[1] R. Stacey, *A Report on the Erroneous Fingerprint Individualisation in the Madrid Train Bombing Case*, in "Journal of Forensic Identification", 54(2004), pp. 707-718.

[2] M. Specter, *Do Fingerprints Lie?*, in "The New Yorker", 27 maggio 2002. Corsivo nostro.

[3] I.E. Dror, R. Rosenthal, *Meta-analytically Quantifying the Reliability and Biasability of Forensic Experts*, in "Journal of Forensic Science", 53(2008), pp. 900-903.

[4] I.E. Dror, D. Charlton, A.E. Péron, *Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications*, in "Forensic Science International", 156(2006), pp. 74-78.

[5] I.E. Dror, D. Charlton, *Why Experts Make Errors*, in "Journal of Forensic Identification", 56(2006), pp. 600-616.

[6] I.E. Dror, S.A. Cole, *The Vision in 'Blind' Justice: Expert Perception, Judgment, and Visual Cognition in Forensic Pattern Recognition*, in "Psychonomic Bulletin and Review", 17(2010), pp. 161-167, in part. 165. Vedi anche I.E. Dror, *A Hierarchy of Expert Performance (HEP)*, in "Journal of Applied Research in Memory and Cognition", (2016), pp. 1-6.

[7] I.E. Dror *et al.*, *Cognitive Issues in Fingerprint Analysis: Inter- and Intra-Expert Consistency and the Effect of a 'Target' Comparison*, in "Forensic Science International", 208(2011), pp. 10-17.

[8] B.T. Ulery *et al.*, *Changes in Latent Fingerprint Examiners' Markup Between Analysis and Comparison*, in "Forensic Science International", 247(2015), pp. 54-61.

[9] I.E. Dror, G. Hampikian, *Subjectivity and Bias in Forensic DNA Mixture Interpretation*, in "Science and Justice", 51(2011), pp. 204-208.

[10] M.J. Saks *et al.*, *Context Effects in Forensic Science: A Review and Application of the Science of Science to Crime Laboratory Practice in the United States*, in "Science Justice Journal of Forensic Science Society", 43(2003), pp. 77-90.

[11] President's Council of Advisors on Science and Technology (PCAST), *Report to the President: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*, Executive Office of the President, PCAST, Washington, DC 2016.

[12] R. Stacey, *Erroneous Fingerprint*, cit.

[13] I.E. Dror, S.A. Cole, *Vision in 'Blind' Justice*, cit.

[14] I.E. Dror, *Biases in Forensic Experts*, in "Science", 360(2018), p. 243.

[15] I.E. Dror, D. Charlton, *Why Experts Make Errors*, cit.

[16] B.T. Ulery *et al.*, *Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners*, "PLoS One", 7(2012).

[17] Innocence Project, *Overturning Wrongful Convictions Involving Misapplied Forensics*, in "Misapplication of Forensic Science" (2018), pp. 1-7, [https://bit.ly/3enr8Ga]. Vedi anche S.M. Kassin *et al.*, *The Forensic Confirmation Bias: Problems, Perspectives, and Proposed Solutions*, in "Journal of Applied Research in Memory and Cognition", 2(2013), pp. 42-52.

[18] PCAST, *Report to the President*, cit.

[19] B.T. Ulery *et al.*, *Accuracy and Reliability of Forensic Latent Fingerprint Decisions*, in "Proceedings of the National Academy of Sciences", 108(2011), pp. 7733-7738.

[20] PCAST, *Report to the President*, cit., p. 95. Corsivo nell'originale.

[21] I. Pacheco, B. Cerchiai, S. Stoiloff, *Miami-Dade Research Study for the Reliability of the ACE-V Process: Accuracy & Precision in Latent Fingerprint Examinations*, relazione finale, Miami-Dade Police Department Forensic Services Bureau, 2014, [www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf].

[22] B.T. Ulery *et al.*, *Factors Associated with Latent Fingerprint Exclusion Determinations*, in "Forensic Science International", 275(2017), pp. 65-75.

[23] R.N. Haber, I. Haber, *Experimental Results of Fingerprint Comparison Validity and Reliability: A Review and Critical Analysis*, in "Science & Justice", 54(2014), pp. 375-389.

[24] I.E. Dror, *Hierarchy of Expert Performance*, cit., p. 3.

[25] M. Leadbetter, lettera alla redazione, "Fingerprint World", 33(2007), p. 231.

[26] L. Butt, *The Forensic Confirmation Bias: Problems, Perspectives and Proposed Solutions – Commentary by a Forensic Examiner*, in "Journal of Applied Research in Memory and Cognition", 2(2013), pp. 59-60. Corsivi nostri.

[27] R. Stacey, *Erroneous Fingerprint*, cit., p. 713. Corsivi nostri.

[28] J. Kukucka *et al.*, *Cognitive Bias and Blindness: A Global Survey of Forensic Science Examiners*, in "Journal of Applied Research in Memory and Cognition", 6(2017).

[29] I.E. Dror *et al.*, lettera alla redazione: *Context Management Toolbox: A Linear Sequential Unmasking (LSU) Approach for Minimizing Cognitive Bias in Forensic Decision Making*, in "Journal of Forensic Science", 60(2015), pp. 1111-1112.

21

Selezione e aggregazione nelle previsioni

Molti giudizi si basano su previsioni. Quale sarà il tasso di disoccupazione nel prossimo trimestre? Quante auto elettriche verranno vendute l'anno prossimo? Quali saranno gli effetti del cambiamento climatico nel 2050? Quanto tempo occorrerà per completare un nuovo edificio? Quali saranno gli introiti annuali di una data società? Come si comporterà un nuovo dipendente? Quali costi genererà una nuova norma sull'inquinamento atmosferico? Chi vincerà le elezioni? Le risposte a queste domande hanno conseguenze enormi, e spesso saranno all'origine di scelte fondamentali da parte di istituzioni pubbliche e private.

Gli esperti di previsioni, che analizzano quando e perché queste si rivelano errate, operano una netta distinzione tra bias e rumore (definito anche "incoerenza" o "inaffidabilità"). Tutti concordano sul fatto che in certi contesti si formulano previsioni affette da bias: le agenzie governative, per esempio, mostrano un ottimismo inverosimile nella programmazione del bilancio.[1] In media, prospettano una crescita economica irrealisticamente alta e un deficit altrettanto irrealisticamente basso; ai fini pratici, conta poco che il loro ottimismo inverosimile derivi da un bias cognitivo o da considerazioni politiche.

Per di più, i previsori tendono a fidarsi troppo dei propri giudizi: se si chiede loro di formulare previsioni sotto forma di intervalli di confidenza più che come stime puntuali, tendono a selezionare intervalli più ristretti del dovuto.[2] Per fare un esempio, una ricerca effettuata a cadenza

trimestrale chiede ai direttori finanziari delle società americane di elaborare una stima del rendimento annuo dell'indice Standard & Poor 500 per l'anno successivo.[3] I soggetti esprimono due stime: un valore minimo al di sotto del quale credono vi sia una possibilità su dieci che si collochi il rendimento effettivo, e un valore massimo, che ritengono la soglia oltre la quale il rendimento avrà solo una probabilità su dieci di arrivare. Queste due cifre, quindi, sono i due limiti di un intervallo di confidenza dell'80%. Eppure il rendimento realizzato ricade in quell'intervallo solo il 36% delle volte: i direttori finanziari hanno troppa fiducia nella precisione delle proprie previsioni.

Inoltre i previsori sono affetti da rumore. Un manuale di riferimento della disciplina, *Principles of Forecasting*, a cura di J. Scott Armstrong, sostiene che anche tra gli esperti «l'inaffidabilità è una fonte di errore nelle previsioni basate su giudizi».[4] In effetti, il rumore è una *grande* fonte di errore. Il rumore occasionale è molto comune: non sempre i previsori concordano con le loro stesse previsioni; quello interpersonale è altrettanto diffuso, poiché spesso i previsori sono in disaccordo tra loro, anche se si tratta di specialisti. Se chiedete a dei docenti di diritto di prevedere le decisioni della Corte suprema degli Stati Uniti, troverete un alto livello di rumore.[5] Se chiedete a degli specialisti di fare una proiezione dei benefici annuali in termini economici delle norme sull'inquinamento atmosferico, troverete una variabilità enorme nelle risposte, con un intervallo che potrà andare, per esempio, dai tre ai nove miliardi di dollari.[6] Se chiedete a un gruppo di economisti di effettuare una previsione sulla disoccupazione e sulla crescita, anche lì troverete una grande variabilità. Abbiamo già visto molti esempi di previsioni affette da rumore, e le ricerche sulle previsioni ne rivelano molti altri ancora.[7]

Migliorare le previsioni

La ricerca offre anche alcuni suggerimenti per ridurre il rumore e il bias. Qui non li tratteremo in maniera esaustiva, ma ci concentreremo su due strategie di riduzione del rumore di ampia applicabilità. Una non è altro che un'attuazione del principio già menzionato nel capitolo 18: selezionare giudici migliori porta a giudizi migliori. L'altra è una delle strategie di igiene decisionale più universalmente applicabili: l'aggregazione di diverse stime indipendenti.

Il modo più semplice per aggregare più previsioni consiste nel calcolo della loro media. È matematicamente certo che questa operazione riduca il rumore: nello specifico, lo divide per la radice quadrata del numero dei giudizi di cui si è calcolata la media. Ciò significa che effettuando la media di cento giudizi ridurrete il rumore del 90%, mentre con la media di quattrocento giudizi del 95%, quindi in pratica lo eliminerete. Questa legge statistica costituisce il fondamento dell'approccio della saggezza della folla discusso nel capitolo 7.

Poiché il calcolo della media non riduce affatto il bias, il suo effetto sull'errore totale (MSE) dipende dalle proporzioni di bias e rumore all'interno di quest'ultimo. Per questo la saggezza della folla dà risultati migliori con giudizi indipendenti, in cui è meno probabile riscontrare bias condivisi. A livello empirico, ampie evidenze dimostrano che calcolando la media di più previsioni si aumenta decisamente l'accuratezza, come accade per esempio nel cosiddetto "consensus", ovvero la media delle previsioni che gli analisti finanziari effettuano sull'andamento di una società o un titolo quotati.[8] Nelle previsioni commerciali ed economiche, ma anche in quelle del meteo, la media non ponderata dei giudizi di un gruppo di esperti dà risultati migliori della maggior parte – e a volte perfino di tutte – le previsioni prese singolarmente.[9] Calcolare la media di previsioni ottenute

con metodi differenti ha lo stesso effetto: da un'analisi di trenta confronti empirici effettuati in diversi campi è emerso che le previsioni combinate hanno ridotto gli errori in media del 12,5%.[10]

La media semplice non è l'unico metodo per aggregare le previsioni. Una strategia basata su una "folla selezionata", che appunto seleziona i soggetti migliori in virtù dell'accuratezza dei loro giudizi recenti e calcola la media dei giudizi di un piccolo numero di soggetti (per esempio, cinque), può essere altrettanto efficace.[11] Per i decisori che apprezzano la competenza, sarà inoltre più facile comprendere e adottare una strategia che si affidi non solo all'aggregazione, ma anche alla selezione.

Un metodo per produrre previsioni aggregate consiste nell'impiego dei cosiddetti "mercati di previsione", in cui singoli individui scommettono su esiti probabili e vengono quindi incentivati a effettuare previsioni corrette. Più volte questa tecnica ha dato ottimi risultati: se il prezzo su questi mercati suggerisce che certi eventi hanno, per esempio, il 70% di probabilità di verificarsi, si verificano il 70% delle volte.[12] Molte società di vari settori impiegano i mercati di previsione per aggregare prospettive diverse.[13]

Un altro processo formale utilizzato per aggregare prospettive diverse è noto come "metodo Delphi".[14] Nella sua versione classica, prevede più round in cui i partecipanti sottopongono le proprie stime (o i loro voti) a un moderatore, senza sapere chi sono gli altri soggetti coinvolti. In ogni nuovo round i partecipanti motivano le proprie stime e reagiscono alle motivazioni fornite dagli altri, sempre in forma anonima. Il processo ha l'obiettivo di far convergere le stime, talvolta anche in maniera forzata, ovvero chiedendo di far ricadere i propri giudizi in un preciso intervallo della distribuzione dei giudizi emersi nella fase precedente. Questa tecnica sfrutta tanto l'aggregazione quanto l'apprendimento collettivo.

Il metodo Delphi si è rivelato efficace in molte situazioni, ma non è sempre facile da attuare.[15] Una versione più semplice, chiamata "mini Delphi", può essere concentrata in un unico incontro.[16] È anche chiamata *estimate-talk-estimate*, in quanto richiede ai partecipanti prima di formulare una stima indipendente (senza comunicarla), poi di spiegarla e giustificarla, infine di effettuare una nuova stima che tenga conto delle stime e delle spiegazioni degli altri. Il giudizio di consenso consisterà nella media delle stime individuali ottenute nella seconda fase.

Il Good Judgment Project

Il lavoro forse più innovativo sulla qualità delle previsioni, che supera di gran lunga tutto ciò che abbiamo visto finora, è stato avviato nel 2011 con la fondazione del Good Judgment Project da parte di tre noti scienziati comportamentali. Philip Tetlock (di cui abbiamo già parlato nel capitolo 11, a proposito della sua valutazione delle previsioni a lungo termine sugli eventi politici), sua moglie Barbara Mellers e Don Moore hanno formato un gruppo che si poneva l'obiettivo di comprendere meglio le previsioni e, in particolare, di capire perché ad alcuni riescono così bene.

Il Good Judgment Project ha lanciato una campagna di reclutamento di decine di migliaia di volontari, ma non tra specialisti o esperti, bensì tra persone comuni di ogni estrazione sociale, a cui è stato chiesto di rispondere a centinaia di domande di questo tipo:

- La Corea del Nord lancerà un'arma nucleare entro la fine di quest'anno?
- La Russia procederà a un'annessione ufficiale di altri territori dell'Ucraina nei prossimi tre mesi?
- L'India o il Brasile diventeranno membri permanenti del Consiglio di sicurezza delle Nazioni Unite nei prossimi due anni?
- Nel corso del prossimo anno vi saranno paesi che usciranno dall'Eurozona?

Come mostrano questi esempi, il progetto ha posto quesiti molto ampi relativi a eventi globali. Si noti che il tentativo di rispondere a queste domande solleva molti problemi che si riscontrano anche in previsioni più comuni: quando un avvocato si chiede se un suo cliente vincerà una causa, o uno studio televisivo si chiede se una certa trasmissione avrà successo, entrano in gioco le loro capacità previsionali. Tetlock e i suoi colleghi volevano capire se esistono persone particolarmente brave nell'effettuare previsioni, e se questa capacità si potesse apprendere o almeno migliorare.

Per comprendere i principali risultati di questa ricerca, occorre spiegare alcuni aspetti fondamentali del metodo per la valutazione dei previsori adottato da Tetlock e dal suo gruppo. Innanzitutto, hanno impiegato un ampio numero di previsioni il cui successo o fallimento potesse dipendere da una pura questione di fortuna. Se prevedete che la vostra squadra vincerà la prossima partita, ed effettivamente vince, non è detto che siate dei buoni previsori; forse prevedete *sempre* che la vostra squadra vincerà: se questa è la vostra strategia, e se la vostra squadra vince solo metà delle volte, la vostra capacità previsionale non è particolarmente elevata. Per ridurre il ruolo della fortuna, i ricercatori hanno esaminato i risultati medi dei partecipanti in numerose previsioni.

In secondo luogo, i ricercatori hanno chiesto ai partecipanti di effettuare le loro previsioni in termini di probabilità che un evento sarebbe accaduto, non mediante un'indicazione binaria del tipo "accadrà / non accadrà". Per molti, infatti, fare previsioni consiste nello schierarsi da una parte o dall'altra; tuttavia, considerando la nostra ignoranza oggettiva degli eventi futuri, è molto meglio formulare previsioni probabilistiche. Se qualcuno nel 2016 ha affermato: «C'è il 70% di probabilità che Hillary Clinton vinca le elezioni presidenziali americane», non è detto che fosse un cattivo previsore: se si dice che vi è una probabilità del 70% che un evento accada,

si afferma allo stesso tempo che il 30% delle volte non accadrà. Per sapere se un previsore sia valido, dovremmo chiederci se le sue stime di probabilità coincidono con la realtà. Poniamo che un particolare previsore che chiameremo Margaret dica che vi è una probabilità del 60% che si verifichino cinquecento eventi diversi: se trecento di questi dovessero davvero accadere, potremmo concludere che il livello di confidenza di Margaret sia ben *calibrato*. Una buona calibratura è essenziale per una buona previsione.

In terzo luogo, aggiungendo un'ulteriore finezza, Tetlock e i suoi colleghi non si sono limitati a chiedere ai loro previsori di effettuare *una sola* stima della probabilità che un evento accadesse o meno nel giro di un anno, ma hanno dato ai partecipanti l'opportunità di rivedere costantemente le proprie previsioni alla luce di nuove informazioni. Supponiamo che aveste previsto, nel 2016, che il Regno Unito avesse appena il 30% di possibilità di lasciare l'Unione Europea entro la fine del 2019. Avendo appreso dai nuovi sondaggi che il "Leave" stava guadagnando terreno, probabilmente avreste rivisto al rialzo la vostra previsione. Quando è stato reso noto il risultato del referendum, non era ancora certo che il Regno Unito avrebbe lasciato l'Unione in quell'arco di tempo, ma certo sembrava molto più probabile (tecnicamente la Brexit è avvenuta nel 2020.)

Tetlock e i suoi colleghi hanno consentito ai previsori di aggiornare le proprie stime ogni volta che acquisiscono una nuova informazione e, ai fini del punteggio, ciascuno di questi aggiornamenti viene trattato alla stregua di una nuova previsione. In questo modo i partecipanti al Good Judgment Project sono incentivati a monitorare le notizie e ad aggiornare di continuo le proprie stime. Questo approccio rispecchia ciò che ci si aspetta dai previsori in ambito aziendale e governativo, che dovrebbero aggiornare di

frequente le proprie previsioni sulla base delle nuove informazioni, pur rischiando di essere criticati per aver cambiato idea. (Una nota risposta a queste critiche, attribuita a John Maynard Keynes, è la seguente: «Quando cambiano i fatti, io cambio idea. E *voi*?».)

Inoltre, per valutare la performance dei previsori, il Good Judgment Project adotta un sistema sviluppato da Glenn W. Brier nel 1950: il cosiddetto "punteggio di Brier" misura la distanza tra ciò che una persona prevede e ciò che accade realmente.

Si tratta di uno strumento intelligente per aggirare un problema molto frequente nelle previsioni probabilistiche: come incentivare i previsori a non fare mai passi avventati per tenere aperte più alternative. Pensiamo a Margaret, che abbiamo definito un previsore ben calibrato perché ha previsto che vi fosse una probabilità del 60% che si verificassero cinquecento eventi, e trecento di questi sono poi accaduti. Questo risultato è forse meno convincente di quanto sembri. Se Margaret fosse una meteorologa che prevede *sempre* un rischio di pioggia del 60% e i giorni di pioggia sono trecento su cinquecento, le previsioni di Margaret sarebbero sì ben calibrate, ma di fatto anche piuttosto inutili. In pratica, vi sta dicendo che sarebbe meglio portarsi sempre dietro l'ombrello, nel caso dovesse piovere. Confrontiamola con Nicholas, che prevede una probabilità del 100% di pioggia nei trecento giorni in cui pioverà e una dello 0% nei duecento giorni asciutti. Nicholas ha la stessa calibratura perfetta di Margaret: quando entrambi prevedono che in una certa percentuale di giorni pioverà, la pioggia cade esattamente in quella percentuale di giorni. Ma le previsioni di Nicholas sono molto più apprezzabili: invece di tenere aperte più alternative, vi dice quando dovreste portarvi dietro l'ombrello. In termini tecnici, possiamo dire che Nicholas ha un'alta *risoluzione*, oltre che una buona calibratura.

Il punteggio di Brier premia tanto la buona calibratura quanto la buona risoluzione. Per ottenere una valutazione alta, bisogna non solo fare previsioni mediamente corrette (essere ben calibrati), ma anche essere disposti a prendere posizione e a differenziare le proprie previsioni (avere un'alta risoluzione). Questo tipo di punteggio si basa sulla logica degli errori quadratici medi, quindi più è basso, meglio è: lo zero indica una previsione perfetta.

Ora che sappiamo come sono stati assegnati i punteggi, quali risultati hanno ottenuto i volontari del Good Judgment Project? Si è visto che la stragrande maggioranza di loro ha dato pessimi risultati, mentre circa il 2% si è distinto. Come abbiamo già visto, Tetlock definisce questi ultimi "superprevisori". Non che fossero infallibili, ma le loro previsioni erano assai superiori a un risultato puramente dettato dal caso. È interessante notare che un funzionario governativo ha dichiarato che il gruppo ha dato risultati decisamente «migliori della media degli analisti dell'intelligence, che disponevano di intercettazioni e altri dati segreti».[17] Vale la pena di soffermarsi su questo paragone. Gli analisti dell'intelligence vengono addestrati per effettuare previsioni accurate; non sono dei dilettanti, e inoltre hanno accesso a informazioni classificate. Eppure non sono bravi quanto i superprevisori.

Perpetual beta

Cosa rende così bravi i superprevisori? In linea con l'argomentazione avanzata nel capitolo 18, potremmo ipotizzare a buon diritto che siano insolitamente intelligenti. Questa ipotesi non è sbagliata. Nei test sulla capacità mentale generale, i superprevisori ottengono risultati migliori della media dei volontari del Good Judgment Project (a sua volta

decisamente più elevata della media nazionale). Ma la differenza non è poi tanta, e molti volontari che ottengono ottimi risultati nei test di intelligenza non rientrano tra i superprevisori. A parte l'intelligenza generale, potremmo aspettarci che siano insolitamente bravi in matematica, e in effetti è così, ma il vero vantaggio non sta tanto nel loro talento con i numeri, quanto nella loro propensione al pensiero analitico e probabilistico.

Pensiamo alla volontà e alla capacità dei superprevisori di strutturare e disaggregare i problemi: invece di formarsi un giudizio olistico su un grande tema geopolitico (se una nazione uscirà dall'Unione Europea oppure no, se in un certo luogo scoppierà una guerra, se un funzionario pubblico verrà assassinato), lo scompongono nelle sue parti. Si chiedono: «Cosa dovrebbe succedere per far accadere questo evento? Cosa per non farlo accadere?». Invece di dare una risposta istintiva o basata su un'impressione generale, si pongono una serie di domande accessorie e cercano di arrivare a una risposta.

I superprevisori sono anche molto bravi ad assumere una visione esterna, e prendono in seria considerazione i tassi di base: come dicevamo a proposito del caso Gambardi nel capitolo 13, prima di concentrarci sui dettagli del suo profilo sarebbe utile conoscere la probabilità che l'amministratore delegato medio venga licenziato o si dimetta entro due anni dall'assunzione. I superprevisori guardano in maniera sistematica i tassi di base. Alla domanda se l'anno prossimo si verificherà uno scontro armato tra Cina e Vietnam per una disputa territoriale, non si concentrano solo o immediatamente sui rapporti attuali tra i due stati. Forse su questo avranno già un'intuizione, alla luce delle notizie e delle analisi che hanno letto, ma sanno che la loro intuizione su un certo evento in genere non è affidabile, e quindi si mettono subito alla ricerca del tasso di base, ovvero si

chiedono con quale frequenza in passato le dispute territoriali hanno fatto scoppiare uno scontro armato. Se simili scontri sono rari, i superprevisori ne terranno conto, e solo a quel punto passeranno ad analizzare i dettagli della situazione tra Cina e Vietnam.

In sostanza, ciò che distingue i superprevisori non è la loro intelligenza in quanto tale, ma il modo in cui la applicano. Le capacità che mettono in campo riflettono il tipo di stile cognitivo che, come dicevamo nel capitolo 18, con ogni probabilità porterà a giudizi migliori, ovvero un alto livello di "apertura mentale attiva". Ricorderete che il test sul pensiero aperto in modo attivo propone affermazioni come: «I dati che contraddicono le proprie credenze andrebbero presi in considerazione» ed «È più utile prestare attenzione a qualcuno con cui siamo in disaccordo che a chi è d'accordo con noi». Chiaramente le persone che ottengono un punteggio alto in questo test sono anche quelle che non esitano ad aggiornare i propri giudizi (senza prendere decisioni avventate) quando vengono a conoscenza di nuove informazioni.

Per caratterizzare lo stile di pensiero dei superprevisori, Tetlock usa l'espressione *perpetual beta*, un termine informatico che indica un programma di cui non è prevista una versione finale, ma che viene costantemente utilizzato, analizzato e perfezionato. Tetlock ritiene che «il predittore più forte di chi ascenderà al rango di superprevisore è il *perpetual beta*, cioè la misura di quanto ci si sforza di aggiornare le proprie credenze e automigliorarsi».[18] Dal suo punto di vista, «a renderli così abili non è tanto chi sono, ma cosa fanno: il grande lavoro di ricerca, le attente riflessioni e l'autocritica, la raccolta e la sintesi di altre prospettive, i giudizi particolareggiati e l'incessante aggiornamento». Pensano in maniera circolare: «tentativo, fallimento, analisi, correzione, nuovo tentativo».[19]

Rumore e bias nelle previsioni

A questo punto sarete tentati di pensare che sia possibile formare qualcuno affinché diventi un superprevisore, o almeno dia risultati simili, è in effetti è ciò che Tetlock e i suoi collaboratori hanno cercato di fare. Il tentativo può essere considerato la seconda fase della ricerca per comprendere le ragioni alla base del talento dei superprevisori e migliorare ulteriormente le loro prestazioni.

In un importante studio, i ricercatori hanno distribuito in maniera casuale in tre gruppi alcuni previsori comuni (quindi non superprevisori) per verificare quale effetto avessero degli interventi di vario tipo sulla qualità dei loro giudizi successivi. Tali interventi rispecchiano tre delle strategie volte al miglioramento dei giudizi descritte in precedenza:

1. *Formazione*. Diversi previsori hanno seguito un tutorial pensato per migliorare le loro capacità mediante il ragionamento probabilistico: imparavano a riconoscere vari bias (compresi la fallacia del tasso di base, l'eccesso di fiducia e il bias di conferma), e capivano quanto fosse importante calcolare la media di più previsioni provenienti da fonti diverse e considerare le classi di riferimento.
2. *Raggruppamento (una forma di aggregazione).* Ad alcuni previsori veniva chiesto di lavorare in gruppi in cui potevano conoscere e discutere le previsioni altrui. Il raggruppamento poteva aumentare l'accuratezza dei giudizi, incoraggiando i soggetti ad affrontare argomenti contrastanti con una mente aperta e attiva.
3. *Selezione*. Tutti i previsori ricevevano un punteggio per la loro accuratezza e, dopo un anno, i migliori (il 2% di loro) venivano segnalati come superprevisori e l'anno successivo avevano l'opportunità di lavorare insieme in gruppi d'élite.

Tutti e tre gli interventi si rivelarono efficaci, nel senso che migliorarono il punteggio di Brier dei partecipanti. La formazione ebbe un buon impatto, il raggruppamento incise ulteriormente e la selezione ebbe un effetto ancora maggiore.

Questi importanti risultati confermano il valore dell'aggregazione dei giudizi e della selezione di soggetti validi. Ma c'è dell'altro. Sulla scorta dei

dati sugli effetti di ciascun intervento, Ville Satopää, uno dei collaboratori di Tetlock e Mellers, sviluppò una raffinata tecnica statistica per carpire in che modo, esattamente, ogni intervento migliorasse le previsioni.[20] In linea di principio, rifletté, tre sono i principali motivi per cui alcuni previsori hanno una performance migliore o peggiore degli altri:

1. Potranno essere più abili nel cercare e analizzare i dati nel campo relativo alla previsione che devono compiere. Questa spiegazione verte sull'importanza delle informazioni.
2. Alcuni previsori potranno avere una tendenza generale a sbagliare sempre nella stessa direzione rispetto al valore reale di una previsione: se, in centinaia di previsioni, sistematicamente sovrastimiamo o sottostimiamo la probabilità che si verifichino certi cambiamenti rispetto allo status quo, è possibile dire che siamo soggetti a una forma di bias in favore o del cambiamento o della stabilità.
3. Alcuni previsori potranno essere meno suscettibili al rumore (o agli errori casuali). Nelle previsioni, come in tutti i giudizi, il rumore può essere scatenato da diversi fattori. I previsori potrebbero avere reazioni spropositate a notizie di un certo tipo (è un esempio di quello che abbiamo chiamato "errore strutturale"), potranno essere soggetti al rumore occasionale, o affetti da rumore nell'uso della scala delle probabilità. Tutti questi errori (e molti altri) hanno dimensioni e direzioni imprevedibili.

Satopää, Tetlock, Mellers e il loro collega Marat Salikhov hanno chiamato il loro modello di previsione BIN (da *bias, information and noise*, ovvero "bias, informazioni e rumore"). Si sono quindi prefissi di misurare quanto incidesse ognuna di queste componenti sul miglioramento della performance in ciascuno dei tre interventi, giungendo a una semplice risposta: tutti e tre agivano innanzitutto riducendo il rumore. I ricercatori hanno spiegato che «quando un intervento incrementava l'accuratezza, funzionava innanzitutto eliminando gli errori casuali di giudizio, anche se, curiosamente, l'intento originario della formazione era in realtà la riduzione del bias».[21]

Poiché la formazione era stata progettata per ridurre i bias, un previsore non eccelso avrebbe previsto che il suo principale effetto sarebbe stato appunto la riduzione del bias; al contrario, la formazione aveva agito

riducendo il rumore. Questo esito imprevisto si può spiegare facilmente: la formazione di Tetlock puntava a contrastare i bias *psicologici* ma, come sapete, non sempre questi si traducono in un bias statistico. Quando toccano individui diversi che esprimono giudizi diversi in modi diversi, i bias psicologici producono rumore. Questo è senz'altro il caso in questo studio, dal momento che gli eventi da prevedere sono piuttosto vari: gli stessi bias possono portare un previsore ad avere una reazione eccessiva o insufficiente, a seconda dell'argomento. Non dovremmo quindi aspettarci un bias *statistico*, ovvero la tendenza generale dei previsori a credere che certi eventi accadano o meno. Di conseguenza, formare i previsori per contrastare i propri bias psicologici è un intervento efficace perché riduce il rumore.

Il raggruppamento ha avuto un effetto altrettanto forte nel ridurre il rumore, ma ha anche migliorato in misura significativa la capacità dei gruppi di cogliere le informazioni. Questo risultato è coerente con la logica dell'aggregazione: più cervelli che lavorano insieme sono più capaci di trovare informazioni di quanto non lo sia un unico cervello. Se Alice e Brian collaborano, e Alice individua dei segnali che Brian non ha notato, la loro previsione congiunta sarà migliore. Quando lavorano insieme, sembra che i superprevisori siano capaci di evitare i pericoli della polarizzazione di gruppo e delle cascate informative. Al contrario, condividono i loro dati e le loro opinioni e, con un'apertura mentale attiva, traggono il massimo dalla combinazione di queste informazioni. Satopää e i suoi colleghi spiegano così il vantaggio che ne deriva: «Il raggruppamento, diversamente dalla formazione, consente ai previsori di sfruttare al meglio le informazioni».[22]

La selezione ha avuto l'effetto totale più consistente. Alcuni miglioramenti sono dovuti a un uso più efficace delle informazioni, e i superprevisori riescono meglio degli altri a trovarne di pertinenti, forse

perché sono più sagaci, più motivati e più esperti nell'avanzare questo tipo di previsioni rispetto al partecipante medio. Ma l'effetto principale della selezione, ancora una volta, è la riduzione del rumore: i superprevisori sono meno soggetti al rumore dei previsori comuni e perfino dei gruppi che hanno svolto la formazione. Anche questo esito ha sorpreso Satopää e gli altri ricercatori: «È possibile che i "superprevisori" debbano il loro successo più a una maggiore dedizione alla riduzione degli errori di misurazione che a una lettura critica delle notizie», virtù non replicabile dagli altri.[23]

Quando la selezione e l'aggregazione funzionano

Il successo del progetto sottolinea il valore di due strategie di igiene decisionale: la *selezione* (i superprevisori svettano sugli altri) e l'*aggregazione* (quando lavorano in gruppo, tutti i previsori danno risultati migliori). Le due strategie sono ampiamente applicabili in molti giudizi, e, ove possibile, bisognerebbe puntare a combinarle, formando gruppi di soggetti (per esempio, previsori, investitori professionisti, responsabili delle assunzioni) selezionati per essere competenti nel proprio lavoro *e* complementari tra loro.

Finora abbiamo considerato come ottenere una maggiore precisione attraverso la media di più giudizi indipendenti, come negli esperimenti sulla saggezza della folla. Aggregare le stime dei soggetti più validi migliorerà ulteriormente l'accuratezza, ma combinare giudizi indipendenti e complementari darà risultati ancora più incisivi. Immaginate che quattro persone siano testimoni di un reato: è cruciale, naturalmente, che non si influenzino a vicenda, ma se, per giunta, hanno assistito al reato da quattro

angolazioni diverse, la qualità delle informazioni che forniranno sarà ancora migliore.

Il compito di assemblare un gruppo di professionisti per esprimere giudizi congiunti non è diverso da quello di assemblare una batteria di test per prevedere le performance future di chi si candida per accedere a un corso di laurea o a un ruolo professionale. Lo strumento standard per compiti simili è la regressione multipla (presentata nel capitolo 9), che opera una selezione delle variabili in successione. Il test che predice meglio i risultati verrà selezionato per primo, ma il test successivo a essere scelto non è detto che sia il secondo in termini di validità. È invece quello che *aggiunge* il maggior potere predittivo al primo test, fornendo previsioni che siano insieme valide e non ridondanti con quelle del primo. Supponete, per esempio, di avere due test attitudinali con una correlazione di 0,50 e 0,45 con le prestazioni future, e un test di personalità che ha una correlazione di appena 0,30, ma non è legato in alcun modo ai test attitudinali. La soluzione migliore sarebbe selezionare prima il test attitudinale più valido, poi il test di personalità, che darà più informazioni nuove.

Analogamente, se state assemblando un gruppo di decisori, naturalmente dovreste scegliere per primo il migliore. Ma la seconda scelta dovrebbe ricadere su un individuo abbastanza valido che metta in campo nuove competenze, invece di un soggetto più valido ma molto simile al primo. Un gruppo così selezionato sarà superiore, in quanto la validità dei giudizi collettivi aumenta più rapidamente quando questi ultimi non sono correlati tra loro che quando sono ridondanti. Il rumore strutturale sarà relativamente alto in un gruppo del genere, perché i giudizi individuali su ogni caso saranno diversi, ma, paradossalmente, la media di questo gruppo, per quanto rumoroso, sarà comunque più accurata della media di un gruppo unanime.

Occorre tuttavia aggiungere una precisazione. A prescindere dalla diversità, l'aggregazione può ridurre il rumore solo se i giudizi sono davvero indipendenti. Come è emerso dal discorso sul rumore nei gruppi, la riflessione di gruppo spesso aggiunge più errore in termini di bias di quanto non ne elimini in termini di rumore. Le organizzazioni che vogliono sfruttare il potenziale della diversità dovranno accettare il disaccordo che sorgerà quando i membri del gruppo formuleranno i propri giudizi indipendenti. Promuovere e aggregare giudizi indipendenti e variegati spesso costituisce la strategia di igiene decisionale più semplice, economica e ampiamente applicabile.

A proposito di selezione e aggregazione

«Calcoliamo la media di quattro giudizi indipendenti. Così facendo, dimezzeremo il rumore.»

«Dovremmo sforzarci di essere in uno stato di perpetual beta*, come i superprevisori.»*

«Prima di affrontare il problema, qual è qui il tasso di base?»

«Abbiamo una buona squadra, ma come possiamo arrivare a una maggiore diversità di opinioni?»

[1] J.A. Frankel, *Over-optimism in Forecasts by Official Budget Agencies and Its Implications*, documento 17239, National Bureau of Economic Research, dicembre 2011, [www.nber.org/papers/w17239].

[2] H.R. Arkes, *Overconfidence in Judgmental Forecasting*, in J. Scott Armstrong (a cura di), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, vol. 30, International Series in Operations Research & Management Science, Springer, Boston 2001.

[3] I. Ben-David, J. Graham, C. Harvey, *Managerial Miscalibration*, in "The Quarterly Journal of Economics", 128(2013), n. 4, pp. 1547-1584.

[4] T.R. Stewart, *Improving Reliability of Judgmental Forecasts*, in J.S. Armstrong (a cura di), *Principles of Forecasting*, cit., p. 82.

[5] T.W. Ruger *et al.*, *The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decision-Making*, in "Columbia Law Review", 104(2004), pp. 1150-1209.

[6] C. Sunstein, *Maximin*, in "Yale Journal of Regulation" (bozza; 3 maggio 2020), [papers.ssrn.com/sol3/papers.cfm?abstract_id=3476250].

[7] Per numerosi esempi, vedi J.S. Armstrong (a cura di), *Principles of Forecasting*, cit.

[8] J.S. Armstrong, *Combining Forecasts*, in Id. (a cura di), *Principles of Forecasting*, cit., pp. 417-439.

[9] T.R. Stewart, *Improving Reliability of Judgmental Forecasts*, cit., p. 95.

[10] J.S. Armstrong, *Combining Forecasts*, cit.

[11] A.E. Mannes *et al.*, *The Wisdom of Select Crowds*, in "Journal of Personality and Social Psychology", 107(2014), n. 2, pp. 276-299.

[12] J. Wolfers, E. Zitzewitz, *Prediction Markets*, in "Journal of Economic Perspectives", 18(2004), pp. 107-126.

[13] C.R. Sunstein, R. Hastie, *Wiser: Getting Beyond Groupthink to Make Groups Smarter*, Harvard Business Review Press, Boston 2014.

[14] G. Rowe, G. Wright, *The Delphi Technique as a Forecasting Tool: Issues and Analysis*, in "International Journal of Forecasting", 15(1999), pp. 353-375. Vedi anche D. Bang, C.D. Frith, *Making Better Decisions in Groups*, in "Royal Society Open Science", 4(2017), n. 8.

[15] R. Hastie, *Review Essay: Experimental Evidence on Group Accuracy*, in B. Grofman, G. Guillermo (a cura di), *Information Pooling and Group Decision Making*, JAI Press, Greenwich, CT 1986, pp. 129-157.

[16] A.H. Van De Ven, A.L. Delbecq, *The Effectiveness of Nominal, Delphi, and Interacting Group Decision Making Processes*, in "Academy of Management Journal", 17(2017), n. 4.

[17] P. Tetlock, D. Gardner, *Superforecasting: The Art and Science of Prediction*, Random House, New York 2016, p. 95.

[18] Ivi, p. 231.

[19] Ivi, p. 273.

[20] V.A. Satopää *et al.*, *Bias, Information, Noise: The BIN Model of Forecasting*, 19 febbraio 2020, p. 23, [dx.doi.org/10.2139/ssrn.3540864].

[21] *Ibid.*

[22] Ivi, p. 22.

[23] Ivi, p. 24.

22

Linee guida in medicina

Alcuni anni fa un nostro amico, che chiameremo Paul, ricevette una diagnosi di ipertensione dal suo medico di base, che chiameremo dottor Jones. Il dottore gli consigliò di provare dei farmaci, prescrivendogli un diuretico, che però non ebbe alcun effetto: la pressione di Paul restava alta. Dopo qualche settimana gli prescrisse un secondo farmaco, un calcio-antagonista, ma anche qui l'effetto fu limitato.

Questi risultati lasciarono perplesso il dottor Jones. Dopo tre mesi di visite settimanali i valori della pressione di Paul calarono leggermente, ma erano ancora alti. Non si sapeva come procedere. Paul era preoccupato e il dottore turbato, anche perché il suo paziente era un uomo relativamente giovane e in buona salute. Prese anche in considerazione l'idea di provare con un terzo farmaco.

A quel punto avvenne che Paul si trasferì in un'altra città, dove consultò un altro medico di base, che chiameremo dottor Smith. Paul raccontò al dottore dei suoi costanti problemi di ipertensione, al che lui gli rispose subito: «Compri un misuratore di pressione e controlli i suoi valori. A me non sembra che abbia la pressione alta. Probabilmente ha solo la sindrome da camice bianco: la sua pressione schizza in alto quando entra in uno studio medico!».

Paul seguì la sua indicazione e, come previsto dal medico, a casa la sua pressione era normale, e così rimase da allora (un mese dopo aver appreso

della sindrome da camice bianco, la pressione si normalizzò anche nello studio del dottore).

Uno dei compiti principali dei medici sta nell'effettuare delle diagnosi per stabilire se un paziente sia affetto da malattia e, nel caso, identificarla. Spesso le diagnosi richiedono un certo tipo di giudizio, e per molti disturbi sono un'operazione di routine, quasi meccanica, che prevede l'applicazione di regole e procedure per ridurre al minimo il rumore. Di solito è facile per un medico stabilire se una persona si è lussata una spalla o si è fratturata un piede. Lo stesso si può dire che valga per problemi più tecnici: la quantificazione del grado di degenerazione di un tendine non produce molto rumore;[1] di fronte a un'agobiopsia delle lesioni al seno, la valutazione del patologo è relativamente semplice e poco soggetta a rumore.[2]

Vi sono poi diagnosi che non richiedono alcun giudizio. Spesso le cure mediche pervengono a un'eliminazione dell'elemento del giudizio, passando da quest'ultimo al calcolo: per una faringite, il medico comincerà con un rapido test antigenico a partire da un prelievo di muco dalla gola del paziente, e in breve tempo il test sarà in grado di rilevare la presenza dello streptococco. (Senza il risultato del test rapido antigenico, e a volte anche con, le diagnosi della faringite sono soggette a rumore.)[3] Con un livello di glicemia a digiuno pari o superiore a 126 milligrammi per decilitro, o un'emoglobina glicata (o HbA1c, che permette di misurare la glicemia media degli ultimi tre mesi) di almeno 6,5, verrà diagnosticato il diabete.[4] Nelle prime fasi della pandemia di Covid-19, alcuni medici effettuavano diagnosi sulla base di un giudizio sui sintomi, poi con il tempo ci si è affidati sempre più ai tamponi, che l'hanno reso superfluo.

Molti sanno che quando i medici esercitano il proprio giudizio possono essere soggetti a rumore, e possono sbagliare; normalmente si consiglia ai

pazienti di chiedere un secondo parere, e anzi in alcuni ospedali è addirittura una prassi obbligatoria.[5] Ogni volta che il secondo parere diverge dal primo, si è in presenza di rumore, anche se naturalmente potrebbe non essere chiaro quale dei due medici abbia ragione. Alcuni pazienti (compreso Paul) restano basiti dalla discrepanza tra il primo e il secondo parere, ma ciò che dovrebbe sorprenderci non è l'esistenza del rumore nella professione medica, quanto la sua pervasività.

In questo capitolo ci proponiamo di motivare quest'affermazione e descrivere alcuni approcci per la riduzione del rumore impiegati in medicina, concentrandoci su una particolare strategia di igiene decisionale: lo sviluppo di linee guida diagnostiche. Siamo assolutamente consapevoli che si potrebbe dedicare un libro intero al rumore in questo campo, e alle varie azioni intraprese da medici, infermieri e ospedali per porvi rimedio. Vale la pena notare che la sua presenza non è circoscritta ai giudizi diagnostici, sui quali qui ci soffermeremo: anche i trattamenti possono essere soggetti a rumore, e vi è un'ampia letteratura anche su questo tema. Di fronte a un paziente con un problema cardiaco, i giudizi dei medici sul miglior trattamento da adottare variano in maniera sconcertante, che si tratti del giusto farmaco, del giusto tipo di intervento o della necessità stessa di un intervento. Il Dartmouth Atlas Project da più di vent'anni documenta «divergenze lampanti nella distribuzione e nell'impiego delle risorse mediche negli Stati Uniti».[6] Simili conclusioni si possono trarre per molte altre nazioni.[7] Per le finalità di questo libro, però, basterà un breve excursus sul rumore nei giudizi diagnostici.

Una panoramica

La letteratura sul rumore in medicina è estesissima. Se molti degli studi sono empirici, tesi a individuare la presenza del rumore, molti altri sono anche normativi: nel settore sanitario si è sempre alla ricerca di strategie di riduzione del rumore, che assumono molte forme e sono una miniera di idee che varrebbe la pena prendere in considerazione in diversi campi.

Quando subentra il rumore, è possibile che un medico abbia chiaramente ragione e un altro chiaramente torto (e forse sia anche soggetto a qualche bias). È ovvio che le competenze hanno un ruolo fondamentale: uno studio sulle diagnosi di polmonite dei radiologi, per esempio, ha riscontrato un tasso significativo di rumore, derivante in gran parte da livelli di competenza diversi.[8] Nello specifico, «la disparità di competenze spiega il 44% delle variazioni nelle decisioni diagnostiche», a indicare che «le azioni per il miglioramento delle competenze personali danno risultati migliori rispetto a linee guida uniformi sulle decisioni». Qui come altrove, formazione e selezione sono evidentemente cruciali per ridurre l'errore, e per eliminare rumore e bias.[9]

In alcune specialità, come la radiologia e la patologia, i medici sono ben consapevoli della presenza del rumore: i radiologi, per esempio, definiscono la variazione diagnostica il loro «tallone d'Achille».[10] Non è chiaro se il rumore in questi due campi riceva una particolare attenzione perché sia realmente più accentuato che in altri o perché qui è più facile documentarlo, ma sospettiamo prevalga la seconda motivazione. In radiologia è più facile condurre test semplici e chiari del rumore (e talvolta dell'errore): per esempio, si può tornare su lastre o immagini digitali per rivedere un giudizio precedente.

In medicina, il rumore interpersonale – o *affidabilità inter-rater* – normalmente si misura con la *statistica kappa*.[11] Più è alto questo valore, minore sarà il rumore: un kappa di 1 riflette un accordo perfetto, mentre lo

0 indica l'accordo che ci si aspetterebbe di trovare tra scimmie che lanciano freccette su una lista di possibili diagnosi. In certi tipi di diagnosi mediche l'affidabilità misurata da questo coefficiente si è dimostrata "scarsa" o "insufficiente", a significare che il rumore è molto alto. Spesso risulta "discreta", che naturalmente è già meglio, ma indica ancora una notevole presenza di rumore. Sul problema importante di quali interazioni tra farmaci siano clinicamente significative, i medici generici che hanno esaminato cento interazioni selezionate in maniera casuale hanno mostrato un «basso grado di accordo».[12] Per i non esperti e per molti medici, le diagnosi di malattie renali di diversa gravità potrebbero sembrare relativamente limpide, ma i nefrologi mostrano solo «un accordo scarso o moderato» nei propri giudizi sul significato degli esami standard effettuati dai loro pazienti.[13]

Alla domanda se una lesione al seno fosse cancerosa, uno studio ha riscontrato un accordo soltanto "discreto" tra i patologi,[14] così come "discreto" era anche quello sulla diagnosi di lesioni proliferative al seno.[15] Sempre "discreto" era l'accordo dei medici sulle valutazioni delle risonanze magnetiche rispetto al livello di stenosi spinale.[16] Vale la pena di soffermarsi su questi dati. Abbiamo detto che, in certi tipi di diagnosi, il livello di rumore in medicina è molto basso, ma anche in aree piuttosto tecniche i medici ne risentono. Che a un paziente venga o non venga diagnosticata una malattia grave come un cancro potrebbe dipendere da una sorta di lotteria, determinata dal particolare medico da cui verrà visitato.

Consideriamo alcuni altri risultati riportati nella letteratura scientifica, relativi a campi in cui il livello di rumore sembra particolarmente rilevante. Descriviamo tali dati non per dare giudizi perentori sullo stato attuale della prassi medica, che continua a evolversi e a migliorare (in certi casi anche

rapidamente), ma per fornire un'idea generale della pervasività del rumore, nel recente passato come nel presente.

1. Le cardiopatie sono la prima causa di morte per uomini e donne negli Stati Uniti.[17] L'angiografia coronarica, uno dei metodi principali per la rilevazione della cardiopatia, valuta la presenza di occlusioni nelle arterie coronarie in fase acuta e non. Nelle fasi non acute, quando un paziente presenta un dolore al petto ricorrente, spesso viene seguito un trattamento – per esempio il posizionamento di uno stent – se più del 70% di una o più arterie risulta ostruito. Tuttavia, si è riscontrato un certo grado di variabilità nell'interpretazione delle angiografie, che potrebbe condurre a procedure non necessarie: da uno studio iniziale, infatti, è emerso che nel 31% dei casi tra i medici vi è un disaccordo sul fatto che in un grande vaso vi sia un'ostruzione superiore al 70%.[18] Malgrado la diffusa consapevolezza da parte dei cardiologi della potenziale variabilità nella lettura delle angiografie, e nonostante i continui sforzi e correttivi attuati in questo senso, il problema non è ancora stato risolto.

2. L'endometriosi è una malattia determinata dall'accumulo fuori dall'utero di cellule endometriali che dovrebbero trovarsi invece al suo interno. Si tratta di un disturbo doloroso, che può comportare problemi di sterilità. Spesso viene diagnosticato mediante laparoscopia, una tecnica che prevede l'inserimento di una piccola videocamera all'interno del corpo attraverso un intervento chirurgico. I filmati digitali di laparoscopie eseguite su tre pazienti, due affette da endometriosi di diversa gravità e una no, sono stati mostrati a centootto chirurghi ginecologi, ai quali è stato chiesto di esprimere un giudizio sul numero e la posizione delle lesioni dell'endometrio. Si è riscontrato un fortissimo disaccordo, con deboli correlazioni sia rispetto al numero sia alla posizione.[19]

3. La tubercolosi è una delle malattie più diffuse e mortali al mondo: solo nel 2016 ne erano affette più di dieci milioni di persone, per un totale di due milioni di vittime. Un metodo ampiamente adottato per l'individuazione della patologia è la radiografia toracica, volta a rilevare le lesioni cavitarie polmonari provocate dai batteri. La variabilità delle diagnosi di tubercolosi è ben documentata da circa settantacinque anni, ma, malgrado i progressi compiuti nel corso dei decenni, gli studi continuano a riscontrarne un grado elevato, con un accordo inter-rater "moderato" o appena "discreto".[20] Vi è inoltre una variabilità nelle diagnosi di tubercolosi tra radiologi provenienti da paesi diversi.[21]

4. Quando alcuni patologi hanno analizzato diverse lesioni cutanee per la diagnosi del melanoma – la forma più grave di tumore della pelle –, si è registrato un accordo appena "moderato": gli otto medici a cui sono stati sottoposti i casi hanno dato parere unanime o mostrato un unico punto di disaccordo soltanto nel 62% delle valutazioni.[22] Un altro studio condotto presso un istituto oncologico ha riscontrato che l'accuratezza delle diagnosi di melanoma non superava

il 64%, a indicare diagnosi errate per una lesione su tre.[23] Da un terzo studio è emerso che i dermatologi della New York University non erano stati in grado di diagnosticare un melanoma mediante biopsia nel 36% dei casi; i ricercatori concludono che «l'insuccesso dei clinici nella diagnosi del melanoma ha gravi implicazioni per la sopravvivenza dei pazienti affetti da una patologia potenzialmente fatale».[24]

5. Vi è variabilità nei giudizi dei radiologi sulle mammografie per l'individuazione del cancro al seno. Un ampio studio ha indicato che l'intervallo dei falsi negativi tra diversi radiologi variava dallo 0% (il giudizio era sempre corretto) a oltre il 50% (il giudizio dei radiologi si era dimostrato errato in più della metà dei casi). Analogamente, il tasso di falsi positivi variava da meno dell'1% al 64% (a indicare che in circa due terzi dei casi il radiologo sosteneva che la mammografia indicasse la presenza di un cancro laddove non era presente).[25] I falsi negativi e i falsi positivi di diversi radiologi indicano la presenza di rumore.

Questi casi di rumore interpersonale prevalgono nelle ricerche esistenti, ma vi sono anche dati che attestano il rumore occasionale. Nel rivalutare la stessa immagine talvolta i radiologi danno un'indicazione diversa che contrasta con quella già espressa da loro stessi (anche se capita più spesso che siano in disaccordo con altri).[26] Valutando il livello di ostruzione in un'angiografia, ventidue medici si sono dimostrati in disaccordo con la loro stessa analisi tra il 63% e il 92% delle volte.[27] In campi caratterizzati da criteri vaghi e giudizi complessi, l'affidabilità intra-rater, ovvero quella legata alla variabilità delle decisioni di uno stesso valutatore, può essere scarsa.[28]

Questi studi non offrono una chiara spiegazione del rumore occasionale, ma un'altra ricerca, non basata sulle diagnosi, identifica una semplice fonte di rumore occasionale in medicina, un risultato che sia i medici sia i pazienti dovrebbero tenere a mente: in breve, è molto più probabile che i medici prescrivano uno screening per la prevenzione del cancro la mattina presto che nel tardo pomeriggio.[29] In un ampio campione, il tasso più alto di prescrizioni di screening al seno e al colon, pari al 63,7%, si riscontrava alle otto di mattina, per poi ridursi nel corso della mattinata fino a

raggiungere il 48,7% alle undici. Dopodiché aumentava di nuovo, arrivando a toccare il 56,2% a mezzogiorno, e si riduceva fino al 47,8% alle diciassette. Ne consegue che i pazienti che avevano fissato una visita a fine giornata avevano meno probabilità di sottoporsi a uno screening contro il cancro, raccomandato dalle linee guida mediche.

Come si spiegano simili risultati? Una possibile risposta è che inevitabilmente i medici accumulano un ritardo nel loro programma giornaliero dopo aver visitato pazienti con problemi medici complessi, che richiedono tempi superiori ai consueti venti minuti. Abbiamo già fatto riferimento al ruolo dello stress e della stanchezza nell'attivazione del rumore occasionale (vedi capitolo 7): ebbene, questi elementi sembrano intervenire anche qui. Per stare al passo con l'orario di ambulatorio, alcuni medici saltano il colloquio sulle misure di prevenzione. Un'altra indicazione del ruolo giocato dalla stanchezza nella pratica clinica è data dal basso tasso di lavaggi appropriati delle mani verso la fine dei turni ospedalieri.[30] (Scopriamo così che anche lavarsi le mani è un'operazione affetta da rumore.)

Medici meno soggetti a rumore: il valore delle linee guida

Un resoconto esaustivo dell'esistenza e della portata del rumore in diversi contesti medici sarebbe di grande utilità non solo per la medicina, ma per il sapere umano in generale.[31] Non siamo a conoscenza di studi di questo tipo, e speriamo che, a tempo debito, ne vengano effettuati. Tuttavia, anche i risultati già a disposizione ci offrono alcuni indizi.

A un estremo si collocano le diagnosi di tipo sostanzialmente meccanico di alcuni disturbi o malattie, che non lasciano spazio al giudizio. In altri casi, pur non essendo meccanica, la diagnosi è comunque semplice, ed è

molto probabile che chiunque abbia una formazione medica pervenga alle stesse conclusioni. Vi sono poi situazioni in cui un certo grado di specializzazione – poniamo, tra gli specialisti di cancro ai polmoni – sarà sufficiente per garantire che il rumore, seppur presente, sia minimo. All'altro estremo troviamo casi che lasciano ampio spazio al giudizio, con criteri diagnostici così indefiniti che il rumore sarà consistente e difficile da ridurre: come vedremo, è ciò che accade per molti disturbi psichiatrici.

Come ridurre il rumore in medicina? Come si diceva, la formazione può accrescere le competenze,[32] le quali sono senz'altro d'aiuto, come lo è l'aggregazione dei giudizi di più esperti (seconde opinioni e così via).[33] Gli algoritmi costituiscono una strada particolarmente promettente, e oggi i medici impiegano intelligenze artificiali e algoritmi di apprendimento profondo: questi ultimi, per esempio, sono stati impiegati per identificare metastasi linfonodali in donne affette da cancro al seno. I migliori si sono dimostrati superiori ai patologi più validi, e naturalmente non sono soggetti a rumore.[34] Sono stati impiegati algoritmi di apprendimento profondo con notevole successo anche nell'identificazione di problemi agli occhi associati al diabete,[35] e oggi l'intelligenza artificiale dà risultati almeno pari a quelli dei radiologi nell'individuazione del cancro a partire dalle mammografie; probabilmente ulteriori progressi ne attesteranno la superiorità.[36]

È ipotizzabile che in futuro la professione medica si avvarrà sempre di più degli algoritmi, che promettono di ridurre il bias e il rumore, e quindi di salvare vite e risparmiare risorse. Qui, però, ci concentreremo sulle linee guida come ausilio ai giudizi umani, perché il campo della medicina illustra in maniera molto istruttiva come siano in grado di produrre risultati buoni o anche eccellenti in talune applicazioni, e risultati più eterogenei in altre.

Forse l'esempio più famoso di linee guida diagnostiche è l'indice di Apgar, sviluppato nel 1952 dall'anestetista ostetrica Virginia Apgar. Fino ad allora, la valutazione delle condizioni di salute del neonato alla nascita era assegnata al giudizio clinico di medici e ostetriche. Il nuovo strumento fornì al personale sanitario delle linee guida standard per valutare il neonato sulla base di cinque parametri: il colorito della pelle, la frequenza cardiaca, i riflessi, il tono muscolare e l'attività respiratoria. Questi parametri vengono riassunti in inglese proprio nell'acronimo Apgar, dal cognome dell'ideatrice dell'indice: *appearance* (colore della pelle), *pulse* (frequenza cardiaca), *grimace* (riflessi), *activity* (tono muscolare) e *respiration* (frequenza respiratoria e sforzo respiratorio). Nell'esame di Apgar a ciascuno di questi parametri è assegnato un punteggio di 0, 1 o 2. Il massimo totalizzabile, molto raro, è 10, mentre un punteggio pari o superiore a 7 è considerato un indice di buona salute (vedi tabella 3).

Tabella 3. *Linee guida dell'indice di Apgar*[37]

Categoria	*Numero di punti assegnati*
Colorito	0: Corpo interamente pallido o cianotico 1: Buon colore corporeo ma estremità cianotiche 2: Colorito roseo o normale
Frequenza cardiaca	0: Assente 1: < 100 battiti/min 2: > 100 battiti/min
Riflessi	0: Assenti dopo la stimolazione delle vie respiratorie 1: Smorfia durante la stimolazione 2: Smorfia e tosse o starnuto durante la stimolazione
Tono muscolare	0: Assente (atonia) 1: Flessione accennata degli arti 2: Movimenti attivi

Respirazione	0: Assente 1: Pianto debole e irregolare 2: Vigorosa, con pianto

Si noti che la frequenza cardiaca è l'unico parametro strettamente numerico del punteggio, mentre tutti gli altri prevedono un elemento di giudizio; ma poiché questo viene scomposto in elementi discreti semplici da valutare, anche il personale con un livello mediocre di preparazione difficilmente si troverà in forte disaccordo, quindi si può dire che l'indice Apgar produca poco rumore.[38]

Questo strumento illustra bene come funzionano le linee guida e perché riducono il rumore. Contrariamente alle regole e agli algoritmi, non eliminano la necessità del giudizio: la decisione non si riduce a un mero calcolo. Il disaccordo è ancora possibile su ciascun parametro, e quindi sulle conclusioni che ne derivano, ma le linee guida riescono a ridurre il rumore perché scompongono una decisione complessa in un certo numero di sottogiudizi in merito a dimensioni predefinite.

I vantaggi di questo approccio sono chiari se si considera il problema nell'ottica dei modelli di predizione semplici discussi nel capitolo 9. Un clinico che esprime un giudizio sulla salute di un neonato si basa su diversi segnali predittivi, e quindi c'è il rischio che subentri del rumore occasionale: un certo giorno, o perché è di un certo umore, potrebbe prestare attenzione a predittori relativamente insignificanti o trascurarne altri importanti. L'indice di Apgar, invece, induce il personale sanitario a concentrarsi sui cinque che, empiricamente, si sono dimostrati importanti, e inoltre fornisce una chiara descrizione di come valutare ciascun segnale, il che semplifica notevolmente ogni singolo giudizio, riducendo in tal modo il rumore. Infine, l'indice specifica come pesare meccanicamente i predittori per arrivare al giudizio complessivo richiesto, laddove i clinici

assegnerebbero pesi differenti ai vari segnali. Concentrarsi sui predittori pertinenti, semplificare il modello predittivo e procedere a un'aggregazione meccanica: tutto questo riduce il rumore.

Approcci analoghi sono stati impiegati in molti campi della medicina: un esempio è rappresentato dai criteri Centor nella diagnosi di faringite streptococcica. Al paziente viene attribuito un punto per ciascuno dei seguenti sintomi o segnali (che, come per l'indice di Apgar, in inglese costituiscono l'acronimo inverso del cognome di Robert Centor, colui che per primo, insieme ai suoi colleghi, elaborò queste linee guida): assenza di tosse (*cough*), presenza di essudati tonsillari (*exudates*, placche biancastre in fondo alla gola), linfoadenite cervicale (*nodes*), presenza di febbre superiore ai 38 gradi (*temperature*). A seconda del punteggio assegnato al paziente, può essere consigliato un test rapido con un tampone da sfregare nel cavo orale per la diagnosi di faringite streptococcica. La valutazione e il punteggio sono relativamente semplici utilizzando questa scala, che in effetti ha ridotto il numero di test e trattamenti non necessari per questa infezione.[39]

Analogamente, sono state sviluppate delle linee guida per la diagnosi di cancro al seno note come Breast Imaging Reporting and Data System (BI-RADS), che riducono il rumore nell'interpretazione delle mammografie. Uno studio ha dimostrato che questo sistema ha condotto a un aumento dell'accordo nella valutazione degli screening mammografici, dimostrando che le linee guida possono essere efficaci nella riduzione del rumore in un campo in cui la variabilità è significativa.[40] In ambito patologico, molti tentativi di impiego delle linee guida per finalità analoghe hanno dato ottimi risultati.[41]

I risultati deprimenti della psichiatria

In termini di rumore, la psichiatria è un caso estremo. Quando si tratta di assegnare una diagnosi agli stessi pazienti adottando gli stessi criteri diagnostici, spesso gli psichiatri sono in disaccordo tra loro, ed è per questo che, almeno dagli anni quaranta, la riduzione del rumore è una grande priorità nel settore.[42] E, come vedremo, nonostante i continui ritocchi le linee guida non si sono dimostrate molto d'aiuto in questo senso.

Uno studio del 1964 che coinvolgeva novantuno pazienti e dieci psichiatri esperti rilevò che la probabilità di trovare un accordo tra due opinioni non superava il 57%.[43] Un altro di questi primi studi sul tema, in cui erano coinvolti 426 pazienti ospedalieri che erano stati visitati da due psichiatri, riscontrò un accordo solo nel 50% dei casi per le diagnosi indipendenti di malattia mentale dei due specialisti. Una ricerca che coinvolse 153 pazienti ambulatoriali rilevò un accordo del 54%. In questi studi la fonte di rumore non era specificata, ma è interessante notare che alcuni psichiatri risultarono inclini ad assegnare i pazienti a specifiche categorie diagnostiche. Qualcuno, per esempio, era particolarmente propenso a diagnosticare casi di depressione, qualcun altro casi di ansia.

Come vedremo, continuano a esserci alti livelli di rumore in psichiatria. Ma quale può essere la causa? Tra gli specialisti non si è arrivati a un'unica risposta chiara (il che significa che anche spiegare la presenza di rumore è un processo affetto da rumore). L'ampia serie di categorie diagnostiche è senz'altro un fattore da considerare. In un primo tentativo di trovare risposta a questa domanda, i ricercatori chiesero a uno psichiatra di avere un colloquio con un paziente, e poi a un secondo psichiatra, dopo una breve pausa, di averne un altro con la stessa persona, dopodiché i due specialisti si riunivano e, se non erano d'accordo, si interrogavano sulle motivazioni.[44]

Spesso uno dei motivi era «l'incostanza del medico»: diverse scuole di pensiero, diversi percorsi formativi, diverse esperienze cliniche, diversi stili di colloquio. Mentre un «clinico con una formazione in psichiatria dell'età evolutiva poteva spiegare l'esperienza allucinatoria in termini di esperienza post traumatica di un maltrattamento subìto», un altro «di orientamento biomedico poteva ricondurre le stesse allucinazioni al processo schizofrenico».[45] Tali differenze sono tipici esempi di rumore strutturale.

Al di là delle differenze tra i medici, però, la principale ragione alla base del rumore era «l'inadeguatezza della nomenclatura». Queste osservazioni, unite alla diffusa insoddisfazione professionale rispetto alla nomenclatura psichiatrica, condussero a una revisione del *Manuale diagnostico e statistico dei disturbi mentali* (*DSM-III*) nella terza edizione del 1980. Il testo includeva per la prima volta criteri espliciti e dettagliati per la diagnosi di questo tipo di patologie, un primo passo verso l'introduzione di linee guida diagnostiche.

Il *DSM-III* condusse a un netto aumento delle ricerche volte a indagare l'eventuale presenza di rumore nelle diagnosi,[46] e si dimostrò utile nel ridurlo. Eppure non fu un grande successo.[47] Anche dopo la significativa revisione della quarta edizione del 2000, o *DSM-IV* (pubblicato originariamente nel 1994), le ricerche mostrarono che il livello di rumore restava elevato.[48] Da una parte, Ahmed Aboraya e i suoi colleghi concludevano che «l'impiego di criteri diagnostici per i disturbi psichiatrici ha portato a un comprovato incremento dell'affidabilità delle diagnosi psichiatriche».[49] Dall'altra, c'era ancora un alto rischio che «le schede di ricovero per un singolo paziente facciano emergere diagnosi multiple per lo stesso individuo».[50]

Nel 2013 venne pubblicata un'altra versione del manuale, il *DSM-5*.[51] La American Psychiatric Association sperava che avrebbe ridotto il rumore, in

quanto la nuova edizione si affidava a criteri più oggettivi e chiaramente disposti in scala,[52] ma le diagnosi degli psichiatri continuavano a esserne affette a un livello significativo.[53] Samuel Lieblich e i suoi colleghi, per esempio, hanno riscontrato che «gli psichiatri hanno grosse difficoltà a convenire su chi sia affetto da un grave disturbo depressivo e chi no».[54] La sperimentazione clinica del *DSM-5* rivelò un «accordo minimo», che «significa che gli alti specialisti di psichiatria coinvolti nello studio concordavano sulla depressione del paziente in una percentuale di casi che andava dal 4 al 15%».[55] Secondo alcune sperimentazioni sul campo, il *DSM-5* non faceva altro che peggiorare le cose, evidenziando un maggior grado di rumore «in tutti i principali ambiti, con alcune diagnosi, quali il disturbo misto ansioso-depressivo, [...] così inaffidabili da dimostrarsi inservibili nella pratica clinica».[56]

Sembra che il motivo principale del limitato successo delle linee guida sia stato il fatto che, in psichiatria, «i criteri diagnostici di alcuni disturbi sono ancora vaghi e difficili da rendere operativi».[57] Alcune linee guida riducono il rumore scomponendo il giudizio in criteri su cui c'è ridotto disaccordo, ma trattandosi di criteri relativamente indefiniti, la possibilità del rumore resta. A partire da ciò, sono state avanzate autorevoli proposte per la definizione di linee guida diagnostiche più standardizzate, che chiedono di (1) chiarire i criteri diagnostici, abbandonando gli standard vaghi; (2) stabilire «definizioni di riferimento» dei sintomi e del loro livello di gravità, partendo dall'assunto che quando «i clinici concordano sulla presenza o sull'assenza di sintomi, saranno più inclini a concordare sulle diagnosi»; (3) impiegare interviste strutturate dei pazienti in aggiunta ai colloqui aperti.[58] È stata proposta una guida per le interviste basata su ventiquattro domande preliminari di screening, che ha condotto a diagnosi più affidabili, per esempio, dei disturbi ansiosi, depressivi e alimentari.

Questi sviluppi sembrano promettenti, ma resta da appurare fino a che punto riescano a ridurre il rumore. Per citare un osservatore, «il fare affidamento sui sintomi soggettivi del paziente, l'interpretazione degli stessi da parte del clinico e l'assenza di misure oggettive (come può esserlo un'analisi del sangue) gettano i semi dell'inaffidabilità diagnostica dei disturbi psichiatrici».[59] In questo senso la psichiatria potrebbe dimostrarsi particolarmente resistente ai tentativi di riduzione del rumore.

Su questo particolare problema è troppo presto per avanzare valide previsioni, ma una cosa è certa: nella medicina in generale le linee guida si sono dimostrate molto efficaci nella riduzione del bias e del rumore. Sono state d'aiuto a medici, infermieri e pazienti, con un netto miglioramento per la salute pubblica. In ambito medico il loro impiego andrebbe incrementato.[60]

A proposito delle linee guida in medicina

«Tra i medici, il livello di rumore è decisamente più alto di quanto potessimo aspettarci. Nelle diagnosi di cancro e cardiopatie (e perfino nella lettura delle radiografie) talvolta gli specialisti sono in disaccordo. Ciò significa che il trattamento prescritto al paziente potrebbe essere il risultato di una lotteria.»

«Ai medici piace pensare che le loro decisioni non varino a seconda del giorno della settimana o dell'ora del giorno, ma si è scoperto che ciò che i medici dicono e fanno potrebbe dipendere dal loro livello di stanchezza.»

«Le linee guida mediche possono ridurre le probabilità che gli specialisti prendano abbagli a spese dei pazienti. Potrebbero inoltre essere d'aiuto a tutte le professioni sanitarie, in quanto riducono la variabilità.»

[1] L. Horton *et al.*, *Development and Assessment of Inter- and Intra-Rater Reliability of a Novel Ultrasound Tool for Scoring Tendon and Sheath Disease: A Pilot Study*, in "Ultrasound", 24(2016), n. 3, p. 134, [www.ncbi.nlm.nih.gov/pmc/articles/PMC5105362].

[2] L.C. Collins *et al.*, *Diagnostic Agreement in the Evaluation of Image-guided Breast Core Needle Biopsies*, in "American Journal of Surgical Pathology", 28(2004), p. 126, [https://bit.ly/3AeBt0r].

[3] J.L. Fierro *et al.*, *Variability in the Diagnosis and Treatment of Group A Streptococcal Pharyngitis by Primary Care Pediatricians*, in "Infection Control and Hospital Epidemiology", 35(2014), n. S3, p. S79, [www.jstor.org/stable/10.1086/677820].

[4] Esame per la diagnosi del diabete, Centers for Disease Control and Prevention, [www.cdc.gov/diabetes/basics/getting-tested.html].

[5] J.D. Kronz *et al.*, *Mandatory Second Opinion Surgical Pathology at a Large Referral Hospital*, in "Cancer", 86(1999), p. 2426, [https://bit.ly/3h7zDaa].

[6] Gran parte del materiale si può trovare online; un libro dedicato al tema è Dartmouth Medical School, *The Quality of Medical Care in the United States: A Report on the Medicare Program; the Dartmouth Atlas of Health Care 1999*, American Hospital Publishers, Washington, DC 1999.

[7] Vedi, per esempio, OECD, *Geographic Variations in Health Care: What Do We Know and What Can Be Done to Improve Health System Performance?*, OECD Publishing, Paris 2014, pp. 137-169; M.P. Hurley *et al.*, *Geographic Variation in Surgical Outcomes and Cost Between the United States and Japan*, in "*American Journal of Managed Care*", 22(2016), p. 600, [https://bit.ly/3xapEXb]; J. Appleby *et al.*, *Variations in Health Care: The Good, the Bad and the Inexplicable*, The King's Fund, London 2011, [https://bit.ly/3qzAljw].

[8] D.C. Chan Jr. *et al.*, *Selection with Variation in Diagnostic Skill: Evidence from Radiologists*, National Bureau of Economic Research, documento NBER, n. 26467, novembre 2019, [www.nber.org/papers/w26467].

[9] P.J. Robinson, *Radiology's Achilles' Heel: Error and Variation in the Interpretation of the Rontgen Image*, in "British Journal of Radiology", 70(1997), p. 1085, [www.ncbi.nlm.nih.gov/pubmed/9536897]. Uno studio sullo stesso tema è Y. Tsugawa *et al.*, *Physician Age and Outcomes in Elderly Patients in Hospital in the US: Observational Study*, in "BMJ", 357(2017), [www.bmj.com/content/357/bmj.j1797], che rileva come i risultati dei medici peggiorano man mano che si allontanano dagli anni della formazione, come se il prezzo da pagare per la maggiore esperienza derivante da anni di pratica fosse la perdita di familiarità con i dati e le

linee guida più recenti. Lo studio fa emergere che i risultati migliori provengono dai medici nei primi anni di internato, che hanno quei dati freschi nella mente.

[10] P.J. Robinson, *Radiology's Achilles' Heel*, cit.

[11] Come il coefficiente di correlazione, il valore kappa può essere negativo, anche se nella pratica questo avviene raramente. Ecco una caratterizzazione del significato di diverse statistiche kappa: «scarsa (κ = da 0,00 a 0,20), discreta (κ = da 0,21 a 0,40), moderata (κ = da 0,41 a 0,60), significativa (κ = da 0,61 a 0,80) e quasi perfetta (κ > 0,80)» (R. Wald *et al.*, *Interobserver Reliability of Urine Sediment Interpretation*, in "Clinical Journal of the American Society of Nephrology", 4 [marzo 2009], n. 3, pp. 567-571, [cjasn.asnjournals.org/content/4/3/567]).

[12] H.R. Strasberg *et al.*, *Inter-Rater Agreement Among Physicians on the Clinical Significance of Drug-Drug Interactions*, in "AMIA Annual Symposium Proceedings", (2013), p. 1325, [www.ncbi.nlm.nih.gov/pmc/articles/PMC3900147].

[13] R. Wald *et al.*, *Interobserver Reliability of Urine Sediment Interpretation*, [cjasn.asnjournals.org/content/4/3/567].

[14] J.P. Palazzo *et al.*, *Hyperplastic Ductal and Lobular Lesions and Carcinomas in Situ of the Breast: Reproducibility of Current Diagnostic Criteria Among Community- and Academic-Based Pathologists*, in "Breast Journal", 4(2003), p. 230, [www.ncbi.nlm.nih.gov/pubmed/21223441].

[15] R.K. Jain *et al.*, *Atypical Ductal Hyperplasia: Interobserver and Intraobserver Variability*, in "Modern Pathology", 24(2011), p. 917, [www.nature.com/articles/modpathol201166].

[16] A.C. Speciale *et al.*, *Observer Variability in Assessing Lumbar Spinal Stenosis Severity on Magnetic Resonance Imaging and Its Relation to Cross-Sectional Spinal Canal Area*, in "Spine", 27(2002), p. 1082, [www.ncbi.nlm.nih.gov/pubmed/12004176].

[17] Centers for Disease Control and Prevention, *Heart Disease Facts*, [www.cdc.gov/heartdisease/facts.htm].

[18] T.A. DeRouen *et al.*, *Variability in the Analysis of Coronary Arteriograms*, in "Circulation", 55(1977), p. 324, [www.ncbi.nlm.nih.gov/pubmed/832349].

[19] O. Buchweitz *et al.*, *Interobserver Variability in the Diagnosis of Minimal and Mild Endometriosis*, in "European Journal of Obstetrics & Gynecology and Reproductive Biology", 122(2005), p. 213, [www.ejog.org/article/S0301-2115(05)00059-X/pdf].

[20] J.P. Zellweger *et al.*, *Intra-observer and Overall Agreement in the Radiological Assessment of Tuberculosis*, in "International Journal of Tuberculosis & Lung Disease", 10(2006), p. 1123, [www.ncbi.nlm.nih.gov/pubmed/17044205]. Sull'accordo "discreto", vedi Y. Balabanova *et al.*, *Variability in Interpretation of Chest Radiographs Among Russian Clinicians and Implications for Screening Programmes: Observational Study*, in "BMJ", 331(2005), p. 379, [www.bmj.com/content/331/7513/379].

[21] S. Sakurada *et al.*, *Inter-Rater Agreement in the Assessment of Abnormal Chest X-Ray Findings for Tuberculosis Between Two Asian Countries*, in "BMC Infectious Diseases", 12(2012), articolo 31,

[bmcinfectdis.biomedcentral.com/articles/10.1186/1471-2334-12-31].

[22] E.R. Farmer *et al.*, *Discordance in the Histopathologic Diagnosis of Melanoma and Melanocytic Nevi Between Expert Pathologists*, in "Human Pathology", 27(1996), p. 528, [www.ncbi.nlm.nih.gov/pubmed/8666360].

[23] A.W. Kopf, M. Mintzis, R.S. Bart, *Diagnostic Accuracy in Malignant Melanoma*, in "Archives of Dermatology", 111(1975), p. 1291, [www.ncbi.nlm.nih.gov/pubmed/1190800].

[24] M. Miller, A.B. Ackerman, *How Accurate Are Dermatologists in the Diagnosis of Melanoma? Degree of Accuracy and Implications*, in "Archives of Dermatology", 128(1992), p. 559, [jamanetwork.com/journals/jamadermatology/fullarticle/554024].

[25] C.A. Beam *et al.*, *Variability in the Interpretation of Screening Mammograms by US Radiologists*, in "Archives of Internal Medicine", 156(1996), p. 209, [www.ncbi.nlm.nih.gov/pubmed/8546556].

[26] P.J. Robinson *et al.*, *Variation Between Experienced Observers in the Interpretation of Accident and Emergency Radiographs*, in "British Journal of Radiology", 72(1999), p. 323, [www.birpublications.org/doi/pdf/10.1259/bjr.72.856.10474490].

[27] K.M. Detre *et al.*, *Observer Agreement in Evaluating Coronary Angiograms*, in "Circulation", 52(1975), p. 979, [www.ncbi.nlm.nih.gov/pubmed/1102142].

[28] L. Horton *et al.*, *Inter- and Intra-Rater Reliability*, cit.; M. Banky *et al.*, *Inter- and Intra-Rater Variability of Testing Velocity When Assessing Lower Limb Spasticity*, in "Journal of Rehabilitation Medicine", 51(2019), [https://bit.ly/3qDeVC0].

[29] E.Y. Hsiang *et al.*, *Association of Primary Care Clinic Appointment Time with Clinician Ordering and Patient Completion of Breast and Colorectal Cancer Screening*, in "JAMA Network Open", 51(2019), [jamanetwork.com/journals/jamanetworkopen/fullarticle/2733171].

[30] H. Dai *et al.*, *The Impact of Time at Work and Time Off from Work on Rule Compliance: The Case of Hand Hygiene in Health Care*, in "Journal of Applied Psychology", 100(2015), p. 846, [www.ncbi.nlm.nih.gov/pubmed/25365728].

[31] A.S. Raja, *The HEART Score Has Substantial Interrater Reliability*, in "NEJM J Watch", 5 dicembre 2018, [https://bit.ly/3hifkpy] (recensione di C.A. Gershon *et al.*, *Inter-rater Reliability of the HEART Score*, in "Academic Emergency Medicine", 26[2019], p. 552).

[32] J.P. Zellweger *et al.*, *Intra-observer and Overall Agreement in the Radiological Assessment of Tuberculosis*, in "International Journal of Tuberculosis & Lung Disease", 10(2006), p. 1123, [www.ncbi.nlm.nih.gov/pubmed/17044205]; I. Abubakar *et al.*, *Diagnostic Accuracy of Digital Chest Radiography for Pulmonary Tuberculosis in a UK Urban Population*, in "European Respiratory Journal", 35(2010), p. 689, [erj.ersjournals.com/content/35/3/689.short].

[33] M.L. Barnett *et al.*, *Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians*, in "JAMA Network Open", 2(2019), e19009, [jamanetwork.com/journals/jamanetworkopen/fullarticle/2726709]; K.H. Allison *et al.*,

Understanding Diagnostic Variability in Breast Pathology: Lessons Learned from an Expert Consensus Review Panel, in "Histopathology", 65(2014), p. 240, [onlinelibrary.wiley.com/doi/abs/10.1111/his.12387].

[34] B.E. Bejnordi *et al.*, *Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer*, in "JAMA", 318(2017), p. 2199, [jamanetwork.com/journals/jama/fullarticle/2665774].

[35] V. Gulshan *et al.*, *Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs*, in "JAMA", 316(2016), p. 2402, [jamanetwork.com/journals/jama/fullarticle/2588763].

[36] M.B. Massat, *A Promising Future for AI in Breast Cancer Screening*, in "Applied Radiology", 47(2018), p. 22, [https://bit.ly/3w6zBDU]; A. Rodriguez-Ruiz *et al.*, *Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison with 101 Radiologists*, in "Journal of the National Cancer Institute", 111(2019), p. 916, [academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djy222/5307077].

[37] Medline Plus, [medlineplus.gov/ency/article/003402.htm].

[38] L.R. Foster *et al.*, *The Interrater Reliability of Apgar Scores at 1 and 5 Minutes*, in "Journal of Investigative Medicine", 54(2006), n. 1, p. 293, [jim.bmj.com/content/54/1/S308.4].

[39] W.J. McIsaac *et al.*, *Empirical Validation of Guidelines for the Management of Pharyngitis in Children and Adults*, in "JAMA", 291(2004), p. 1587, [www.ncbi.nlm.nih.gov/pubmed/15069046].

[40] E.A. Ooms *et al.*, *Mammography: Interobserver Variability in Breast Density Assessment*, in "Breast", 16(2007), p. 568, [https://bit.ly/2UadEHa].

[41] F.P. O'Malley *et al.*, *Interobserver Reproducibility in the Diagnosis of Flat Epithelial Atypia of the Breast*, in "Modern Pathology", 19(2006), p. 172, [www.nature.com/articles/3800514].

[42] Vedi A. Aboraya *et al.*, *The Reliability of Psychiatric Diagnosis Revisited*, in "Psychiatry (Edgmont)", 3(2006), p. 41, [www.ncbi.nlm.nih.gov/pmc/articles/PMC2990547]. Per una panoramica, vedi N. Kreitman, *The Reliability of Psychiatric Diagnosis*, in "Journal of Mental Science", 107(1961), pp. 876-886, [https://bit.ly/3du1wHh].

[43] A. Aboraya *et al.*, *The Reliability of Psychiatric Diagnosis Revisited*, cit., p. 43.

[44] C.H. Ward *et al.*, *The Psychiatric Nomenclature: Reasons for Diagnostic Disagreement*, in "Archives of General Psychiatry", 7(1962), p. 198.

[45] A. Aboraya *et al.*, *The Reliability of Psychiatric Diagnosis Revisited*, cit.

[46] S.M. Lieblich *et al.*, *High Heterogeneity and Low Reliability in the Diagnosis of Major Depression Will Impair the Development of New Drugs*, in "British Journal of Psychiatry Open", 1(2015), p. e5-e7, [https://bit.ly/2UgF6Tu].

[47] *Ibid.*

[48] Vedi E. Cheniaux *et al.*, *The Diagnoses of Schizophrenia, Schizoaffective Disorder, Bipolar Disorder and Unipolar Depression: Interrater Reliability and Congruence Between DSM-IV and ICD-10*, in "Psychopathology", 42(2009), pp. 296-298, in part. 293; M. Chmielewski *et al.*, *Method Matters:*

Understanding Diagnostic Reliability in DSM-IV and DSM-5, in "Journal of Abnormal Psychology", 124(2015), pp. 764, 768-769.

[49] A. Aboraya *et al.*, *The Reliability of Psychiatric Diagnosis Revisited*, cit., p. 47.

[50] *Ibid.*

[51] Vedi M. Chmielewski *et al.*, *Method Matters*, cit.

[52] Vedi, per esempio, H. Chmura Kraemer *et al.*, *DSM-5: How Reliable Is Reliable Enough?*, in "American Journal of Psychiatry", 169(2012), pp. 13-15.

[53] S.M. Lieblich *et al.*, *High Heterogeneity*, cit.

[54] Ivi, e-5.

[55] *Ibid.*

[56] Ivi, e-6.

[57] A. Aboraya *et al.*, *The Reliability of Psychiatric Diagnosis Revisited*, cit., p. 47.

[58] *Ibid.*

[59] *Ibid.*

[60] Alcuni preziosi ammonimenti si possono trovare in C. Worsham, A.B. Jena, *The Art of Evidence-Based Medicine*, in "Harvard Business Review", 30 gennaio 2019, [https://bit.ly/36CD7LO].

23

Definire la scala nelle valutazioni delle prestazioni

Partiamo da un esercizio. Prendete tre persone di vostra conoscenza, magari amici o colleghi, e assegnate loro una valutazione su una scala da 1 a 5, dove 1 è il punteggio più basso e 5 il più alto, sulla base di tre caratteristiche: gentilezza, intelligenza e diligenza. Ora chiedete a qualcuno che le conosce bene – vostra moglie o vostro marito, il vostro migliore amico o un collega fidato – di effettuare la stessa valutazione delle stesse tre persone.

Ci sono buone probabilità che, su alcuni parametri, voi e l'altro valutatore abbiate indicato numeri diversi. Ora potreste confrontarvi sulle ragioni alla base di queste differenze. Forse il motivo risiederà in un diverso impiego della scala, che abbiamo definito rumore di livello: magari voi pensavate che un 5 richiedesse doti straordinarie, mentre l'altro valutatore pensava che dovessero solo essere decisamente buone. Oppure la differenza deriva da un diverso modo di considerare le persone oggetto di valutazione: il fatto di ritenerle gentili o meno, e la vostra definizione della gentilezza stessa, potrebbero far variare il vostro giudizio da quello dell'altro valutatore.

Ora immaginate che per queste tre persone da voi valutate sia in ballo una promozione o un bonus, e che voi e l'altro giudice siate coinvolti in una valutazione delle loro prestazioni presso una società che apprezza la gentilezza (o la collegialità), l'intelligenza e la diligenza. Le vostre valutazioni sarebbero diverse? Sarebbero altrettanto alte o ancora di più? Indipendentemente dalla risposta a queste domande, è probabile che le

differenze nei criteri o nelle scale di valutazione produrranno rumore, ed è proprio ciò che si osserva ampiamente all'interno delle organizzazioni.

Un compito di giudizio

In quasi tutte le grandi organizzazioni, le prestazioni vengono regolarmente valutate in maniera formale, e chi finisce sotto esame non ne è entusiasta. Come titolava un quotidiano qualche anno fa, *Uno studio annuncia che in pratica tutti odiano le valutazioni delle prestazioni.*[1] E tutti sanno anche, secondo noi, che tali valutazioni sono soggette sia a bias sia a rumore, ma quasi nessuno crede che il rumore tocchi anche i propri giudizi.

In un mondo ideale la valutazione delle prestazioni non dovrebbe essere un compito di giudizio: basterebbero dei fatti oggettivi per stabilire se le persone stanno agendo bene. Ma quasi tutte le organizzazioni moderne hanno poco in comune con la fabbrica di spilli di Adam Smith, in cui ogni operaio aveva un rendimento misurabile. Come potrebbe essere misurabile il rendimento di un direttore finanziario o del responsabile della divisione ricerca e sviluppo di un'azienda? I lavoratori della conoscenza di oggi devono trovare un equilibrio tra molteplici obiettivi, talvolta contraddittori: concentrarsi su uno solo potrebbe portare a valutazioni errate, con dannosi effetti di incentivazione. Il numero di pazienti visitati da un medico ogni giorno è un motore importante della produttività ospedaliera, per esempio, ma sarebbe meglio che i medici non si concentrassero unicamente su quell'indicatore, né tantomeno che venissero valutati e premiati solo su tale base. Anche le misure delle prestazioni quantificabili, per esempio le vendite per un commesso o il numero di stringhe di codice per un programmatore, devono essere valutate all'interno del proprio contesto: non tutti i clienti sono ugualmente difficili da servire, come non tutti i progetti di sviluppo software sono identici. Di fronte a impegni di questo tipo, molte persone

non possono essere valutate esclusivamente sulla base di misure delle prestazioni oggettive. Da qui la pervasività delle valutazioni basate sui giudizi.[2]

Un quarto segnale, tre quarti rumore

Sulla pratica della valutazione delle prestazioni sono stati pubblicati migliaia di studi, e la gran parte dei ricercatori ritiene che sia incredibilmente soggetta a rumore.[3] Questa preoccupante conclusione proviene per lo più da studi basati su valutazioni a trecentosessanta gradi, in cui più valutatori forniscono un input sulla stessa persona, valutandone la performance su più dimensioni. Quando viene condotto questo tipo di analisi, il risultato non è piacevole: spesso gli studi indicano che la varianza reale, cioè quella attribuibile alle prestazioni del soggetto, ammonta ad appena il 20 o il 30% della varianza totale. Il resto, cioè il 70 o l'80%, è rumore sistemico.[4]

Da dove viene questo rumore? Grazie a più studi condotti sulla varianza nelle valutazioni delle prestazioni professionali,[5] sappiamo che sono presenti tutte le componenti del rumore sistemico, piuttosto facili da immaginare in questo contesto.

Consideriamo due valutatrici, Lynn e Mary. Se Lynn è indulgente e Mary è rigida, nel senso che Lynn dà valutazioni più alte di Mary, in media, a tutti i soggetti valutati, siamo in presenza di rumore di livello; come osservato a proposito dei giudici, tale rumore potrebbe indicare o che Lynn e Mary hanno impressioni davvero diverse, o che le due valutatrici semplicemente usano in maniera diversa le scale di valutazione per esprimere una stessa impressione.

Ora, se Lynn sta valutando una certa persona e si dà il caso che abbia una bassa opinione di lei e del suo lavoro, la sua generale indulgenza potrebbe

essere controbilanciata dalla sua reazione idiosincratica negativa nei confronti del soggetto in questione; è quello che viene definito uno schema stabile: la reazione di uno specifico valutatore a una specifica persona. Poiché questo schema è un tratto peculiare di Lynn (e del suo giudizio nei confronti del soggetto), è fonte di rumore strutturale.

Infine, appena prima di compilare un modulo di valutazione, Mary potrebbe aver scoperto che qualcuno le ha ammaccato la macchina nel parcheggio della società, o Lynn potrebbe avere appena ricevuto un bonus sorprendentemente lauto, che l'ha messa di buonumore – fatto per lei insolito – mentre valutava le prestazioni di un soggetto. Questi eventi, naturalmente, potrebbero produrre rumore occasionale.

Diversi studi giungono a diverse conclusioni sulla scomposizione del rumore sistemico nelle sue tre componenti (rumore di livello, strutturale e occasionale), e possiamo sicuramente immaginare i motivi per cui esso varierà da un'organizzazione all'altra. Ma il rumore andrebbe evitato in qualsiasi forma si manifesti. Il succo di queste ricerche è semplicissimo: la maggior parte delle valutazioni delle prestazioni ha meno a che fare con le effettive performance dei soggetti valutati di quanto vorremmo. Per usare le parole di un ricercatore: «Il rapporto tra prestazioni e valutazioni professionali è tendenzialmente debole, o quantomeno incerto».[6]

Inoltre vi sono molte ragioni per cui le valutazioni all'interno delle organizzazioni potrebbero non riflettere la percezione delle reali prestazioni di un dipendente da parte del valutatore:[7] per esempio, i valutatori potrebbero non cercare neanche di valutare le prestazioni in maniera attenta, limitandosi a formulare giudizi in termini "strategici".[8] Oppure potrebbero volutamente gonfiare una valutazione per evitare un difficile confronto successivo, per favorire una persona che attende da tempo una promozione o anche, paradossalmente, per liberarsi di un

membro poco produttivo di un gruppo a cui serva una buona valutazione per potersi trasferire in un'altra divisione.

Calcoli strategici del genere senz'altro influiscono sulle valutazioni, ma non sono l'unica fonte di rumore. Questo ci è noto grazie a una sorta di "esperimento naturale": alcuni sistemi di feedback a trecentosessanta gradi vengono impiegati esclusivamente per finalità di sviluppo del personale, e viene detto ai rispondenti che il feedback non verrà usato per scopi valutativi. Posto che credano davvero a questa indicazione, tale approccio scoraggia i valutatori dal gonfiare – o sgonfiare – i punteggi che assegnano. In effetti, le valutazioni finalizzate allo sviluppo fanno la differenza nella qualità del feedback, ma il rumore sistemico resta alto e produce comunque una percentuale più elevata della varianza totale rispetto alla prestazione della persona che è oggetto di valutazione. Anche quando si tratta esclusivamente di un feedback finalizzato allo sviluppo di un dipendente, il rumore nelle valutazioni non cambia.[9]

Un problema noto da tempo ma ancora irrisolto

Se i sistemi di valutazione delle prestazioni sono così inefficaci, coloro che li utilizzano dovrebbero prenderne atto e migliorarli. In effetti, negli ultimi decenni le organizzazioni hanno sperimentato innumerevoli tentativi di riforma di tali sistemi, sulla base di alcune delle strategie di riduzione del rumore qui presentate. A nostro avviso, tuttavia, si potrebbe fare molto di più.

Quasi tutte le organizzazioni impiegano la strategia di riduzione del rumore che abbiamo definito *aggregazione*. Le valutazioni aggregate sono spesso associate ai sistemi di valutazione a trecentosessanta gradi, che negli anni novanta divennero lo standard nelle grandi aziende. (La rivista

“Human Resources Management” pubblicò uno speciale su questo tipo di feedback già nel 1993.)

Le valutazioni medie di diversi giudici dovrebbero contribuire alla riduzione del rumore sistemico, ma vale la pena notare che i sistemi di feedback a trecentosessanta gradi non sono stati creati come rimedio a tale problema: il loro scopo primario consiste nel misurare molto più di ciò che un capo può facilmente notare da sé. Quando ai vostri pari e subordinati, non solo al vostro capo, viene chiesto di contribuire alla valutazione delle vostre prestazioni, l’oggetto della valutazione cambia. La ratio è che questo cambiamento sia auspicabile, poiché oggi i professionisti non devono limitarsi a compiacere il proprio datore di lavoro; la grande popolarità del feedback a trecentosessanta gradi, non a caso, ha coinciso con la generalizzazione delle organizzazioni fluide basate su progetti.

Alcuni dati indicano che il feedback a trecentosessanta gradi sia uno strumento utile in quanto prevede le prestazioni misurabili in maniera oggettiva.[10] Purtroppo, l’uso di questo sistema ha creato a sua volta dei problemi: dal momento che l’informatizzazione ha facilitato l’aggiunta di domande nei sistemi di feedback, e la proliferazione degli obiettivi e dei vincoli aziendali ha aggiunto nuove dimensioni ai mansionari, molti questionari sono diventati incredibilmente complessi e iperspecifici (per esempio, un questionario prevede quarantasei valutazioni su undici dimensioni per ogni valutatore e soggetto valutato).[11] Ci vorrebbe un valutatore con capacità sovrumane per ricordare ed elaborare fatti accurati e pertinenti relativi a numerosi soggetti su così tante dimensioni. In un certo senso, questo approccio eccessivamente complicato è non solo inutile, ma anche dannoso: come abbiamo visto, l’effetto alone implica che dimensioni teoricamente separate non vengano trattate come tali, e una valutazione fortemente positiva o negativa in uno dei primi quesiti tenderà a trainare nella stessa direzione le risposte alle domande successive.

Ancora più importante è poi il fatto che lo sviluppo dei sistemi di feedback a trecentosessanta gradi ha aumentato esponenzialmente i tempi dedicati alla produzione del feedback stesso. Non è insolito che a un dirigente intermedio venga chiesto di completare decine di questionari sui propri colleghi di ogni grado e livello, e talvolta anche su chi ricopre un ruolo analogo in altre organizzazioni, perché oggi molte società richiedono feedback da clienti, venditori e altri soci aziendali. Pur con le migliori intenzioni, questa esplosione di richieste poste a valutatori con poco tempo a disposizione difficilmente migliorerà la qualità delle informazioni da loro fornite. In questo caso, tentare di ridurre il rumore potrebbe perfino non valere la pena, in quanto l'operazione avrebbe un costo troppo alto – problema, questo, di cui parleremo nella parte 6.

Infine, i sistemi di feedback a trecentosessanta gradi non sono immuni dai difetti praticamente universali di tutti i sistemi di misurazione delle prestazioni: la tendenza strisciante a gonfiare le valutazioni. Una grande società industriale una volta osservò che il 98% dei propri dirigenti risultava «soddisfare pienamente le aspettative».[12] Quando quasi tutti ricevono il voto più alto possibile, è lecito dubitare del valore delle valutazioni.

Elogio dei giudizi relativi

Una soluzione teoricamente efficace al problema delle valutazioni gonfiate consiste nell'introduzione di una forma di standardizzazione, e una prassi comune impiegata a questo scopo è la *distribuzione forzata*, un sistema in cui i valutatori non solo vengono privati della facoltà di dare a tutti la più alta valutazione possibile, ma sono anche costretti ad attenersi a una distribuzione predeterminata.[13] Questo strumento venne proposto da Jack Welch quando ricopriva l'incarico di amministratore delegato di General Electric per bloccare le valutazioni gonfiate e introdurre il massimo

"candore" nelle valutazioni delle prestazioni, e da allora lo adottarono molte società, che poi dovettero però abbandonarlo per gli effetti collaterali indesiderati che aveva sul morale e sulla cooperazione del personale.

A prescindere dalle loro pecche, le distribuzioni in ranghi sono meno soggette a rumore delle valutazioni. Nell'esempio dei danni punitivi abbiamo visto che vi è molto meno rumore nei giudizi relativi che in quelli assoluti, e lo stesso sembra valere anche per le valutazioni delle prestazioni.[14]

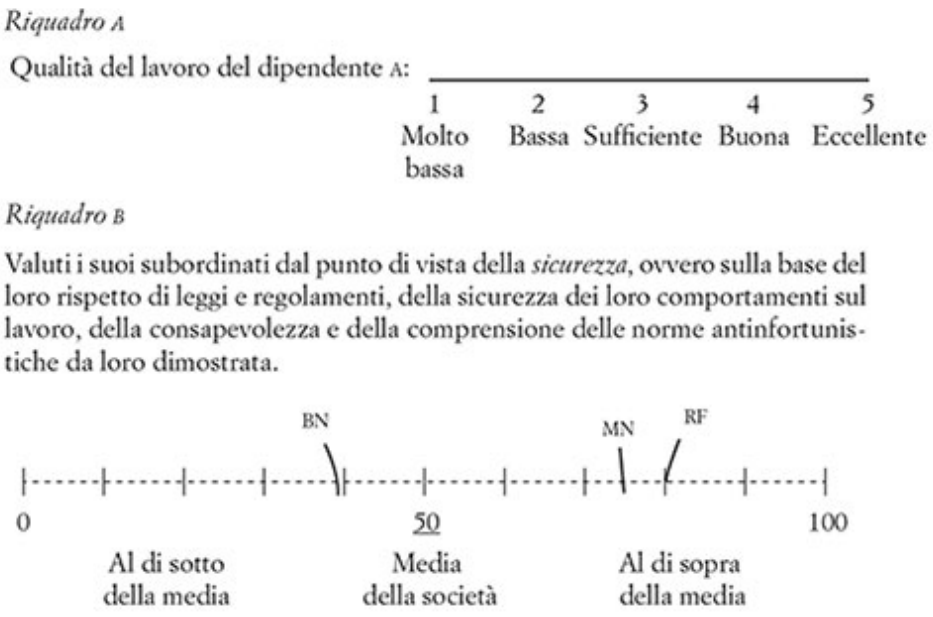


Figura 17. Esempi di scale di valutazione assolute e relative[15]

Per capire perché, osservate la figura 17, che mostra due esempi di scale per la valutazione dei dipendenti. Nel riquadro A, in cui un impiegato viene valutato su una scala assoluta, è richiesta quella che abbiamo definito un'operazione di matching: individuare il punteggio che si avvicina di più alla vostra impressione della "qualità del lavoro" del dipendente. Per contro, il riquadro B richiede che ciascun individuo venga confrontato con un gruppo di altri rispetto a una dimensione specifica: la sicurezza. Al supervisore viene chiesto di stabilire il rango o percentile di un dipendente

all'interno di una popolazione specifica, attraverso una scala percentile. Come vediamo, il supervisore ha posizionato tre dipendenti su questa semplice scala.

L'approccio adottato nel riquadro B ha due vantaggi. Innanzitutto, la valutazione di tutti i dipendenti su una dimensione alla volta (in questo caso, la sicurezza) offre un esempio di una strategia di riduzione del rumore di cui parleremo più nel dettaglio nel prossimo capitolo: la *strutturazione* di un giudizio complesso in più dimensioni. La strutturazione rappresenta un tentativo di limitare l'effetto alone, che di solito concentra in un intervallo ristretto le valutazioni di un individuo su diverse dimensioni. (Naturalmente, funziona solo se la distribuzione viene effettuata separatamente su ogni dimensione, come in questo esempio: classificare i dipendenti sulla base di un vago giudizio aggregato di "qualità del lavoro" non ridurrebbe l'effetto alone.)

In secondo luogo, come si è detto nel capitolo 15, distribuire le valutazioni in ranghi riduce sia il rumore strutturale sia quello di livello: sarete tendenzialmente meno incoerenti (e creerete meno rumore strutturale) se paragonate le prestazioni di due membri del vostro gruppo piuttosto che dare a ciascuno un voto indipendente. Cosa ancora più importante, i ranghi eliminano meccanicamente il rumore di livello: se Lynn e Mary valutano lo stesso gruppo di venti dipendenti e Lynn è più indulgente di Mary, le loro valutazioni medie saranno diverse, ma le loro distribuzioni medie no. Che siano indulgenti o rigide nelle loro distribuzioni, impiegheranno gli stessi ranghi.

In effetti, la riduzione del rumore è il principale obiettivo dichiarato della distribuzione forzata, che garantisce la stessa media e la stessa distribuzione delle valutazioni da parte di tutti i valutatori. Viene definita "forzata" proprio perché impone una certa distribuzione dei giudizi: per esempio, una regola potrebbe stabilire che non più del 20% dei soggetti valutati possa

ricadere nella categoria più alta, e che non meno del 15% venga inserito in quella più bassa.

Distribuire, ma senza forzare

In linea di principio, quindi, la distribuzione forzata dovrebbe condurre ai miglioramenti necessari, eppure spesso è un'arma a doppio taglio. Qui non intendiamo elencare tutti i possibili effetti indesiderati (spesso legati a un'applicazione scorretta, non al principio di fondo), ma due aspetti dei sistemi di distribuzione forzata consentono di avanzare delle riflessioni generali.

Il primo è la confusione tra prestazioni assolute e relative. Certo è impossibile per il 98% dei dirigenti di una società collocarsi nella fascia del 20%, del 50% o anche dell'80% dei migliori del proprio gruppo, ma non è assurdo che tutti «soddisfino le aspettative», se queste sono state definite a priori *in termini assoluti.*

Molti dirigenti si oppongono all'idea che quasi tutti i dipendenti siano in grado di soddisfare le aspettative. Se fosse così, sostengono, vorrebbe dire che ci si aspetta da loro troppo poco, forse per via di una sorta di autocompiacimento collettivo. Bisogna ammettere che questa interpretazione potrebbe anche essere valida, ma è comunque possibile che la maggior parte dei dipendenti soddisfino aspettative *elevate*; anzi, è proprio quanto ci aspetteremmo in un'organizzazione ad alte prestazioni. Non taccereste di eccessiva indulgenza le procedure di gestione delle prestazioni della NASA, se veniste a sapere che tutti gli astronauti di una certa missione spaziale di successo hanno pienamente soddisfatto le aspettative.

Ne consegue che un sistema che dipende da valutazioni relative è appropriato solo se a un'organizzazione interessano le prestazioni relative.

Per esempio, potrebbe avere senso quando, a prescindere dalle prestazioni assolute di un gruppo, solo una percentuale fissa di quest'ultimo può ricevere una promozione (pensiamo ai colonnelli che vengono valutati per la promozione a generali). Ma forzare una distribuzione relativa laddove ci si proponga di misurare un livello *assoluto* di prestazioni, come fanno molte società, non ha senso; e richiedere che una percentuale predefinita di dipendenti venga giudicata non in grado di soddisfare le aspettative (assolute) non è soltanto crudele, è assurdo. Sarebbe sciocco dire che il 10% di un corpo scelto dell'esercito debba essere classificato come "insoddisfacente".

Il secondo problema è che la distribuzione forzata delle valutazioni dovrebbe riflettere quella delle prestazioni reali corrispondenti, che in genere è simile a una distribuzione normale. Eppure, anche se la distribuzione delle prestazioni nella popolazione soggetta a valutazione fosse nota, la stessa potrebbe non riscontrarsi in un gruppo più piccolo, come la porzione di soggetti giudicati da un singolo valutatore. Un esempio chiarirà il concetto: se estraete a caso dieci persone da una popolazione di diverse migliaia, non vi è alcuna garanzia che esattamente due di loro appartengano al 20% dei migliori della popolazione generale. («Non vi è alcuna garanzia» in realtà è un eufemismo: la probabilità che ciò avvenga non supera il 30%.) Nella pratica, il problema è ancora peggiore, perché la composizione dei gruppi non è casuale: alcune unità potranno vantare una quasi totalità di personale ad alte prestazioni, mentre altre saranno composte da dipendenti al di sotto della media.

Inevitabilmente, in un simile scenario la distribuzione forzata è una fonte di errore e iniquità. Poniamo che il gruppo di un valutatore si componga di cinque persone le cui prestazioni siano indistinguibili: forzare una distribuzione differenziata delle valutazioni in questa realtà indifferenziata non riduce l'errore, ma lo aumenta.

I critici della distribuzione forzata spesso attaccano il principio della classificazione, che condannano in quanto brutale, disumano e in fin dei conti controproducente. Che accettiate o meno queste argomentazioni, la pecca fatale di questo tipo di distribuzione non sta nella classificazione, ma nella "forzatura": ogni volta che dei giudizi vengono "forzati" in una scala inadeguata, perché si impiega una scala relativa per misurare una prestazione assoluta o perché i valutatori sono costretti a operare distinzioni nell'indistinguibile, la scelta della scala aggiungerà automaticamente rumore.

Cosa ci aspetta?

Alla luce di tutti gli sforzi compiuti dalle organizzazioni per migliorare la misurazione delle prestazioni, i risultati possono dirsi a dir poco deludenti. Proprio in virtù di questi sforzi, il costo delle valutazioni delle prestazioni è balzato alle stelle: nel 2015 Deloitte ha calcolato di aver dedicato due milioni di ore all'anno alla valutazione dei suoi sessantacinquemila dipendenti.[16] Il rituale della valutazione continua a essere uno dei più temuti all'interno delle organizzazioni, odiato quasi allo stesso livello da chi lo effettua e da chi lo subisce. Da uno studio è emerso che addirittura il 90% di dirigenti, dipendenti e responsabili delle risorse umane ritengono che i propri processi di gestione delle prestazioni non abbiano dato i risultati attesi.[17] Le ricerche hanno confermato ciò che riscontra la maggior parte dei dirigenti: benché il feedback sulle prestazioni, quando è associato a un piano di sviluppo dei dipendenti, possa condurre a dei miglioramenti, le valutazioni della performance, per come spesso vengono svolte, possono motivare ma anche demotivare. Come ben sintetizza un articolo sul tema, «per quanto da decenni si cerchi di migliorare i processi [di gestione delle prestazioni],

questi continuano a generare informazioni inaccurate e non contribuiscono in nessun modo a stimolare le prestazioni».[18]

Per disperazione, oggi un numero esiguo ma crescente di società sta pensando di operare una scelta radicale, eliminando del tutto i sistemi di valutazione. I fautori di questa «rivoluzione della gestione delle prestazioni»,[19] che comprendono molte imprese tecnologiche, alcune società di servizi professionali e una manciata di aziende dei settori più tradizionali, intendono concentrarsi su feedback orientati al futuro e alla crescita del dipendente più che su giudizi valutativi che guardano al passato. Alcune hanno perfino eliminato i numeri dalle proprie valutazioni, abbandonando quindi le forme più tradizionali di valutazione delle prestazioni.

Ma per le società che non intendono farne a meno (ovvero la schiacciante maggioranza), come è possibile migliorarle? Una delle strategie di riduzione del rumore si basa, ancora una volta, sulla scelta della scala giusta. L'intento è di garantire un *quadro comune di riferimento*: le ricerche indicano che associare migliori modelli di valutazione a una formazione dei valutatori può contribuire a raggiungere una maggiore coerenza nell'impiego della scala.

Come minimo, le scale di valutazione delle prestazioni devono essere ancorate a descrittori sufficientemente specifici da poter essere interpretati in maniera coerente. Molte organizzazioni adottano *scale ad ancoraggio comportamentale* in cui ogni livello corrisponde a una descrizione di comportamenti specifici. Il riquadro di sinistra della figura 18 ne fornisce un esempio.

I dati indicano, tuttavia, che le scale di questo tipo non sono sufficienti a eliminare il rumore.[20] Una fase ulteriore, la *formazione sul quadro di riferimento*, si è dimostrata utile per garantire la coerenza tra i giudizi. In questa fase i valutatori vengono formati per riconoscere le diverse dimensioni delle

prestazioni attraverso brevi filmati e poi imparano a confrontare le proprie valutazioni con quelle "veritiere" fornite dagli esperti.[21] I filmati fungono da casi di riferimento: ognuno rappresenta un punto di ancoraggio sulla scala delle prestazioni, che diviene quindi una *scala di casi*, come quella mostrata nel riquadro di destra della figura 18.

In una simile scala, ogni nuova valutazione di un individuo viene svolta mediante il confronto con un caso di ancoraggio, diventando così un giudizio relativo. Poiché i giudizi comparativi sono meno soggetti a rumore rispetto alle valutazioni, le scale di casi sono più affidabili di quelle basate su numeri, aggettivi o descrizioni comportamentali.

La formazione sul quadro di riferimento è nota da decenni, ed è appurato che riduca il rumore e dia valutazioni più accurate. Eppure non ha fatto molta strada, ed è facile intuire perché: così come le scale di casi e altri mezzi rivolti agli stessi obiettivi, è uno strumento complesso e richiede un grande dispendio in termini di tempo. Per essere validi, questi strumenti spesso devono essere adattati all'azienda e perfino all'unità che conduce le valutazioni, e vanno aggiornati spesso per seguire l'evoluzione dei requisiti professionali; inoltre, richiedono un ulteriore investimento nei sistemi di gestione delle prestazioni da parte dell'azienda. Per questo oggi ci si muove nella direzione opposta. (Nella parte 6 torneremo a soffermarci sui costi della riduzione del rumore.)

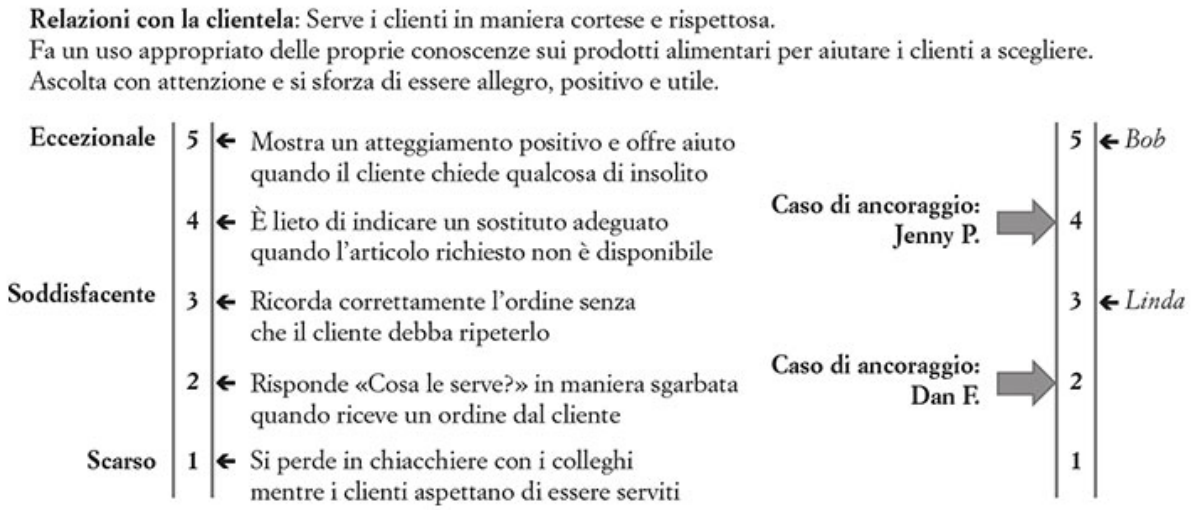


Figura 18. Esempi di una scala di valutazione ad ancoraggio comportamentale (sinistra) e di una scala di casi (destra)[22]

Per giunta, ogni organizzazione che riesca a contenere il rumore attribuibile ai valutatori ridurrà anche la loro capacità di influenzare le valutazioni per i propri fini; chiedere ai manager di svolgere una formazione supplementare come valutatori, investire di più nel processo di valutazione e rinunciare in parte al proprio controllo sui risultati non potrà che scontrarsi con una notevole resistenza. Non è un caso se finora la maggior parte degli studi sulla formazione sul quadro di riferimento dei valutatori è stata condotta sugli studenti, non sui veri manager.[23]

L'ampio tema della valutazione delle prestazioni solleva molti interrogativi, di tipo pratico e filosofico. Alcuni si chiedono, per esempio, fino a che punto il concetto di prestazioni individuali abbia senso nelle organizzazioni contemporanee, in cui spesso i risultati dipendono dall'interazione reciproca del personale. Se questa idea invece ci sembra sensata, dovremmo chiederci su quanti livelli vada a distribuirsi il personale di una data organizzazione se si considerano le prestazioni individuali – per esempio, se le prestazioni seguano una curva di distribuzione normale o se esistano "talenti stellari" che danno un contributo del tutto sproporzionato

rispetto agli altri.[24] E se il vostro obiettivo è tirare fuori il meglio dalle persone, fareste bene a chiedervi se la misurazione delle prestazioni individuali e il suo impiego per motivare il personale facendo leva sulla paura e sull'avidità sia l'approccio migliore (o anche solo se sia davvero efficace).

Se state progettando o rivedendo un sistema di gestione delle prestazioni, dovrete porvi queste e altre domande. Non ci proponiamo di esaminarle in questa sede, ma vorremmo avanzare un modesto suggerimento: se vi avvalete di una misurazione delle prestazioni, probabilmente le vostre valutazioni saranno impregnate di rumore sistemico e, per questa ragione, è possibile che siano del tutto inutili e forse anche controproducenti. Ridurre il rumore è una sfida che non può essere risolta attraverso semplici correzioni tecnologiche, ma richiede una riflessione lucida sui giudizi che ai valutatori viene chiesto di esprimere. Con ogni probabilità vi accorgerete che è possibile migliorare i giudizi facendo chiarezza sulla scala di valutazione e formando il personale affinché la utilizzi in maniera coerente. Questa strategia di riduzione del rumore è applicabile anche in molti altri campi.

A proposito della definizione della scala

«Dedichiamo molto tempo alla valutazione delle prestazioni, eppure i risultati dipendono per un quarto dalle effettive performance dei dipendenti e per tre quarti dal rumore sistemico.»

«Abbiamo provato il feedback a trecentosessanta gradi e la distribuzione forzata per risolvere il problema, ma forse abbiamo solo peggiorato le cose.»

«Se c'è così tanto rumore di livello è perché valutatori diversi hanno idee completamente diverse sul significato di parole come "buono", "ottimo" e altre. Troveranno un accordo solo se presentiamo loro dei casi concreti che fungano da ancoraggio sulla scala di valutazione.»

[1] J. McGregor, *Study Finds That Basically Every Single Person Hates Performance Reviews*, in "Washington Post", 27 gennaio 2014.

[2] La trasformazione digitale in atto in molte organizzazioni potrebbe creare nuove possibilità in questo senso. In teoria, le società oggi possono raccogliere grandi quantità di informazioni dettagliate in tempo reale sulle prestazioni di ogni professionista, che potrebbero aprire la strada a valutazioni algoritmiche delle performance per certi ruoli. Qui, tuttavia, ci concentriamo sulle posizioni per le quali il giudizio non può essere totalmente eliminato dalla misurazione delle prestazioni. Vedi E.D. Pulakos, R. Mueller-Hanson, S. Arad, *The Evolution of Performance Management: Searching for Value*, in "Annual Review of Organizational Psychology and Organizational Behavior", 6(2018), pp. 249-271.

[3] S.E. Scullen, M.K. Mount, M. Goff, *Understanding the Latent Structure of Job Performance Ratings*, in "Journal of Applied Psychology", 85(2000), pp. 956-970.

[4] Una piccola percentuale, pari al 10% della varianza totale in alcuni studi, costituisce quella che i ricercatori chiamano *prospettiva del valutatore*, o effetto "di livello", nel senso del livello all'interno dell'organizzazione, non di quello che abbiamo qui definito "rumore di livello". La prospettiva del valutatore indica che, nel valutare la stessa persona, un capo differisce sistematicamente da un pari, e un pari da un subordinato. Se si vuole dare un'interpretazione benevola dei sistemi di valutazione a trecentosessanta gradi, si potrebbe sostenere che non si tratti di rumore: se più persone a diversi livelli di un'organizzazione vedono sistematicamente aspetti diversi delle prestazioni della stessa persona, il loro giudizio su quella persona dovrebbe essere sistematicamente diverso, e ciò dovrebbe riflettersi nelle loro valutazioni.

[5] S.E. Scullen, M.K. Mount, M. Goff, *Latent Structure*, cit.; C. Viswesvaran, D.S. Ones, F.L. Schmidt, *Comparative Analysis of the Reliability of Job Performance Ratings*, in "Journal of Applied Psychology", 81(1996), pp. 557-574; G.J. Greguras, C. Robie, *A New Look at Within-Source Interrater Reliability of 360-Degree Feedback Ratings*, in "Journal of Applied Psychology", 83(1998), pp. 960-968; G.J. Greguras *et al.*, *A Field Study of the Effects of Rating Purpose on the Quality of Multisource Rating*, in "Personnel Psychology", 56(2003), pp. 1-21; C. Viswesvaran, F.L. Schmidt, D.S. Ones, *Is There a General Factor in Ratings of Job Performance? A Meta-Analytic Framework for Disentangling Substantive and Error Influences*, in "Journal of Applied Psychology", 90(2005), pp. 108-131; B. Hoffman *et al.*, *Rater Source Effects Are Alive and Well After All*, in "Personnel Psychology", 63(2010), pp. 119-151.

[6] K.R. Murphy, *Explaining the Weak Relationship Between Job Performance and Ratings of Job Performance*, in "Industrial and Organizational Psychology", 1(2008), pp. 148-160, in part. 151.

[7] Nella disamina delle fonti di rumore abbiamo tralasciato la possibilità del rumore di caso derivante da bias sistematici nella valutazione di certi dipendenti o categorie di dipendenti. Nessuno degli studi che abbiamo reperito sulla variabilità delle valutazioni delle prestazioni le ha confrontate con prestazioni "vere" valutate esternamente.

[8] E.D. Pulakos, R.S. O'Leary, *Why Is Performance Management Broken?*, in "Industrial and Organizational Psychology", 4(2011), pp. 146-164; M.M. Harris, *Rater Motivation in the Performance Appraisal Context: A Theoretical Framework*, in "Journal of Management", 20(1994), pp. 737-756; K.R. Murphy, J.N. Cleveland, *Understanding Performance Appraisal: Social, Organizational, and Goal-Based Perspectives*, Sage, Thousand Oaks, CA 1995.

[9] G.J. Greguras *et al.*, *Field Study*, cit.

[10] P.W. Atkins, R.E. Wood, *Self-Versus Others' Ratings as Predictors of Assessment Center Ratings: Validation Evidence for 360-Degree Feedback Programs*, in "Personnel Psychology", 55(2006), n. 4, pp. 871-904.

[11] *Ibid.*

[12] C.A. Olson, G.M. Davis, *Pros and Cons of Forced Ranking and other relative performance ranking systems*, citato in P.G. Dominick, *Forced Ranking: Pros, Cons and Practices*, in J.W. Smither, M. London (a cura di), *Performance Management: Putting Research into Action*, Jossey-Bass, San Francisco 2009, pp. 411-443.

[13] P.G. Dominick, *Forced Ranking*, cit.

[14] B.R. Nathan, R.A. Alexander, *A Comparison of Criteria for Test Validation: A Meta-Analytic Investigation*, in "Personnel Psychology", 41(1988), n. 3, pp. 517-535.

[15] Adattata da R.D. Goffin e J.M. Olson, *Is It All Relative? Comparative Judgments and the Possible Improvement of Self-Ratings and Ratings of Others*, in "Perspectives on Psychological Science", 6(2011), n. 1, pp. 48-60.

[16] M. Buckingham, A. Goodall, *Reinventing Performance Management*, in "Harvard Business Review", aprile 2015, pp. 1-16, doi:ISSN: 0017-8012.

[17] Corporate Leadership Council, citato in S. Adler *et al.*, *Getting Rid of Performance Ratings: Genius or Folly? A Debate*, in "Industrial and Organizational Psychology", 9(2016), pp. 219-252.

[18] E.D. Pulakos, R. Mueller-Hanson, S. Arad, *The Evolution of Performance Management*, cit., p. 250.

[19] A. Tavis, P. Cappelli, *The Performance Management Revolution*, in "Harvard Business Review", ottobre 2016, pp. 1-17.

[20] F.J. Landy, J.L. Farr, *Performance Rating*, in "Psychological Bulletin", 87(1980), n. 1, pp. 72-107.

[21] D.J. Woehr, A.I. Huffcutt, *Rater Training for Performance Appraisal: A Quantitative Review*, in "Journal of Occupational and Organizational Psychology", 67(1994), pp. 189-205; S.G. Roch *et al.*, *Rater Training Revisited: An Updated Meta-Analytic Review of Frame-of-Reference Training*, in "Journal of Occupational and Organizational Psychology", 85(2012), pp. 370-395; M.H. Tsai, S. Wee, B. Koh,

Restructured Frame-of-Reference Training Improves Rating Accuracy, in "Journal of Organizational Behavior" (2019), pp. 1-18, doi:10.1002/job.2368.

[22] Il riquadro di sinistra è adattato da R.D. Goffin, J.M. Olson, *Is It All Relative? Comparative Judgments and the Possible Improvement of Self-Ratings and Ratings of Others*, in "Perspectives on Psychological Science", 6(2011), n. 1, pp. 48-60.

[23] S.G. Roch *et al.*, *Rater Training Revisited*, cit.

[24] E. O'Boyle, H. Aguinis, *The Best and the Rest: Revisiting the Norm of Normality of Individual Performance*, in "Personnel Psychology", 65(2012), n. 1, pp. 79-119; H. Aguinis, E. O'Boyle, *Star Performers in Twenty-First Century Organizations*, in "Personnel Psychology", 67(2014), n. 2, pp. 313-350.

24

Strutturare le assunzioni

Chiunque abbia ricoperto un incarico di qualsiasi tipo ricorderà lo stress del colloquio di assunzione. I colloqui di lavoro, in cui il candidato incontra il proprio futuro supervisore o un professionista delle risorse umane, sono un rito di passaggio imprescindibile per accedere a molte organizzazioni.

Nella maggior parte dei casi si segue una routine consolidata: dopo uno scambio di convenevoli i reclutatori chiedono ai candidati di descrivere la propria esperienza o di dilungarsi su un aspetto specifico della stessa; vengono poste domande su sfide e traguardi raggiunti, sulle motivazioni per intraprendere quel lavoro o su idee per migliorare la società in cui andranno a inserirsi; spesso i reclutatori chiedono ai candidati di descrivere la propria personalità e di spiegare perché sarebbero adatti a una certa posizione o alla loro cultura aziendale, e talvolta si parla di hobby e interessi personali. Verso la fine, si chiede al candidato di porre a sua volta qualche domanda, che spesso viene valutata in base alla sua pertinenza e al suo acume.

Se ora vi trovate nella posizione di assumere dei dipendenti, è probabile che i vostri metodi di selezione prevedano una qualche versione di questo rituale. Come ha osservato uno psicologo del lavoro: «È raro, e forse addirittura impensabile, che qualcuno venga assunto senza che vi sia prima un colloquio di qualche tipo».[1] E quasi tutti i professionisti si affidano in una certa misura ai propri giudizi intuitivi nel prendere decisioni sulle assunzioni durante questi incontri.[2]

L'altissima diffusione dei colloqui di lavoro riflette una profonda fiducia nel valore del giudizio quando si tratta di scegliere le persone con cui andremo a lavorare. Come compito di giudizio, la selezione del personale ha un grosso vantaggio: essendo così diffusa e importante, è stata studiata nel dettaglio dagli psicologi del lavoro. Il primo numero del "Journal of Applied Psychology" del 1917 identificava l'assunzione come «il problema supremo [...] perché le capacità umane, dopotutto, sono la principale risorsa di una nazione».[3] Oggi, a distanza di un secolo, abbiamo molte informazioni sull'efficacia di varie tecniche di selezione (compresi i colloqui classici): non vi è altro compito di giudizio complesso su cui siano state svolte altrettante ricerche sul campo. Per questo rappresenta un perfetto banco di prova da cui poter trarre conclusioni estendibili a molti giudizi che presuppongono una scelta tra più opzioni disponibili.

I rischi dei colloqui

Se non siete al corrente delle ricerche sui colloqui di lavoro, resterete sorpresi da quanto segue. In sostanza, se il vostro obiettivo è stabilire quali candidati avranno successo in una certa professione e quali no, i colloqui classici (chiamati anche "interviste non strutturate", per distinguerli da quelle strutturate, su cui presto torneremo) non vi saranno molto d'aiuto. Anzi, per dirlo senza mezzi termini, spesso sono inutili.

Questa conclusione è frutto di innumerevoli studi che hanno stimato la correlazione tra la valutazione data a un candidato dopo un colloquio e il successo dello stesso in quel ruolo. Se la correlazione è alta, i colloqui – o qualsiasi altra tecnica di reclutamento la cui correlazione si calcoli nello stesso modo – possono essere definiti un buon predittore delle prestazioni future dei candidati.

Qui occorre fare una precisazione. La definizione di *successo* è un problema non da poco. In genere le prestazioni vengono misurate sulla base delle valutazioni di un supervisore, mentre altre volte si valuta l'anzianità di servizio. Tali misurazioni, naturalmente, sollevano dei dubbi, soprattutto se si considera la discutibile validità delle valutazioni delle prestazioni osservata nel precedente capitolo. Tuttavia, allo scopo di valutare la qualità dei giudizi di un datore di lavoro al momento della selezione dei dipendenti, sembra ragionevole prendere in considerazione i giudizi espressi dallo stesso nel valutare in un secondo momento i dipendenti da lui assunti. Qualsiasi analisi della qualità delle decisioni di assunzione dovrà partire da tale presupposto.

A quale conclusione sono giunte queste analisi? Nel capitolo 11 abbiamo accennato a una correlazione di 0,28 tra le valutazioni dei colloqui classici e quelle delle prestazioni lavorative; altri studi riportano correlazioni che variano da 0,20 a 0,33.[4] Come abbiamo visto, questa è un'ottima correlazione per gli standard delle scienze sociali, ma non abbastanza elevata per usarla come base per le proprie decisioni. Impiegando la percentuale di coppie concordanti (PC) presentata nella parte 3, possiamo calcolare la probabilità: dati i livelli di correlazione appena menzionati, se l'unica cosa che sapete di due candidati è che uno vi ha fatto un'impressione migliore dell'altro in sede di colloquio, le possibilità che tale candidato dimostri di avere prestazioni migliori si collocano tra il 56 e il 61%. Un po' meglio che lanciare una moneta, certo, ma non proprio un metodo a prova di errore quando si tratta di prendere decisioni importanti.

A dire il vero, i colloqui non servono solo a formulare giudizi sui candidati: costituiscono anche un'occasione per "vendere" la società a profili promettenti e per cominciare a stabilire un primo contatto con i futuri colleghi. Eppure, nella logica di un'organizzazione che investe

tempo e risorse nella selezione dei talenti, la loro prima finalità è appunto la selezione. E, in questo senso, i colloqui non si possono definire un grande successo.

Rumore nei colloqui

È facile vedere per quale motivo i colloqui tradizionali portino a errori nella previsione delle prestazioni professionali. Alcuni di questi errori hanno a che fare con quella che abbiamo chiamato ignoranza oggettiva (vedi capitolo 11).[5] Le prestazioni lavorative dipendono da tanti elementi, compresa la velocità con cui il neoassunto si adatta alla nuova posizione o l'influenza di varie circostanze della vita sul suo lavoro; molti di questi fattori non si possono prevedere al momento dell'assunzione, e questa incertezza limita la validità predittiva dei colloqui e, in generale, di ogni tecnica di selezione del personale.

I colloqui sono inoltre un campo minato di bias psicologici. In anni recenti si è constatato che i reclutatori tendono, spesso senza volerlo, a favorire i candidati culturalmente simili a loro o con cui hanno qualcosa in comune, compresi genere, razza o istruzione.[6] Molte società oggi riconoscono i rischi posti dai bias e cercano di affrontarli offrendo una formazione specifica a reclutatori professionisti e altri dipendenti. Da decenni si conoscono anche altri bias: l'aspetto fisico, per esempio, ha un grande peso nella valutazione dei candidati, anche per posizioni in cui dovrebbe avere poca o nessuna importanza. Tali bias accomunano quasi tutti i reclutatori e, quando applicati a una data categoria, tenderanno a produrre un errore condiviso, ovvero un bias positivo o negativo nella valutazione del candidato.

Come a questo punto potrete intuire, nei colloqui subentra anche il rumore: selezionatori diversi reagiranno in maniera diversa allo stesso candidato e perverranno a diverse conclusioni. Le misure della correlazione tra le valutazioni di due reclutatori dopo un colloquio con lo stesso candidato variano tra 0,37 e 0,44 (PC = 62-65%).[7] Una spiegazione possibile è che il candidato potrebbe non comportarsi esattamente nello steso modo con reclutatori diversi, ma anche nelle cosiddette *panel interviews*, in cui più selezionatori assistono allo stesso comportamento del candidato, la correlazione tra le loro valutazioni non è affatto perfetta. Una meta-analisi stima una correlazione di 0,74 (PC = 76%). Ciò significa che, dopo aver visto gli stessi due candidati nella stessa *panel interview*, in un caso su quattro due selezionatori non saranno d'accordo su chi è il migliore tra i due.

Questa variabilità deriva più che altro dal rumore strutturale, la differenza nelle reazioni idiosincratiche dei reclutatori a un dato candidato. Quasi tutte le organizzazioni si aspettano una simile variabilità e, per questo motivo, richiedono più colloqui con lo stesso candidato, aggregando poi i risultati in un modo o nell'altro. (In genere l'opinione aggregata si forma attraverso una discussione in cui occorre pervenire a un consenso – una procedura di per sé problematica, come abbiamo visto).

Un dato forse più sorprendente è la presenza di molto rumore occasionale nei colloqui. Vi sono evidenze schiaccianti, per esempio, che i pareri sui candidati siano in rapporto con le impressioni che ci si forma nella fase informale del colloquio in cui si stabilisce un primo contatto, i primi due o tre minuti in cui si scambiano quattro chiacchiere per mettere il candidato a proprio agio. Questa prima impressione conta, e parecchio.[8]

Forse penserete che giudicare sulla base della prima impressione non sia un problema, e in effetti almeno una parte di ciò che apprendiamo in

maniera immediata è significativo: sappiamo tutti che già nei primi secondi di interazione con una persona appena conosciuta si riesce a capire qualcosa di lei. Non si può negare che questo valga in modo particolare per i selezionatori più abili. Ma i primi secondi di un colloquio riflettono proprio le qualità superficiali che associamo alle prime impressioni: le percezioni iniziali si basano per lo più sull'estroversione e sulle capacità verbali del candidato. Perfino la stretta di mano è un importante predittore dei pareri che verranno espressi sui candidati![9] Forse tutti noi apprezziamo una stretta di mano decisa, ma pochi reclutatori la sceglierebbero in maniera consapevole come criterio chiave per l'assunzione.

La psicologia dei selezionatori

Perché la prima impressione finisce per segnare il corso di un colloquio molto più lungo? Un motivo è che in un colloquio tradizionale i selezionatori sono liberi di portare l'intervista nella direzione che ritengono più appropriata, e probabilmente porranno delle domande che confermano la loro impressione iniziale. Se un candidato appare timido e riservato, per esempio, il selezionatore vorrà fargli qualche domanda sulle sue esperienze all'interno di un gruppo di lavoro, ma forse non riterrà fondamentale porre le stesse domande a qualcuno che si presenta come sorridente e socievole, con il risultato che i dati raccolti sui due candidati non saranno gli stessi. Uno studio basato sul monitoraggio del comportamento di alcuni selezionatori che si erano formati una prima impressione positiva o negativa a partire dai curricula e dai risultati dei test condotti dai candidati ha riscontrato che quell'impressione iniziale ha un grande effetto sull'andamento del colloquio. I selezionatori che

ricevono fin da subito una sensazione positiva, per esempio, pongono meno domande e tendono a "vendere" la società al candidato.[10]

Ma il potere della prima impressione non è l'unico aspetto problematico dei colloqui: vi è anche il fatto che i selezionatori vogliono che il candidato seduto di fronte a loro *abbia senso* (è un'altra manifestazione della nostra tendenza eccessiva a ricercare la coerenza dappertutto, discussa nel capitolo 13). In un interessante esperimento, i ricercatori hanno assegnato ad alcuni studenti il ruolo di candidati o di selezionatori, dicendo a entrambi che il colloquio doveva consistere esclusivamente in domande chiuse a cui rispondere con un sì o con un no.[11] Poi hanno chiesto ad alcuni candidati di rispondere in maniera *casuale*. (La prima lettera della prima parola con cui veniva formulata una domanda determinava la risposta affermativa o negativa.) Come osservano ironicamente i ricercatori: «Alcuni candidati inizialmente temevano che quei colloqui casuali sarebbero presto stati smascherati perché non aveva alcun senso. Ma questo problema non c'è stato e i colloqui si sono svolti normalmente». Sì, avete letto bene: *nessun selezionatore* si è reso conto che i candidati stavano dando risposte del tutto casuali. Peggio ancora, quando è stato chiesto loro di ipotizzare se fossero «in grado di dedurre molte informazioni su questa persona in virtù del tempo trascorso con lei», i selezionatori di questo studio randomizzato tendevano a rispondere positivamente non meno di chi aveva sostenuto un colloquio con candidati che avevano dato risposte veritiere. Alla base c'è la nostra grande capacità di creare delle storie coerenti. Come spesso troviamo degli schemi immaginari in dati casuali o vediamo una forma nei contorni di una nuvola, siamo capaci di trovare una logica in risposte del tutto insensate.

Per un'altra illustrazione meno estrema, considerate il seguente caso. Uno degli autori di questo libro ha dovuto gestire un colloquio con un

candidato che, in precedenza, era stato direttore finanziario di una società di medie dimensioni. Si è accorto che il candidato aveva lasciato quella posizione dopo qualche mese e gliene ha chiesto il motivo, e quest'ultimo gli ha spiegato che si era trattato di «una divergenza strategica con l'amministratore delegato». Poi il candidato ha avuto un colloquio con un altro selezionatore, che gli ha posto la stessa domanda e ha ottenuto la stessa risposta. Nello scambio di impressioni successivo, tuttavia, i due selezionatori hanno manifestato pareri radicalmente diversi. L'uno, che fino a quel punto era pervenuto a una valutazione positiva del candidato, ha inteso la decisione di lasciare la società come un'indicazione di integrità morale e coraggio, mentre l'altro, la cui prima impressione era stata negativa, ha interpretato lo stesso fatto come un indice di inflessibilità e forse perfino di immaturità. Questa storia illustra che, per quanto vogliamo credere che il nostro giudizio su un candidato si basi sui fatti, il modo in cui interpretiamo tali fatti è condizionato dalle nostre attitudini pregresse.

I limiti dei colloqui tradizionali gettano seri dubbi sulla nostra capacità di trarne conclusioni sensate. Eppure le impressioni che ci si forma nel corso di un colloquio sono molto vivide, e di solito il selezionatore le ritiene affidabili. Nell'associare l'idea cui si giunge durante il colloquio con altri segnali riguardanti il candidato si tende ad attribuire troppo peso all'incontro personale e troppo poco ad altri dati che potrebbero essere più predittivi, come i risultati dei test.

Un aneddoto potrà dare maggior concretezza a questa osservazione. Ai professori che si sottopongono a un colloquio per una docenza universitaria spesso viene chiesto di tenere una lezione di fronte a una commissione composta da altri docenti, per appurare che le capacità di insegnamento dei candidati siano conformi agli standard dell'istituzione. Naturalmente qui la posta in gioco è molto più alta rispetto a una normale

lezione. Uno degli autori di questo libro una volta ha assistito alla prova di un candidato che ha dato una brutta impressione, evidentemente per lo stress della situazione: il suo curriculum riportava valutazioni straordinarie e diversi premi per l'eccellenza nell'insegnamento, eppure la vivida impressione prodotta dal suo insuccesso in una situazione altamente artificiale ha avuto un peso maggiore sulla decisione finale rispetto ai dati astratti sulle sue eccellenti prestazioni pregresse nell'attività di docente.

Un ultimo punto: quando i colloqui non sono l'unica fonte di informazioni sui candidati – per esempio quando vi sono anche test, referenze o altri input – tutti questi elementi devono essere associati in un giudizio complessivo. Si ripresenta dunque la consueta domanda: questi input andrebbero combinati mediante un giudizio (aggregazione clinica) o una formula (aggregazione meccanica)? Come abbiamo visto nel capitolo 9, l'approccio meccanico è superiore, sia in generale sia nel caso specifico della previsione delle prestazioni lavorative. Purtroppo i sondaggi indicano che la stragrande maggioranza dei professionisti delle risorse umane opta per l'aggregazione clinica, una pratica che aggiunge un'ulteriore fonte di rumore a un processo di per sé già molto rumoroso.[12]

Migliorare la selezione del personale attraverso la strutturazione

Se i colloqui tradizionali e le decisioni sulle assunzioni basate sui giudizi hanno una limitata validità predittiva, come possiamo agire? Fortunatamente le ricerche offrono dei suggerimenti per migliorare la selezione del personale, e alcune aziende stanno iniziando a metterli in pratica.

Un esempio di società che ha migliorato le proprie pratiche di selezione rendendone noti i risultati è Google, come racconta nel libro *Work Rules!* Laszlo Bock, ex vicepresidente senior delle risorse umane del colosso americano. Pur concentrandosi sull'assunzione di talenti di altissimo calibro e destinando risorse considerevoli alla ricerca delle persone giuste, l'azienda era in difficoltà. Un controllo della validità predittiva dei suoi colloqui di assunzione non trovò «nessuna relazione, [...] un caos totale».[13] I cambiamenti introdotti da Google per risolvere questi problemi riflettono diversi princìpi emersi in decenni di ricerche, e illustrano alcune utili strategie di igiene decisionale.

Una di queste dovrebbe ormai esservi nota: l'aggregazione. Il suo impiego in questo contesto non sorprende, dal momento che quasi tutte le società aggregano i giudizi di più selezionatori su uno stesso candidato. Per non essere da meno, talvolta Google infliggeva ai candidati fino a venticinque colloqui! Bock giunse alla conclusione che andavano ridotti a quattro, perché aveva riscontrato che quello era il numero di incontri oltre il quale ogni ulteriore colloquio non avrebbe aggiunto validità predittiva. Per garantire tale livello di validità, tuttavia, Google impone una regola stringente che non tutte le società osservano: si assicura che i selezionatori valutino i candidati separatamente, *prima* di comunicare tra loro. Vale la pena ribadirlo: l'aggregazione funziona, ma solo se i giudizi sono indipendenti.

Google adottò inoltre una strategia di igiene decisionale che non abbiamo ancora descritto nel dettaglio: *la strutturazione dei giudizi complessi.* Il termine "struttura" può significare molte cose, ma, in questa sede, un giudizio complesso strutturato verte su tre princìpi: scomposizione, indipendenza e giudizio olistico differito.

Il primo principio, la *scomposizione*, comporta la suddivisione della decisione in più componenti, o *valutazioni intermedie*. Questa fase ha la stessa finalità dell'identificazione dei sottogiudizi nelle linee guida: fa sì che i valutatori si concentrino sui segnali importanti. La scomposizione funge da piano d'azione per specificare quali dati sono necessari e lascia fuori le informazioni irrilevanti.

Nel caso di Google, la scomposizione consta di quattro valutazioni intermedie: capacità cognitiva generale, leadership, allineamento alla cultura aziendale (chiamato *googleyness*) e conoscenze associate al ruolo. (Alcune di queste valutazioni vengono poi a loro volta scomposte in componenti più piccole.) Si noti che l'aspetto gradevole del candidato, la parlantina, gli hobby interessanti e qualsiasi altro aspetto, positivo o negativo, che un reclutatore potrebbe osservare in un colloquio non strutturato non rientrano nell'elenco.

Creare questa sorta di struttura per un compito di reclutamento potrebbe sembrare una pura questione di buonsenso, ed effettivamente, se dovete assumere un ragioniere alle prime armi o un assistente amministrativo, esistono mansionari standard in cui vengono specificate le competenze richieste. Come sanno i reclutatori professionisti, tuttavia, una definizione degli aspetti fondamentali da valutare diventa difficile per posizioni insolite o di alto livello, e spesso la sua formulazione viene trascurata. Un famoso "cacciatore di teste" afferma che definire le competenze richieste in maniera sufficientemente specifica è un compito complicato e quindi spesso tralasciato, e sottolinea quanto sia importante che i decisori «investano nella definizione del problema», dedicando il tempo necessario, prima di incontrare i candidati, alla formulazione congiunta di una descrizione chiara e dettagliata dell'incarico.[14] Molti selezionatori, al contrario, usano descrizioni enfatiche frutto di accordi e

compromessi, che in questo modo diventano nient'altro che vaghe liste di desiderata comprensive di tutte le caratteristiche che dovrebbe possedere il candidato ideale, senza indicare in quale modo calibrarle o soppesarle.

Il secondo principio del giudizio strutturato, l'*indipendenza*, richiede che le informazioni relative a ciascun aspetto della valutazione vengano raccolte in forma indipendente. Non basta elencare le componenti della descrizione dell'incarico: molti dei reclutatori che conducono colloqui tradizionali sanno quali sono i quattro o cinque elementi da ricercare in un candidato, ma il problema è che, nel corso del colloquio, non li valutano singolarmente. In questo modo, ciascun aspetto influenza gli altri, il che rende ogni valutazione molto soggetta a rumore.

Per risolvere questo problema, Google ha orchestrato dei sistemi per compiere le valutazioni sulla base dei fatti e in forma indipendente. Forse il gesto più eclatante è stato l'introduzione delle *interviste comportamentali strutturate*.[15] Il compito del selezionatore in questo caso non è stabilire se il candidato gli faccia nel complesso una buona impressione, ma raccogliere dati su ogni singolo aspetto all'interno della struttura di valutazione, e assegnare un punteggio al candidato sulla base di ciascuno. A questo scopo, ai selezionatori viene chiesto di porre domande predefinite sul comportamento del candidato in situazioni passate. Devono inoltre registrare le risposte e assegnare un punteggio su una scala predeterminata, seguendo una griglia di valutazione unica che offre esempi di risposte medie, buone o ottime per ogni domanda. Questa scala condivisa (un esempio delle scale di valutazione ad ancoraggio comportamentale di cui abbiamo parlato nel precedente capitolo) contribuisce a ridurre il rumore nei giudizi.

Un approccio del genere vi sembrerà molto diverso dalle chiacchierate su cui si basano i colloqui tradizionali, e in effetti lo è. Anzi, potrebbe

sembrare più un esame o un interrogatorio che un incontro professionale, ed è emerso che tanto i selezionatori quanto i candidati non apprezzano le interviste strutturate (o comunque preferiscono quelle non strutturate). Vi è un costante dibattito su come dovrebbe presentarsi esattamente un'intervista per considerarsi strutturata,[16] eppure uno dei dati più ricorrenti riportati nella letteratura sul tema è che interviste di questo tipo sono molto più predittive delle performance future rispetto a quelle tradizionali non strutturate.[17] Le correlazioni con le prestazioni lavorative variano tra 0,44 e 0,57; se convertiamo queste cifre nella nostra misurazione in PC, le possibilità di scegliere il miglior candidato con un'intervista strutturata si collocano tra il 65 e il 69%, un sensibile miglioramento rispetto all'intervallo tra il 56 e il 61% dei colloqui classici.

Google utilizza altri dati come input per alcune delle dimensioni che ha più a cuore. Per verificare le conoscenze legate al lavoro, per esempio, si affida in parte a prove pratiche – o *work sample tests* – come chiedere a un candidato per un lavoro da programmatore di scrivere delle righe di codice.[18] Le ricerche dimostrano che queste prove sono tra i migliori predittori delle prestazioni lavorative. Google usa anche le "referenze di backdoor", ovvero quelle fornite non dalle persone nominate come referenti dai candidati, ma da suoi dipendenti che hanno incrociato il candidato nell'arco del suo percorso professionale.

Il terzo principio del giudizio strutturato, il *giudizio olistico differito*, si può riassumere in un semplice imperativo: non escludete l'intuizione, rimandatela. In Google, il parere finale sulle assunzioni viene pronunciato in forma collegiale da un comitato apposito, che vaglia un fascicolo completo di tutte le valutazioni ottenute dal candidato su ciascun aspetto in ogni singola intervista, e altre informazioni pertinenti a sostegno di tali

valutazioni. Sulla base di queste informazioni, il comitato decide poi se procedere con l'offerta.

Malgrado la cultura di quest'azienda sia notoriamente orientata ai dati, e a dispetto dell'evidenza per cui una combinazione meccanica degli stessi dà risultati migliori di una clinica, la decisione finale sull'assunzione *non* è meccanica. Resta un giudizio, in cui la commissione tiene conto di tutti i dati raccolti e li pesa in maniera olistica, ponendosi la domanda: «Questa persona avrà successo in Google?». La decisione non è frutto di un mero calcolo.

Nel prossimo capitolo spiegheremo perché riteniamo che questo approccio alla decisione finale abbia senso. Ma notate che, anche se non sono meccaniche, le decisioni finali di Google sono comunque ancorate al punteggio medio assegnato dai quattro selezionatori, e sono inoltre informate dalle evidenze soggiacenti. In altre parole, Google consente il giudizio e l'intuizione nel suo processo decisionale solo dopo la raccolta e l'analisi di ogni informazione utile. Pertanto, la tendenza di ciascun selezionatore (e membro della commissione selezionatrice) a formarsi impressioni veloci e intuitive e precipitarsi a formulare un giudizio viene tenuta sotto controllo.

I tre princìpi – che, ripetiamo, sono scomposizione, valutazione indipendente di ciascuna dimensione e giudizio olistico differito – non necessariamente costituiranno un modello per tutte le organizzazioni che cercano di migliorare il proprio processo di selezione, ma sono piuttosto coerenti con le raccomandazioni formulate dagli psicologi negli anni. Tali princìpi, in effetti, presentano una certa somiglianza con il metodo di selezione che uno degli autori di questo libro (Kahneman) introdusse nell'esercito israeliano nel lontano 1956, descritto in *Pensieri lenti e veloci.*[19] Questo processo, come quello utilizzato da Google, formalizzava una

struttura di valutazione (l'elenco delle dimensioni legate alla personalità e alla competenza che andavano valutate), richiedendo ai selezionatori di carpire dati oggettivi pertinenti a ogni dimensione e assegnare un punteggio a ciascuna, prima di passare alla successiva; inoltre, consentiva ai reclutatori di usare il proprio giudizio e il proprio intuito per arrivare a una decisione finale, ma solo dopo aver svolto la valutazione strutturata.

Vi sono evidenze schiaccianti sulla superiorità dei processi di giudizio strutturati (tra cui le interviste strutturate) nelle assunzioni. I dirigenti aziendali che necessitassero di suggerimenti per la loro adozione li troveranno in letteratura.[20] Come illustra l'esempio di Google e come osservato da altri ricercatori, i metodi di giudizio strutturati sono anche meno costosi, perché gli incontri faccia a faccia sono quanto di più oneroso un'azienda possa permettersi.

Detto ciò, molti dirigenti restano convinti dell'insostituibilità dei metodi informali basati sui colloqui, e, fatto interessante, lo stesso vale per molti candidati, secondo cui solo in un colloquio tradizionale potranno mostrare al futuro datore di lavoro il proprio valore. I ricercatori parlano in questo caso di «persistenza dell'illusione».[21] Una cosa è certa: sia i selezionatori sia i candidati sottovalutano nettamente la portata del rumore nei giudizi sulle assunzioni.

A proposito della strutturazione nelle assunzioni

«Nei colloqui informali tradizionali spesso abbiamo l'irresistibile sensazione istintiva di aver capito il candidato e di sapere se è la persona giusta. Dobbiamo imparare a diffidare di queste intuizioni.»

«I colloqui tradizionali sono pericolosi non solo per via dei bias, ma anche a causa del rumore.»

«Dobbiamo strutturare di più le nostre interviste e, in senso lato, il nostro processo di selezione. Cominciamo con il definire in maniera più chiara e specifica cosa cerchiamo nei candidati, e

assicuriamoci di valutarli in maniera indipendente su ognuna di queste dimensioni.»

[1] A.I. Huffcutt, S.S. Culbertson, *Interviews*, in S. Zedeck (a cura di), *APA Handbook of Industrial and Organizational Psychology*, American Psychological Association, Washington, DC 2010, pp. 185-203.

[2] N.R. Kuncel, D.M. Klieger, D.S. Ones, *In Hiring, Algorithms Beat Instinct*, in "Harvard Business Review", 92(2014), n. 5, p. 32.

[3] R.E. Ployhart, N. Schmitt, N.T. Tippins, *Solving the Supreme Problem: 100 Years of Selection and Recruitment at the Journal of Applied Psychology*, "Journal of Applied Psychology", 102(2017), pp. 291-304.

[4] M. McDaniel *et al.*, *Meta Analysis of the Validity of Employment Interviews*, in "Journal of Applied Psychology", 79(1994), pp. 599-616; A. Huffcutt, W. Arthur, *Hunter and Hunter (1984) Revisited: Interview Validity for Entry-Level Jobs*, in "Journal of Applied Psychology", 79(1994), p. 2; F.L. Schmidt, J.E. Hunter, *The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings*, in "Psychology Bulletin", 124(1998), pp. 262-274; F.L. Schmidt, R.D. Zimmerman, *A Counterintuitive Hypothesis About Employment Interview Validity and Some Supporting Evidence*, in "Journal of Applied Psychology", 89(2004), pp. 553-561. Si noti che le percentuali di validità sono più alte quando si considerano certe sottocategorie di studi, specialmente se le ricerche usano valutazioni delle prestazioni create *ad hoc*, piuttosto che valutazioni amministrative esistenti.

[5] S. Highhouse, *Stubborn Reliance on Intuition and Subjectivity in Employee Selection*, in "Industrial and Organizational Psychology", 1(2008), pp. 333-342; D.A. Moore, *How to Improve the Accuracy and Reduce the Cost of Personnel Selection*, in "California Management Review", 60(2017), pp. 8-17.

[6] L.A. Rivera, *Hiring as Cultural Matching: The Case of Elite Professional Service Firms*, in "American Sociology Review", 77(2012), pp. 999-1022.

[7] F.L. Schmidt, R.D. Zimmerman, *A Counterintuitive Hypothesis*, cit.; T.A. Judge, C.A. Higgins, D.M. Cable, *The Employment Interview: A Review of Recent Research and Recommendations for Future Research*, in "Human Resource Management Review", 10(2000), pp. 383-406; e A.I. Huffcutt, S.S. Culbertson, W.S. Weyhrauch, *Employment Interview Reliability: New Meta-Analytic Estimates by Structure and Format*, in "International Journal of Selection and Assessment", 21(2013), pp. 264-276.

[8] M.R. Barrick *et al.*, *Candidate Characteristics Driving Initial Impressions During Rapport Building: Implications for Employment Interview Validity*, in "Journal of Occupational and Organizational Psychology", 85(2012), pp. 330-352; M.R. Barrick, B.W. Swider, G.L. Stewart, *Initial Evaluations in the Interview: Relationships with Subsequent Interviewer Evaluations and Employment Offers*, in "Journal of Applied Psychology", 95(2010), p. 1163.

[9] G.L. Stewart *et al.*, *Exploring the Handshake in Employment Interviews*, in "Journal of Applied Psychology", 93(2008), pp. 1139-1146.

[10] T.W. Dougherty, D.B. Turban, J.C. Callender, *Confirming First Impressions in the Employment Interview: A Field Study of Interviewer Behavior*, in "Journal of Applied Psychology", 79(1994), pp. 659-665.

[11] J. Dana, R. Dawes, N. Peterson, *Belief in the Unstructured Interview: The Persistence of an Illusion*, in "Judgment and Decision Making", 8(2013), pp. 512-520.

[12] N.R. Kuncel *et al.*, *Mechanical versus Clinical Data Combination in Selection and Admissions Decisions: A Meta-Analysis*, in "Journal of Applied Psychology", 98(2013), n. 6, pp. 1060-1072.

[13] L. Bock, intervista con A. Bryant, "The New York Times", 19 giugno 2013. Vedi anche L. Bock, *Work Rules!: Insights from Inside Google That Will Transform How You Live and Lead*, Hachette, New York 2015.

[14] C. Fernandez-Araoz, *Hiring Without Firing*, in "Harvard Business Review", 1 luglio 1999.

[15] Per una guida accessibile alle interviste strutturate, vedi M.A. Campion, D.K. Palmer, J.E. Campion, *Structuring Employment Interviews to Improve Reliability, Validity and Users' Reactions*, in "Current Directions in Psychological Science", 7(1998), n. 3, pp. 77-82.

[16] J. Levashina *et al.*, *The Structured Employment Interview: Narrative and Quantitative Review of the Research Literature*, in "Personnel Psychology", 67(2014), pp. 241-293.

[17] M. McDaniel *et al.*, *Meta Analysis*, cit.; A. Huffcutt, W. Arthur, *Hunter and Hunter (1984) Revisited*, cit.; F.L. Schmidt, J.E. Hunter, *Validity and Utility*, cit.; F.L. Schmidt, R.D. Zimmerman, *A Counterintuitive Hypothesis*, cit.

[18] F.L. Schmidt, J.E. Hunter, *Validity and Utility*, cit.

[19] D. Kahneman, *Pensieri lenti e veloci*, cit., pp. 307-308.

[20] N.R. Kuncel, D.M. Klieger, D.S. Ones, *In Hiring, Algorithms Beat Instinct*, cit. Vedi anche M.A. Campion, D.K. Palmer, J.E. Campion, *Structuring Employment Interviews*, cit.

[21] J. Dana, R. Dawes, N. Peterson, *Belief in the Unstructured Interview*, cit.

25

Il protocollo a valutazioni intermedie

Qualche tempo fa due di noi (Kahneman e Sibony), insieme al nostro amico Dan Lovallo, hanno descritto un metodo per prendere decisioni all'interno delle organizzazioni che aveva come obiettivo primario la riduzione del rumore, definendolo *protocollo a valutazioni intermedie*.[1] Questo metodo racchiude molte delle strategie di igiene decisionale presentate nei capitoli precedenti, e può avere un largo impiego ogni volta che la valutazione di un progetto o di una scelta richieda di considerare e soppesare più dimensioni. Può essere impiegato e adattato in molti modi da organizzazioni di ogni tipo, comprese società, ospedali, università e agenzie governative.

Illustreremo qui il protocollo attraverso un esempio stilizzato che è in realtà un collage di diversi casi reali, quello di una società fittizia che chiameremo Mapco. Seguiremo le azioni intraprese da Mapco per sondare la possibilità di effettuare una grossa acquisizione che avrebbe un enorme impatto sulla società, e sottolineeremo come tali azioni differiscano da quelle normalmente intraprese da aziende che si trovano nella medesima situazione. Come vedrete, le differenze sono significative ma sottili: un osservatore poco attento potrebbe non coglierle.

La prima riunione: stabilire un approccio comune

Nella società Mapco si fece strada l'idea di acquisire la concorrente Roadco, e maturò al punto che i leader della compagnia pensarono di organizzare una riunione del consiglio d'amministrazione per discuterne. Joan Morrison, amministratrice delegata di Mapco, riunì il comitato strategico per un colloquio preliminare sulla possibile acquisizione e su quali azioni fosse necessario intraprendere per migliorare la qualità delle riflessioni che sarebbero emerse dal consiglio. All'inizio della riunione, Joan avanzò una proposta inaspettata:

«Vorrei proporvi di provare una nuova procedura nella riunione del consiglio d'amministrazione in cui si deciderà in merito all'acquisizione di Roadco. Ha un nome forse poco allettante, "protocollo a valutazioni intermedie", ma l'idea è semplicissima: si basa sull'analogia tra la valutazione di un'opzione strategica e quella di un candidato in un colloquio di lavoro.

«Come sicuramente saprete, le ricerche indicano che le interviste strutturate danno risultati migliori rispetto a quelle non strutturate, per cui possiamo trarre la conclusione generale che la strutturazione delle decisioni dà risultati migliori nelle assunzioni. Saprete anche che il nostro dipartimento Ricerca e Sviluppo ha adottato questi princìpi nel processo di selezione del personale. Numerose ricerche dimostrano che la strutturazione delle interviste porta a un livello di accuratezza molto più alto rispetto ai colloqui tradizionali che effettuavamo in passato.

«Ritengo che vi sia un'evidente somiglianza tra la valutazione dei candidati e quella delle opzioni tra cui scegliere quando si tratta di prendere grosse decisioni; insomma, *le opzioni sono assimilabili ai candidati.* È stata questa somiglianza a farmi pensare che dovremmo adattare al nostro compito, ovvero la valutazione delle opzioni strategiche, il metodo che si è dimostrato valido nella selezione dei candidati a un posto di lavoro.»

I membri del comitato inizialmente restarono perplessi di fronte a quell'analogia: il processo di reclutamento, fecero notare, è un ingranaggio ben oliato in cui vengono prese molte decisioni simili senza una forte pressione dal punto di vista del tempo a disposizione; una decisione strategica, d'altro canto, richiede un grande lavoro specifico e va presa in tempi brevi. Alcuni membri del comitato resero chiaro a Joan che avrebbero osteggiato qualsiasi proposta che ritardasse la decisione. Inoltre, temevano di appesantire troppo i requisiti di *due diligence* richiesti ai ricercatori di Mapco, ovvero la loro attività di approfondimento e raccolta di informazioni in vista della trattativa.

Joan impugnò subito queste obiezioni, assicurando ai colleghi che il processo strutturato non avrebbe ritardato il momento della scelta. «Si tratta solo di stabilire i punti all'ordine del giorno per la riunione del consiglio in cui discuteremo dell'operazione», spiegò. «Dovremo stabilire preventivamente un elenco delle valutazioni da ricavare riguardo a diversi aspetti dell'accordo, come un selezionatore parte da una descrizione del lavoro che funge da elenco dei tratti e delle caratteristiche di cui il candidato deve essere in possesso. Faremo in modo che il consiglio affronti queste valutazioni in forma separata, una alla volta, come i selezionatori nelle interviste strutturate valutano il candidato sulle varie dimensioni in maniera sequenziale. Solo allora avvieremo una discussione sull'eventualità di accettare o meno l'offerta. Questa procedura ci consentirà di trarre vantaggio dalla saggezza collettiva del consiglio in modo molto più efficace.

«Se opteremo per questo approccio, naturalmente, questo inciderà sul modo di presentare le informazioni e sulla preparazione della riunione da parte del *deal team*, la squadra che si occuperà di portare avanti la trattativa. Per questo vorrei sapere subito cosa ne pensate.»

Un membro del comitato, ancora scettico, chiese a Joan quali benefici quella strutturazione avesse apportato alla qualità delle decisioni di assunzione e perché ritenesse che fossero trasferibili alle decisioni strategiche. Lei gli illustrò la logica della proposta: attraverso il protocollo a valutazioni intermedie, spiegò, si massimizza il valore delle informazioni tenendo le dimensioni da giudicare indipendenti l'una dall'altra. «I dibattiti del nostro consiglio d'amministrazione sono molto simili alle interviste non strutturate», osservò. «Siamo sempre consapevoli che l'obiettivo finale è arrivare a una decisione, e vagliamo tutte le informazioni alla luce di tale obiettivo. Cominciamo pensando già al momento di chiudere, e ci arriviamo il prima possibile. Come un selezionatore in un'intervista non strutturata, rischiamo di sviluppare un dibattito che si limiti a confermare le nostre prime impressioni.

«Un approccio strutturato, invece, ci costringerà a posporre l'obiettivo del raggiungimento della decisione finché non avremo compiuto tutte le valutazioni, affrontandone una per volta come se fossero obiettivi intermedi: in questo modo, considereremo tutte le informazioni di cui disponiamo e saremo certi che la nostra conclusione in merito a un certo aspetto dell'accordo non cambi la nostra lettura di un altro aspetto non collegato.»

I membri del comitato accettarono di sperimentare quell'approccio, ma chiesero in cosa consistessero quelle valutazioni intermedie. Forse Joan aveva già in mente un elenco predefinito? «No», rispose lei. «Faremmo così se dovessimo applicare il protocollo a una decisione di routine, ma in questo caso dobbiamo essere noi a definire ogni passaggio intermedio di valutazione. Questo è un punto cruciale: sta a noi decidere quali sono i principali aspetti dell'acquisizione che andrebbero valutati.» Il comitato strategico decise di riunirsi il giorno seguente a tale scopo.

La seconda riunione: definire le valutazioni intermedie

«La prima cosa da fare», spiegò Joan, «è stendere un elenco completo delle valutazioni indipendenti sull'accordo, alle quali il gruppo di ricerca di Jeff Schneider assegnerà poi dei valori. Oggi il nostro compito sarà stilare questo elenco, che definisco "completo" in quanto ogni voce pertinente a cui riuscite a pensare dovrebbe essere inclusa e pesare su almeno una delle valutazioni. Quando dico "indipendenti" intendo che ciascuna voce dovrebbe di preferenza pesare su una sola delle valutazioni, per ridurre al minimo la ridondanza.»

Il gruppo si mise al lavoro e produsse un lungo elenco di fatti e dati che sembravano pertinenti, per poi organizzarli in un elenco di valutazioni. La sfida, scoprirono i partecipanti, stava nel creare un elenco breve, completo e composto da valutazioni che non si sovrapponessero. Difficile, ma non impossibile. In effetti, l'elenco finale delle sette valutazioni del gruppo era in apparenza simile all'indice che il consiglio si sarebbe aspettato di trovare in una normale relazione per la presentazione di una proposta di acquisizione. Oltre ai previsti modelli finanziari, l'elenco comprendeva, per esempio, una valutazione della qualità del gruppo dirigente della società target e una valutazione della probabilità che venissero colte le sperate sinergie.

Alcuni membri del comitato strategico si dissero delusi perché la riunione non aveva prodotto nuove idee su Roadco, ma Joan spiegò che non era quello il suo scopo: l'obiettivo immediato era dare istruzioni al deal team incaricato dello studio dell'acquisizione. Ogni valutazione, disse, sarebbe stata trattata in una sezione distinta della relazione del deal team e discussa in forma separata dal consiglio.

L'intento del deal team, secondo Joan, non era quello di dire al consiglio cosa pensasse nel complesso dell'accordo – o almeno, non in questa fase;

doveva invece esprimere un parere oggettivo e indipendente su ciascuna valutazione intermedia. In sostanza, spiegò Joan, ogni sezione della relazione del deal team doveva concludersi con un giudizio che rispondesse a una semplice domanda: «A prescindere dal peso da dare a questo aspetto nella decisione finale, in quale misura i risultati di questa valutazione depongono a favore o a sfavore dell'accordo?».

Il deal team

Il leader del team responsabile della valutazione dell'accordo, Jeff Schneider, quel pomeriggio riunì il suo gruppo per organizzare il lavoro. I cambiamenti da introdurre rispetto al loro consueto metodo non erano molti, ma lui ne sottolineò l'importanza.

Innanzitutto, spiegò, gli analisti del gruppo dovevano cercare di effettuare delle analisi il più possibile oggettive. Le valutazioni dovevano basarsi sui fatti – e fin qui niente di nuovo – ma anche adottare una *visione esterna*, ove possibile. Poiché i membri del gruppo non sapevano bene cosa intendesse con quell'espressione, Jeff fece due esempi, utilizzando due delle valutazioni intermedie identificate da Joan. Per valutare la probabilità che l'accordo ricevesse l'approvazione normativa, disse, dovevano partire da una ricerca del *tasso di base*, ovvero la percentuale di operazioni comparabili che vengono approvate. Questo compito, a sua volta, avrebbe richiesto la definizione di una relativa *classe di riferimento*, un gruppo di trattative di acquisizione considerate sufficientemente comparabili.

Jeff spiegò poi come valutare le competenze tecnologiche del dipartimento di sviluppo del prodotto della società target, un altro punto importante dell'elenco di Joan. «Non basta descrivere i recenti successi dell'azienda sulla base dei fatti e definirli "buoni" o "ottimi". Mi aspetto

piuttosto una valutazione di questo tipo: "Il dipartimento di sviluppo del prodotto è nel secondo quintile del suo gruppo di pari, sulla base della misurazione dei suoi lanci di prodotti recenti".» Nel complesso, spiegò, l'obiettivo era arrivare a valutazioni quanto più possibile comparative, perché i giudizi relativi sono migliori di quelli assoluti.

Jeff aveva un'altra richiesta. Secondo le istruzioni di Joan, disse, le valutazioni dovevano essere il più possibile indipendenti l'una dall'altra, per ridurre il rischio che si influenzassero a vicenda. Di conseguenza, assegnò a diversi analisti le diverse valutazioni, e chiese loro di lavorare autonomamente.

Alcuni di loro si dissero sorpresi. «Non è preferibile lavorare in gruppo?» gli chiesero. «Che senso ha formare un gruppo se non possiamo comunicare tra noi?»

Jeff si rese conto di dover spiegare perché fosse necessario esprimere valutazioni indipendenti. «Probabilmente saprete che nella selezione del personale è presente l'effetto alone», disse. «È ciò che succede quando l'impressione generale che dà un candidato influenza la valutazione delle sue capacità rispetto a una dimensione specifica: esattamente quello che stiamo cercando di evitare.» Poiché alcuni analisti ritenevano che non fosse un problema importante, Jeff ricorse a un'altra analogia: «Se ci fossero quattro testimoni di un crimine, sarebbe giusto che si consultassero prima di testimoniare? Ovviamente no. Non vogliamo che un testimone influenzi gli altri». Gli analisti non trovarono quel paragone particolarmente lusinghiero, ma il messaggio era passato, pensò Jeff.

Si dà il caso, però, che Jeff non disponesse di un numero sufficiente di analisti per arrivare a valutazioni perfettamente indipendenti. A Jane, un membro esperto del team, vennero affidate due valutazioni: Jeff scelse le due che gli sembravano maggiormente diverse tra loro e le chiese di

completare la prima valutazione, redigendo la relativa relazione, e solo dopo passare alla seconda. Un altro problema era il giudizio sulla qualità del gruppo dirigente dell'azienda target; Jeff temeva che i suoi analisti avrebbero faticato a dissociare la propria valutazione delle qualità intrinseche dei dirigenti dai giudizi sui risultati recenti della società (che, naturalmente, il team avrebbe analizzato nel dettaglio). Per risolvere il problema, chiese a un esperto esterno di pronunciarsi sulla qualità del gruppo dirigente; in questo modo, pensò, avrebbe ottenuto un input più indipendente.

Jeff diede poi un'altra istruzione che il team trovò in un certo senso insolita. Ogni capitolo del rapporto finale doveva concentrarsi su una dimensione e, come richiesto da Joan, pervenire a una conclusione sotto forma di punteggio. Tuttavia, aggiunse, gli analisti dovevano includere in ciascun capitolo tutte le informazioni fattuali pertinenti ai fini della valutazione. «Non occultate nulla», precisò. «Il tono generale del capitolo naturalmente sarà in linea con la valutazione proposta, ma se ci sono informazioni che sembrano incoerenti o anche contraddittorie rispetto alla valutazione principale, non nascondetele sotto il tappeto. Il punto non è dare credito alla vostra raccomandazione, ma rappresentare la verità. Se è complicata, che lo sia pure – non sarebbe affatto insolito.»

Nello stesso spirito, Jeff invitò gli analisti alla massima trasparenza sul proprio livello di fiducia in ciascuna valutazione. «Il consiglio di amministrazione sa che non siete in possesso di tutte le informazioni, e li aiuterà sapere quando siete all'oscuro di qualcosa. Se vi imbattete in qualcosa che davvero vi fa pensare – un potenziale motivo per non concludere l'accordo – non esitate a riferirlo subito.»

Il deal team seguì le istruzioni, e fortunatamente non trovò grosse ragioni per sconsigliare l'operazione. Il suo rapporto conclusivo per Joan e

il consiglio teneva conto di tutte le valutazioni identificate.

La riunione decisionale

Nel leggere la relazione del team prima della riunione decisionale, Joan notò immediatamente un elemento importante: se le valutazioni erano per lo più a sostegno dell'accordo, il quadro che tratteggiavano non era affatto semplice, né roseo. Alcune delle valutazioni erano forti, altre no. Queste differenze, come sapeva, erano il risultato prevedibile dell'averle tenute indipendenti l'una dall'altra: quando si tiene a freno l'eccesso di coerenza, la realtà è molto meno coerente di quanto la si voglia far sembrare nei consigli d'amministrazione. «Bene», si disse Joan. «Queste discrepanze tra le valutazioni solleveranno dei dubbi e accenderanno il dibattito, proprio quello che ci serve per una discussione proficua in sede di consiglio. I diversi risultati renderanno la decisione non certo più facile, ma sicuramente migliore.»

Convocata la riunione del consiglio d'amministrazione, illustrò l'approccio seguito dal deal team e invitò i membri del consiglio ad applicare lo stesso principio. «Jeff e il suo team hanno lavorato sodo per tenere le valutazioni indipendenti l'una dall'altra, e il nostro compito è di esaminarle in forma altrettanto indipendente. Ciò significa che, prima di metterci a discutere sulla decisione finale, considereremo ogni valutazione in maniera autonoma, trattandola come se fosse un diverso punto all'ordine del giorno.»

I membri del consiglio sapevano che seguire questo approccio strutturato sarebbe stato difficile: Joan stava chiedendo loro di non arrivare a una visione olistica dell'accordo prima di aver analizzato tutte le valutazioni, ma molti di loro erano addetti ai lavori, e *avevano già* un'idea di

Roadco. Non parlarne sembrava innaturale, ma poiché comprendevano l'obiettivo di Joan, accettarono di stare alle sue regole e si astennero temporaneamente dall'esporre le proprie idee di massima.

Con grande sorpresa, scoprirono che quell'approccio era molto efficace. Nel corso della riunione alcuni di loro cambiarono perfino idea sull'accordo (anche se nessuno lo venne a sapere, perché lo tennero per sé). Il modo in cui Joan condusse la riunione fu cruciale: impiegò il metodo *estimate-talk-estimate*, in cui i vantaggi del dibattito si sommano a quelli della media delle opinioni indipendenti.[2]

Si procedette in questo modo: per ogni valutazione, Joan chiese a Jeff, a nome del deal team, di riassumere brevemente i punti chiave (che i membri del consiglio avevano precedentemente letto nel dettaglio). Poi disse ai membri del consiglio di utilizzare un'app precedentemente scaricata sul cellulare per assegnare il proprio punteggio a una data dimensione, che poteva essere uguale o diverso rispetto a quello del deal team. La distribuzione dei punteggi venne proiettata immediatamente sullo schermo, senza identificare i valutatori. «Questo non è un voto», spiegò Joan. «Stiamo solo sondando l'umore dei presenti su ciascun tema.» Dando una lettura immediata dell'opinione indipendente dei membri del consiglio prima di avviare la discussione, Joan ridusse il rischio dell'influenza sociale e delle cascate informative.

Su alcune valutazioni vi era un consenso immediato, su altre emersero visioni contrastanti. Naturalmente Joan gestì la discussione in modo da dedicare più tempo a queste ultime; diede la parola ai membri di entrambi gli schieramenti, incoraggiandoli a esprimere il proprio punto di vista adducendo fatti e argomentazioni, ma anche con la giusta umiltà e apertura alle ragioni degli altri. Una volta, quando un membro del consiglio particolarmente favorevole all'acquisizione si accalorò oltre misura, lei gli

ricordò che «se delle persone ragionevoli come noi non sono d'accordo, vorrà dire che tra persone ragionevoli si può essere in disaccordo su questo tema».

Quando il dibattito su una valutazione volgeva al termine, Joan chiedeva ai membri del consiglio di votare di nuovo: quasi sempre l'accordo era superiore rispetto al primo giro. La stessa sequenza – prima stima, discussione, seconda stima – venne ripetuta per ogni valutazione.

Infine arrivò il momento di tirare le conclusioni sull'accordo. Per facilitare la discussione, Jeff mostrò l'elenco delle valutazioni sulla lavagna e, per ciascuna, la media dei punteggi assegnati dai presenti. I membri del consiglio d'amministrazione si trovarono così di fronte a un profilo dell'accordo, ma su quali basi dovevano decidere?

Uno degli amministratori avanzò un semplice suggerimento: impiegare una media semplice dei punteggi. (Forse conosceva i vantaggi dell'aggregazione meccanica rispetto ai giudizi clinici olistici, di cui abbiamo parlato nel capitolo 9.) Un altro, però, immediatamente obiettò che, secondo lui, ad alcune valutazioni andava attribuito un peso molto più alto che ad altre. Una terza persona non era d'accordo, e propose una gerarchizzazione diversa delle valutazioni.

Joan li interruppe. «Non si tratta semplicemente di fare un calcolo a partire da un insieme di punteggi», disse. «Abbiamo accantonato le intuizioni, ma ora è arrivato il momento di usarle. Ora ci servono i vostri giudizi.»

Joan non motivò la sua logica, ma aveva imparato dall'esperienza che, soprattutto nelle decisioni importanti, la gente rifiuta gli schemi che imbrigliano e non permettono di ricorrere ai propri giudizi. Sapeva che spesso i decisori sono in grado di manipolare il sistema se sanno che verrà impiegata una formula, modificando i propri punteggi per arrivare alla

conclusione desiderata e in questo modo vanificando l'intento dell'intero esercizio. Peraltro, anche se non era questo il caso, era consapevole della possibilità che potessero emergere considerazioni decisive che non erano state previste nella definizione delle valutazioni (l'eccezione della gamba rotta discussa nel capitolo 10). Se si fosse presentata una ragione imprevista per non concludere l'accordo (o, al contrario, un motivo in più per siglarlo), un processo decisionale puramente meccanico basato sulla media delle valutazioni avrebbe potuto causare grossi errori.

Joan sapeva anche che permettere ai membri del consiglio d'amministrazione di usare il proprio intuito in questa fase era molto diverso dal lasciare che lo usassero in una delle fasi precedenti del processo. Ora che le valutazioni erano disponibili e note a tutti, la decisione finale era saldamente ancorata a quei punteggi basati sui fatti su cui avevano discusso a fondo. Un membro del consiglio avrebbe dovuto addurre motivazioni forti per opporsi all'accordo di fronte a un elenco di valutazioni intermedie tendenzialmente favorevoli. In base a questa logica, il consiglio discusse l'acquisizione e votò nel modo consueto.

Il protocollo a valutazioni intermedie nelle decisioni ricorrenti

Abbiamo descritto il protocollo a valutazioni intermedie nel contesto di una singola decisione, unica e irripetibile, ma la stessa procedura si applica anche alle decisioni ricorrenti. Immaginate che Mapco non stia affrontando una sola acquisizione, ma sia un fondo di capitale di rischio che effettui investimenti reiterati nelle startup. Il protocollo sarebbe ugualmente applicabile e il procedimento sarebbe più o meno lo stesso, con due piccole differenze che, semmai, lo renderebbero più semplice.

Innanzitutto, la definizione iniziale dell'elenco di valutazioni intermedie verrà effettuata una volta sola. Il fondo ha dei criteri che applica a tutti i potenziali investimenti: questi corrisponderanno alle singole dimensioni di valutazione, senza che ci sia bisogno di riformularli ogni volta.

In secondo luogo, se il fondo prende molte decisioni dello stesso tipo, può basarsi sull'esperienza per calibrare i propri giudizi. Considerate, per esempio, una valutazione che ogni fondo di capitale di rischio si troverà a effettuare: quella sulla qualità del gruppo dirigente dell'azienda su cui investire. Nelle pagine precedenti abbiamo consigliato di effettuare stime di questo tipo mettendole in relazione con una classe di riferimento. Certo, non possiamo che dirci solidali con gli analisti di Mapco: raccogliere dati su società comparabili, oltre a valutare una specifica società target, è difficile. Tuttavia, i giudizi comparativi diventano molto più semplici nel contesto di decisioni ricorrenti. Se avete valutato il gruppo dirigente di decine o magari anche centinaia di società, potete impiegare questa esperienza condivisa come classe di riferimento. In pratica, potreste creare una scala di casi definita sulla base di casi di ancoraggio, in modo da poter dire, per esempio, che il gruppo dirigente della società target è «valido quanto quello della società ABC che abbiamo acquisito», ma «non altrettanto di quello della compagnia DEF». I casi di ancoraggio, naturalmente, devono essere noti a tutti i partecipanti (e periodicamente aggiornati), e la loro definizione richiede un investimento iniziale in termini di tempo. Ma il valore di questo approccio sta nel fatto che, come abbiamo visto, i giudizi relativi (comparare un dato gruppo dirigente a quelli di ABC e di DEF) sono molto più affidabili delle valutazioni assolute operate su una scala numerica o aggettivale.

Cosa cambia con il protocollo

Per facilità di consultazione, nella tabella 4 riassumiamo i principali cambiamenti introdotti dal protocollo a valutazioni intermedie.

Tabella 4. Le fasi principali del protocollo a valutazioni intermedie

1. All'inizio del processo, strutturare la decisione in valutazioni intermedie. *(Per i giudizi ricorrenti, va fatto una sola volta.)*
2. Fare in modo che, ove possibile, nelle valutazioni intermedie si adotti una visione esterna. *(Per i giudizi ricorrenti, impiegare giudizi relativi, se è possibile utilizzando una scala di casi.)*
3. In fase di analisi, mantenere le valutazioni il più possibile indipendenti l'una dall'altra.
4. Nella riunione decisionale, esaminare separatamente ogni valutazione.
5. Per ciascuna valutazione, fare in modo che i partecipanti esprimano i propri giudizi in forma individuale, dopodiché impiegare il metodo *estimate-talk-estimate*.
6. Per prendere la decisione finale, ritardare il ricorso all'intuito, senza però bandirlo.

Forse avrete riconosciuto un'applicazione di diverse tecniche di igiene decisionale presentate nei capitoli precedenti: sequenziamento delle informazioni, strutturazione delle decisioni in valutazioni indipendenti, impiego di un quadro di riferimento comune fondato sulla visione esterna, e aggregazione dei giudizi indipendenti di più individui. Applicando queste tecniche, il protocollo a valutazioni intermedie punta a cambiare il *processo* decisionale per introdurre più igiene decisionale possibile.

Senz'altro quest'enfasi sul processo più che sul contenuto delle decisioni potrà sollevare qualche dubbio, e le reazioni dei membri del gruppo di ricerca e del consiglio di amministrazione che abbiamo descritto sono tutt'altro che insolite. Il contenuto è specifico, il processo è generico; usare l'intuito e il giudizio è divertente, seguire un procedimento no. È opinione diffusa che le buone decisioni, e in particolar modo quelle che si rivelano le migliori, emergano dalle idee e dalla creatività dei grandi leader – ci piace

crederlo soprattutto quando i leader in questione siamo noi –, e a molti la parola *processo* evocherà subito burocrazia e ritardi.

La nostra esperienza con società e agenzie governative che hanno implementato alcune o tutte le componenti del protocollo ci dice che questi timori sono infondati. Sicuramente aggiungere ulteriore complessità ai processi decisionali di un'organizzazione già burocratizzata non migliorerà le cose. Ma non è detto che l'igiene decisionale debba essere lenta, né tantomeno macchinosa: al contrario, promuove il confronto e il dibattito, non il consenso soffocante tipico della burocrazia.

I vantaggi dell'igiene decisionale sono evidenti. I leader del settore pubblico e privato di solito sono del tutto ignari del rumore presente nelle proprie decisioni più importanti, e di conseguenza non prendono alcuna misura specifica per ridurlo. In questo senso sono come i reclutatori che continuano a fare affidamento su interviste non strutturate come unico strumento di selezione del personale: ignari del rumore nei propri giudizi, più fiduciosi del dovuto in merito alla loro validità, e inconsapevoli delle procedure che potrebbero migliorarli.

Così come lavarci le mani non ci protegge da tutte le malattie, l'igiene decisionale non ci risparmia da tutti gli errori. Non renderà brillante ogni nostra decisione, ma, proprio come lavarsi le mani, è un modo per affrontare un problema invisibile ma pervasivo e dannoso. Dove c'è giudizio, c'è rumore, e noi proponiamo l'igiene decisionale come strumento per ridurlo.

A proposito del protocollo a valutazioni intermedie

«Adottiamo già un processo strutturato per prendere decisioni sulle assunzioni, quindi perché non utilizzarne uno anche per le scelte strategiche? Dopotutto, le opzioni sono come i candidati.»

«Questa è una decisione difficile. Quali sono le valutazioni intermedie su cui dovremmo basarci?»

«Il nostro giudizio olistico intuitivo su questo progetto è molto importante, ma non parliamone ora. Ricorreremo all'intuito solo una volta che le valutazioni indipendenti che abbiamo chiesto ci daranno un quadro più completo su cui basarci.»

[1] D. Kahneman, D. Lovallo, O. Sibony, *A Structured Approach to Strategic Decisions: Reducing Errors in Judgment Requires a Disciplined Process*, in "MIT Sloan Management Review", 60(2019), pp. 67-73.

[2] A.H. Van De Ven, A. Delbecq, *The Effectiveness of Nominal, Delphi, and Interacting Group Decision Making Processes*, in "Academy of Management Journal", 17(1974), n. 4, pp. 605-621. Vedi anche capitolo 21.

SESTA PARTE

Rumore ottimale

Nel 1973 il giudice Marvin Frankel aveva ottime ragioni per richiedere un'azione prolungata finalizzata a ridurre il rumore nelle condanne penali. Il suo controllo del rumore intuitivo e informale, seguito da azioni più formali e sistematiche, rivelò disparità ingiustificate, vergognose e preoccupanti nel trattamento di casi simili.

Gran parte di questo libro può essere inteso come un tentativo di generalizzare le argomentazioni di Frankel e offrire un'interpretazione delle loro basi psicologiche. A qualcuno la presenza del rumore nel sistema di giustizia penale potrà sembrare particolarmente intollerabile e perfino scandalosa, ma anche in innumerevoli altri contesti non si può certo definire tollerabile, perché persone che dovrebbero essere intercambiabili, sia nel settore pubblico sia in quello privato, esprimono giudizi diversi in ambito professionale. Nelle assicurazioni, nella scelta e nella valutazione dei dipendenti, in medicina, nella scienza forense, nell'istruzione, in contesti aziendali e governativi il rumore interpersonale è una grande fonte di errore. Abbiamo inoltre osservato che tutti noi siamo soggetti al rumore occasionale, in quanto elementi che dovrebbero essere irrilevanti possono indurci a esprimere giudizi diversi a seconda che sia mattina o pomeriggio, lunedì o giovedì.

Ma, come indica la reazione decisamente negativa dei giudici alle linee guida sulle condanne, i tentativi di ridurre il rumore spesso incorrono in obiezioni serie e veementi. Molti sostengono che le linee guida siano rigide, disumanizzanti e a loro volta ingiuste; sarà capitato a tutti di avanzare una

richiesta ragionevole a una società, a un datore di lavoro, allo stato, e sentirsi rispondere: «Vorremmo aiutarla, ma abbiamo le mani legate, ci sono precise regole da seguire». Potranno anche sembrare sciocche e addirittura crudeli, ma le regole in questione forse sono state adottate per un buon motivo: ridurre il rumore (e magari anche il bias).

Detto ciò, alcune azioni volte a diminuire il rumore sollevano serie preoccupazioni, soprattutto se riducono o minano il diritto a difendersi delle persone. L'uso degli algoritmi e dell'apprendimento automatico ha gettato una nuova luce su tale obiezione, al di là di ogni fanatismo tecnologico.

Una critica autorevole è stata mossa da Kate Stith della Yale Law School e dal giudice federale José Cabranes, che hanno attaccato con forza le linee guida sulle condanne e, in un certo senso, una delle tesi centrali di questo libro. Il loro attacco era circoscritto all'ambito delle condanne penali, ma la stessa obiezione si può avanzare contro molte strategie di riduzione del rumore nell'ambito dell'istruzione, del commercio, dello sport e in moltissimi altri ancora. Stith e Cabranes sostengono che le linee guida sulle condanne siano animate «da una paura dell'esercizio della discrezionalità – una paura di giudicare – e da una fede tecnocratica negli esperti e nella pianificazione centralizzata». Ritengono che la «paura di giudicare» porti a proibire l'esame dei «particolari di ogni caso». Secondo loro, «nessuna soluzione meccanica può soddisfare il bisogno di giustizia».[1]

Vale la pena di esaminare queste obiezioni. In contesti che richiedono un giudizio di qualsiasi tipo, spesso si crede che il «bisogno di giustizia» sia incompatibile con qualsiasi soluzione meccanica: in questo modo, si consentono o addirittura si impongono processi e approcci che finiscono per garantire la presenza del rumore. Molti invitano a porre attenzione ai «particolari di ogni caso»; negli ospedali, nelle scuole e nelle ditte private,

grandi o piccole che siano, questo richiamo ha un fascino profondo e intuitivo. Ma come abbiamo visto, l'igiene decisionale comprende diverse strategie per la riduzione del rumore, molte delle quali non prevedono soluzioni meccaniche; quando un problema viene scomposto nelle sue componenti, non necessariamente il giudizio dovrà essere meccanico. Nonostante questo, molte persone non vedono di buon occhio l'impiego di strategie di igiene decisionale.

Abbiamo definito il rumore una variabilità indesiderata, e se qualcosa è indesiderato, probabilmente andrebbe eliminato. Ma l'analisi ci porta a una considerazione più complessa e anche più interessante: il rumore potrà essere indesiderato, a parità di altre condizioni, ma non è detto che le altre condizioni restino immutate, e i costi dell'eliminazione del rumore potrebbero superare i benefici. E anche quando un'analisi di costi e benefici indica che il rumore sta avendo effetti onerosi, eliminarlo potrebbe produrre una serie di conseguenze terribili e perfino inaccettabili per le istituzioni pubbliche e private.

Vi sono sette obiezioni fondamentali alle azioni di riduzione o eliminazione del rumore.

Prima obiezione: queste azioni possono avere un costo eccessivo, e allora il gioco non vale la candela. I passi necessari per ridurre il rumore potrebbero essere molto gravosi, e in certi casi addirittura insostenibili.

Seconda: alcune strategie adottate per ridurre il rumore potrebbero introdurre a loro volta degli errori, e in alcuni casi produrre un bias sistematico. Se tutti i previsori di un ufficio governativo partissero dagli stessi presupposti irrealisticamente ottimistici, le loro previsioni non sarebbero affette da rumore, ma nondimeno sarebbero errate. Se tutti i medici di un ospedale prescrivessero l'aspirina per ogni malattia, non produrrebbero rumore, ma commetterebbero molti errori.

Tratteremo queste obiezioni nel capitolo 26, mentre nel capitolo 27 ne affronteremo altre cinque altrettanto comuni, che probabilmente sentiremo avanzare da varie parti nei prossimi anni, quando ci si affiderà sempre di più a regole, algoritmi e apprendimento automatico.

Terza: se vogliamo che le persone sentano di essere trattate con rispetto e dignità, potrebbe essere necessario tollerare un certo grado di rumore. Il rumore, infatti, può essere un effetto collaterale di un processo imperfetto che si finisce per adottare perché permette a tutti (dipendenti, clienti, candidati, studenti, imputati) di ricevere un trattamento personalizzato, di avere l'opportunità di influenzare l'esercizio della discrezionalità e di far sentire la propria voce.

Quarta: il rumore può essere essenziale per accogliere nuovi valori, e quindi consentire uno sviluppo politico e morale. Eliminando il rumore, potremmo ridurre la nostra capacità di reagire quando l'impegno morale e politico ci porta in direzioni del tutto nuove. Un sistema privo di rumore potrebbe cristallizzare i valori esistenti.

Quinta: alcune strategie progettate per ridurre il rumore potrebbero incoraggiare comportamenti opportunistici, permettendo a determinate persone di ingannare il sistema o eludere i divieti. Un po' di rumore, o anche molto, potrebbe essere necessario per prevenire gli illeciti.

Sesta: un processo affetto da rumore può essere un buon deterrente. Se la gente sa che potrebbe incorrere in una pena lieve così come in una severa, è possibile che si astenga dal delinquere, almeno se è avversa al rischio. Un sistema può tollerare il rumore con la finalità di aumentare l'effetto deterrente.

Settima e ultima: le persone non vogliono essere trattate come oggetti o ruote di un ingranaggio. Alcune strategie di riduzione del rumore potrebbero reprimere la creatività e avere un effetto demoralizzante.

Anche se prenderemo in esame queste obiezioni con tutta la comprensione possibile, questo non vuol dire che le facciamo nostre, almeno se intese come motivazioni per bocciare l'obiettivo generale della riduzione del rumore. Per anticipare un'idea che ripeteremo più volte, specifichiamo che un'obiezione diventa più o meno convincente solo se riferita a una particolare strategia di riduzione del rumore. Per esempio, potrete disapprovare l'applicazione di rigide linee guida e invece concordare sul fatto che aggregare giudizi indipendenti sia una buona idea. L'impiego del protocollo a valutazioni intermedie potrà sembrarvi una complicazione inutile, ma allo stesso tempo potrete promuovere a pieni voti l'utilizzo di una scala condivisa che poggi sulla visione esterna. Tenendo conto di tutto questo, siamo giunti alla conclusione generale che, pur dando il giusto peso alle obiezioni elencate, la riduzione del rumore resta un obiettivo meritevole e perfino impellente. Nel capitolo 28 difenderemo la nostra tesi a partire da un dilemma che affrontiamo tutti i giorni, anche se non sempre ne siamo consapevoli.

[1] K. Stith, J.A. Cabranes, *Fear of Judging: Sentencing Guidelines in the Federal Courts*, University of Chicago Press, Chicago 1998, p. 177.

26

I costi della riduzione del rumore

Ogni volta che si chiede di eliminare il rumore, qualcuno obietta che i passi necessari da intraprendere comporterebbero una spesa eccessiva. In circostanze estreme, la riduzione del rumore non è proprio possibile. Ci è stata posta un'obiezione di questo tipo in ambito aziendale, governativo, scolastico e in diversi altri contesti. È una preoccupazione legittima, ma talvolta esagerata e usata sostanzialmente come scusa.

Per dare il massimo credito a quest'obiezione, poniamo il caso di un insegnante delle scuole superiori che ogni settimana valuti venticinque temi svolti da studenti del secondo anno. Se non dedica più di un quarto d'ora a ogni tema, la valutazione potrebbe essere affetta da rumore, quindi inaccurata e iniqua. Potrebbe allora prendere in considerazione una piccola misura di igiene decisionale, ovvero cercare di ridurre il rumore chiedendo anche a un suo collega di valutare gli stessi temi, in modo che ogni elaborato venga letto da due persone. O forse potrebbe raggiungere lo stesso obiettivo dedicando più tempo alla lettura di ogni saggio, strutturando il processo relativamente complesso di valutazione o leggendo i temi più di una volta in ordine diverso; potrebbe servirsi di linee guida dettagliate a mo' di checklist per l'assegnazione dei voti, oppure decidere di leggere tutti i temi nello stesso momento della giornata, in modo da ridurre il rumore occasionale.

Ma se i suoi giudizi sono piuttosto accurati e non troppo inclini al rumore, avrebbe senso anche non fare niente di tutto ciò: forse non ne

varrebbe la pena. L'insegnante potrebbe pensare che impiegare una checklist o chiedere a un collega di rileggere gli stessi temi sia un'esagerazione. Per scoprire se ha ragione o no, potrebbe occorrere un'attenta analisi: quanto ne guadagnerebbe in termini di accuratezza, quanto è importante arrivare a un giudizio più accurato, quanto tempo e denaro richiederebbe lo sforzo di ridurre il rumore? È facile immaginare un limite alle risorse che si è disposti a investire in questo tentativo, ma è altrettanto facile sottolineare che tale limite dovrebbe essere diverso a seconda che si tratti di temi scritti da studenti del secondo anno o di tesine finali, che negli Stati Uniti, per esempio, possono incidere sull'eventuale accesso all'università.

Questa semplice analisi potrebbe essere estesa a situazioni più complesse in cui sono coinvolte organizzazioni pubbliche e private di ogni tipo, e condurre a un rifiuto di alcune strategie di riduzione del rumore. Per alcune malattie gli ospedali e i medici potrebbero avere difficoltà a identificare delle linee guida di base per eliminare la variabilità. Nel caso di diagnosi mediche divergenti, le azioni per ridurre il rumore sono particolarmente utili, in quanto potrebbero contribuire a salvare delle vite umane, ma occorre tenere conto della fattibilità e dei costi di tali azioni. Alcuni esami potrebbero eliminare il rumore nelle diagnosi, ma se si tratta di controlli invasivi, pericolosi e costosi, e se la variabilità delle diagnosi è tutto sommato modesta e ha conseguenze lievi, forse non sarà il caso di prescriverli a tutti i pazienti.

Raramente la valutazione dei dipendenti è una questione di vita o di morte, ma il rumore può essere all'origine di parzialità nei loro confronti e comportare alti costi per l'azienda. Come abbiamo visto, le azioni per ridurre il rumore dovrebbero essere praticabili, ma spesso ci si chiede se valga la pena di intraprenderle; nel caso in cui emergano valutazioni

chiaramente errate, per esempio, potrebbe generarsi imbarazzo, vergogna o peggio ancora. Un'istituzione potrebbe ritenere che non sia il caso di intraprendere complesse azioni correttive; talvolta questa conclusione si rivelerà miope, opportunista e sbagliata, con esiti anche catastrofici, mentre alcune forme di igiene decisionale potrebbero essere molto vantaggiose. Eppure l'idea che ridurre il rumore richieda esborsi eccessivi non è sempre sbagliata.

Insomma, è giusto mettere a confronto i costi e i benefici della riduzione del rumore, ed è per questo che i controlli del rumore sono così importanti: in molte situazioni rivelano un'iniquità scandalosa, costi molto elevati, o entrambe le cose. In questo caso, il costo della riduzione del rumore non è certo un buon motivo per non procedere.

Meno rumore ma più errori?

C'è chi lamenta che determinate azioni per la riduzione del rumore potrebbero a loro volta produrre livelli inaccettabili di errore: è un'obiezione comprensibile se gli strumenti impiegati sono troppo blandi. Effettivamente, alcune di queste azioni potrebbero addirittura aumentare i bias: se un social media come Facebook o Twitter introducesse linee guida ferree per l'eliminazione di tutti i post in cui compaiono determinate espressioni volgari, le sue decisioni sarebbero meno affette da rumore, ma cancellerebbero molti post che non andrebbero eliminati. Questi falsi positivi costituiscono un errore direzionale, cioè un bias.

Il mondo pullula di riforme istituzionali introdotte per ridurre la discrezionalità delle persone, e di pratiche che generano rumore. Molte di queste sono motivate, ma talvolta la cura è peggiore della malattia. In *Retoriche dell'intransigenza*, l'economista Albert Hirschman affronta tre

classiche obiezioni ai tentativi di riforma. Innanzitutto, potrebbero avere effetti perversi, nel senso che aggraverebbero il problema stesso che intendono risolvere; in secondo luogo, potrebbero essere futili, non generando alcun effetto; infine, metterebbero a repentaglio altri importanti valori (come quando si fa notare che un'azione mirante a proteggere i sindacati e i diritti sindacali danneggia la crescita economica).[1] La perversità, la futilità e la messa a repentaglio potrebbero costituire delle obiezioni alla riduzione del rumore, e delle tre accuse sono la prima e la terza quelle che tendono a essere le più forti. A volte queste obiezioni sono soltanto retoriche, e il loro vero obiettivo è far deragliare una riforma che sarebbe invece molto utile, ma è vero che alcune strategie di riduzione del rumore potrebbero mettere a repentaglio importanti valori, e per altre il rischio di perversità potrebbe non essere facile da scongiurare.

I giudici che obiettavano alle linee guida sulle condanne si riferivano proprio a questo rischio: conoscevano bene lo studio del giudice Frankel e non negavano che la discrezionalità fosse una fonte di rumore, ma ritenevano che ridurre quest'ultima avrebbe portato a un aumento anziché a una diminuzione degli errori. Citando Václav Havel, sostenevano che «dobbiamo abbandonare l'arrogante convinzione che il mondo non sia che un enigma da sciogliere, una macchina di cui scoprire le istruzioni per l'uso, un corpus di informazioni da dare in pasto a un computer nella speranza che, presto o tardi, possa emettere una soluzione universale».[2] Un motivo per rifiutare l'idea di soluzioni universali risiede nel vivo convincimento che le situazioni umane siano molto varie, e che un buon giudice debba tenere conto di queste variazioni, il che potrebbe voler dire tollerare il rumore, o almeno rifiutare alcune strategie finalizzate a ridurlo.

Quando si diffusero i primi programmi di scacchi al computer, una grande compagnia aerea ne mise uno a disposizione dei passeggeri dei voli

internazionali. Il programma prevedeva vari livelli, e quello più basso adottava una semplice regola: dare scacco al re dell'avversario ogni volta che era possibile. Il computer non era soggetto a rumore: giocava sempre nello stesso modo, ma spesso quell'unica regola che seguiva lo induceva in errore. Si può dire che era un pessimo giocatore di scacchi: anche uno sfidante inesperto era in grado di sconfiggerlo (e sicuramente era questo l'obiettivo, per tenere contenti i passeggeri).

O ancora, consideriamo le politiche penali adottate in alcuni stati americani, in particolare quella del *three strikes and you're out* ("tre errori e sei fuori").[3] L'idea è che al terzo reato scatta il carcere a vita – punto e basta. Questa politica riduce la variabilità derivante dall'assegnazione casuale del giudice nel processo penale: alcuni dei proponenti erano molto preoccupati per il rumore di livello e la possibilità che alcuni giudici fossero troppo clementi verso i criminali incalliti, e l'eliminazione del rumore è il principio fondamentale alla base di questo criterio.

Ma anche se la dottrina dei tre errori raggiungesse questo obiettivo, è ragionevole obiettare che il prezzo da pagare sarebbe troppo elevato: alcuni di coloro che hanno compiuto tre reati non andrebbero incarcerati a vita. È possibile che si trattasse di crimini non violenti; oppure le loro terribili condizioni di vita potrebbero averli in parte indotti a compierli; forse hanno buone probabilità di affrontare con successo un percorso di riabilitazione. Secondo molti, una condanna all'ergastolo che non tenga conto delle circostanze particolari non solo è troppo severa, ma anche intollerabilmente rigida. Pertanto, il prezzo di questa strategia di riduzione del rumore è troppo elevato.

Analizziamo il caso *Woodson v. North Carolina*,[4] in cui la Corte suprema degli Stati Uniti giudicò incostituzionale la pena di morte obbligatoria non perché fosse troppo brutale, ma *perché era una regola*. Il senso della pena di

morte obbligatoria stava proprio nel suo porsi come una garanzia contro il rumore, stabilendo che, in determinate circostanze, tutti gli assassini avrebbero dovuto essere condannati a morte. Invocando la necessità di un trattamento personalizzato, la Corte concluse che «non prevale più il convincimento che qualsivoglia reato all'interno della medesima categoria legale richieda una pena identica senza tenere conto della vita e delle abitudini pregresse del singolo imputato». Secondo la Corte suprema, un serio elemento di incostituzionalità della pena di morte obbligatoria stava nel «trattare tutte le persone condannate per un reato designato non come esseri umani unici e individuali, ma come membri di una massa anonima e indifferenziata da sottoporre al castigo cieco della pena di morte».

È indubbio che la pena capitale sia un caso particolarmente delicato, ma l'analisi della Corte può essere applicata a molte altre situazioni, gran parte delle quali non è legata alla legge. Gli insegnanti che valutano gli studenti, i medici che valutano i pazienti, i datori di lavoro che valutano i dipendenti, i sottoscrittori che valutano l'entità dei premi assicurativi, gli allenatori che valutano gli atleti: tutte queste persone potrebbero commettere errori se applicassero regole eccessivamente rigide per ridurre il rumore. Se i datori di lavoro impiegassero semplici regole per valutare, promuovere o sospendere dei dipendenti, tali regole potrebbero eliminare il rumore ma trascurare degli aspetti importanti delle prestazioni dei singoli lavoratori. Un sistema a punti privo di rumore ma che non tenga conto di variabili significative potrebbe essere peggiore del ricorso a giudizi individuali, per quanto soggetti a rumore.

Il capitolo 27 considererà l'idea generale di trattare le persone come «uniche e individuali» piuttosto che come «membri di una massa anonima e indifferenziata». Per il momento, tuttavia, ci concentriamo su un aspetto più prosaico: alcune strategie di riduzione del rumore comportano troppi

errori, in maniera non troppo dissimile dall'insulso programma di scacchi menzionato in precedenza.

Eppure l'obiezione sembra molto più convincente di quanto in realtà non sia. Se una strategia di riduzione del rumore è soggetta a errori, non dovremmo rassegnarci a un livello di rumore elevato, ma piuttosto cercare di pensare a una strategia migliore, per esempio aggregando i giudizi invece di adottare regole sciocche, oppure elaborare linee guida sensate anziché stupide. Ai fini della riduzione del rumore un'università potrebbe stabilire, per esempio, che verranno ammessi i candidati con il punteggio più alto agli esami d'ammissione, senza considerare nessun altro parametro. Se questa regola sembra troppo drastica, l'ateneo potrebbe elaborare una formula che tenga conto del punteggio di ammissione, dei voti, dell'età, dei titoli sportivi, della situazione famigliare e altro. Queste regole complesse potrebbero essere più accurate, cioè più in sintonia con la vasta gamma di fattori rilevanti, come quelle adottate dai medici per diagnosticare alcune malattie. Le linee guida e le regole impiegate dai professionisti non sono sempre semplici o drastiche, e in molti casi contribuiscono a ridurre il rumore senza incorrere in costi (o bias) intollerabili. E se non dovessero funzionare, forse si potrebbero introdurre altre forme di igiene decisionale più adatte a quella situazione particolare; pensiamo all'aggregazione dei giudizi o all'impiego di un processo strutturato come il protocollo a valutazioni intermedie.

Algoritmi privi di rumore ma affetti da bias

Spesso si parla dei costi potenzialmente alti della riduzione del rumore a proposito degli algoritmi, e le obiezioni ai "bias algoritmici" sono sempre più frequenti. Come abbiamo visto, poiché eliminano il rumore, spesso

questi strumenti sembrano una soluzione auspicabile, e in effetti gran parte di questo libro potrebbe essere inteso come un'argomentazione a sostegno di un maggiore ricorso agli algoritmi. Ma abbiamo anche osservato che i costi della riduzione del rumore diventano intollerabili se questo maggiore ricorso all'informatica porta a un aumento delle discriminazioni di razza e di genere, o nei confronti di chi appartiene a gruppi svantaggiati.

Vi è un timore diffuso che gli algoritmi abbiano effetti discriminatori, e indubbiamente ciò rappresenta un rischio serio. In *Armi di distruzione matematica*, Cathy O'Neil avverte che il ricorso ai big data e a decisioni prese dagli algoritmi potrebbe introdurre pregiudizi, accrescere le disuguaglianze e costituire una minaccia per la democrazia stessa.[5] Un altro parere scettico sottolinea che «i modelli matematici potenzialmente affetti da bias stanno trasformando le nostre vite, e né le società responsabili del loro sviluppo né il governo hanno interesse a risolvere il problema».[6] Secondo ProPublica, un'organizzazione di giornalismo investigativo indipendente, l'algoritmo COMPAS, che ha largo impiego nelle valutazioni sul rischio di recidive, è soggetto a un forte bias contro le minoranze razziali.[7]

Nessuno dubita che sia possibile, e perfino facile, creare un algoritmo privo di rumore ma al contempo razzista, sessista o altro. Uno strumento che usasse esplicitamente il colore della pelle di un imputato per stabilire se a quella persona debba essere concessa la libertà provvisoria sarebbe discriminatorio (e in molti stati il suo impiego sarebbe considerato illegale), così come uno che tenesse conto della possibilità che chi si candida a una certa posizione lavorativa possa avere una gravidanza discriminerebbe le donne. In questi e altri casi, gli algoritmi potrebbero eliminare la variabilità indesiderata nei giudizi ma anche introdurre bias inaccettabili.

In linea di principio, dovremmo essere in grado di progettare un algoritmo che *non* tenga conto di razza e genere, e in effetti ciò è assolutamente possibile. Il problema più stringente, però, a cui oggi si pone molta attenzione, è che un algoritmo potrebbe essere discriminatorio e, in questo senso, soggetto a bias, anche se non utilizza dichiaratamente la razza e il genere come predittori.

Come abbiamo già indicato, potrebbe essere affetto da bias per due motivi principali. Primo, di proposito o meno potrebbe usare predittori altamente correlati alla razza o al genere: l'altezza e il peso, per esempio, sono correlati al genere, e il luogo in cui qualcuno è cresciuto o in cui vive potrebbe esserlo alla razza.

Secondo, la discriminazione potrebbe anche derivare dai dati d'origine: se un algoritmo viene allenato con un set di dati soggetto a bias, lo sarà anch'esso. Pensiamo per esempio agli algoritmi di "polizia predittiva", che spesso cercano di prevedere i reati per migliorare l'assegnazione delle risorse della polizia.[8] Se i dati esistenti sui reati riflettono l'eccessiva vigilanza di certi quartieri o il relativo eccesso di denunce per certi tipi di reati, i risultanti algoritmi perpetueranno o esaspereranno la discriminazione. Ogni volta che i dati di allenamento sono affetti da bias, è ben possibile che venga elaborato, di proposito o meno, un algoritmo discriminatorio; ne consegue che anche se un algoritmo non considera espressamente razza o genere, potrebbe rivelarsi non meno affetto da bias degli esseri umani. Anzi, in questo senso gli algoritmi potrebbero essere anche peggiori: poiché eliminano il rumore, potrebbero essere più *sistematicamente* soggetti a bias rispetto ai giudici umani.[9]

Secondo molti, andrebbe fatta una considerazione pratica fondamentale sull'eventuale difformità dell'impatto di un algoritmo su gruppi identificabili. Come verificare con esattezza questa difformità e come

decidere cosa rappresenti una discriminazione, un bias o un trattamento equo per un algoritmo sono tutti temi molto complessi, che esulano dalle finalità di questo libro.[10]

Il solo fatto che questo interrogativo possa essere sollevato, tuttavia, costituisce un netto vantaggio degli algoritmi rispetto ai giudizi umani. Per i meno pratici, consigliamo un'attenta valutazione degli algoritmi per assicurarsi che non tengano conto di input inammissibili e per verificare se operino discriminazioni spiacevoli. Molto più difficile sarebbe sottoporre a un esame altrettanto minuzioso degli esseri umani, i cui giudizi sono spesso opachi; talvolta le persone operano discriminazioni senza volerlo, in un modo che gli osservatori esterni, compreso il sistema legale, non sono in grado di identificare facilmente. In un certo senso, quindi, un algoritmo può essere più trasparente di un essere umano.

È indubbio che occorra sottolineare i costi di algoritmi privi di rumore ma affetti da bias, così come quelli di regole prive di rumore ma affette da bias, ma la questione centrale è se sia possibile progettare algoritmi migliori dei giudici "reali" sulla base di una serie di criteri rilevanti: accuratezza, riduzione del rumore, imparzialità ed equità. Molti dati indicano che gli strumenti informatici sono in grado di dare risultati migliori degli esseri umani, a prescindere dai criteri selezionati. (Attenzione, stiamo dicendo che ne sono capaci, non che lo faranno.) Per esempio, come si diceva nel capitolo 10, un algoritmo può essere più accurato dei giudici umani nelle decisioni sulla libertà provvisoria, pur dando risultati meno macchiati dalla discriminazione razziale rispetto ai giudizi di persone in carne e ossa. Allo stesso modo, un algoritmo di selezione dei curricula potrà individuare un bacino di candidati più competenti *e più diversificati* di quanto farebbero i valutatori umani.

Questi esempi e molti altri ci portano a una conclusione incontrovertibile: benché in un mondo caratterizzato dall'incertezza un algoritmo predittivo sarà difficilmente perfetto, potrà essere molto meno imperfetto del giudizio umano, sempre soggetto a rumore e spesso anche a bias. Questa superiorità va intesa sia in termini di validità che di discriminazione (ovvero, dei buoni algoritmi quasi sempre danno previsioni migliori e possono essere meno affetti da bias dei giudici umani). Se gli strumenti informatici sbagliano meno degli esperti umani e, malgrado ciò, tendiamo intuitivamente a preferire questi ultimi, allora le nostre preferenze intuitive meriterebbero un attento esame.

Trarremo quindi delle conclusioni generali semplici e non circoscritte al tema degli algoritmi: è vero che le strategie di riduzione del rumore possono essere costose, ma quasi sempre questa è solo una scusa e non una motivazione sufficiente per tollerare l'ingiustizia e i costi anche maggiori dovuti al rumore. Naturalmente gli sforzi per ridurre il rumore potrebbero a loro volta introdurre degli errori, magari sotto forma di bias; se così fosse avremmo un serio problema, ma la soluzione non sta nel rinunciare alle azioni di riduzione del rumore, quanto piuttosto nel trovarne di migliori.

A proposito dei costi della riduzione del rumore

«Per cercare di eliminare il rumore nel campo dell'istruzione occorrerebbe una grossa spesa. Nel valutare gli studenti, gli insegnanti sono soggetti a rumore, ma non possiamo chiedere a cinque insegnanti diversi di valutare lo stesso elaborato.»

«Se, invece di affidarsi al giudizio umano, un social network dovesse decidere che nessuno può usare determinate parole, a prescindere dal contesto, eliminerà il rumore, ma commetterà anche molti errori. La cura potrebbe essere peggiore della malattia.»

«È vero, ci sono regole e algoritmi affetti da bias, ma le persone non sono da meno. Dovremmo chiederci se sia possibile progettare algoritmi privi di rumore e al contempo meno soggetti a bias.»

«Rimuovere il rumore potrebbe avere un costo, ma spesso vale la pena sostenerlo. Il rumore può portare a terribili ingiustizie, e se un'azione volta a ridurlo si rivela troppo grossolana – se per esempio ci ritroviamo con linee guida eccessivamente rigide o che, senza volerlo, producono bias –, invece di abbandonare l'impresa dovremmo semplicemente riprovare.»

[1] A.O. Hirschman, *Retoriche dell'intransigenza: perversità, futilità, messa a repentaglio*, il Mulino, Bologna 2017.

[2] K. Stith, J.A. Cabranes, *Fear of Judging*, cit.

[3] Vedi, per esempio, Three Strikes Basics, Stanford Law School, [https://stanford.io/3BcTzQT].

[4] 428 U.S. 280(1976).

[5] C. O'Neil, *Armi di distruzione matematica. Come i big data aumentano la disuguaglianza e minacciano la democrazia*, Bompiani, Milano 2017.

[6] W. Knight, *Biased Algorithms Are Everywhere, and No One Seems to Care*, in "MIT Technology Review", 12 luglio 2017.

[7] J. Larson *et al.*, *How We Analyzed the COMPAS Recidivism Algorithm*, "ProPublica", 23 maggio, 2016, [https://bit.ly/2TgorPx]. Il riferimento al bias in quest'esempio è problematico, in quanto diverse definizioni di bias potrebbero condurre a conclusioni opposte. Per altre opinioni su questo caso e, più in generale, sulla definizione e la misurazione del bias algoritmico, rimandiamo alla successiva nota 10.

[8] A. Shapiro, *Reform Predictive Policing*, in "Nature", 541(2017), n. 7638, pp. 458-460.

[9] Sebbene questo timore si sia riaffacciato a proposito dei modelli basati sull'intelligenza artificiale, non è limitato a quest'ultima. Già nel 1972 Paul Slovic osservò che costruire dei modelli delle intuizioni avrebbe preservato e rafforzato, e forse perfino amplificato, i bias cognitivi esistenti. P. Slovic, *Psychological Study of Human Judgment: Implications for Investment Decision Making*, in "Journal of Finance", 27(1972), p. 779.

[10] Per un'introduzione su questo dibattito nel contesto della controversia sull'algoritmo di previsione della recidività COMPAS, vedi J. Larson *et al.*, *COMPAS Recidivism Algorithm*, cit.; W. Dieterich *et al.*, *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*, Northpointe, Inc., 8 luglio 2016, [go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf]; J. Dressel, H. Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, in "Science Advances", 4(2018), n. 1, pp. 1-6; S. Corbett-Davies *et al.*, *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear*, in "Washington Post", 17 ottobre 2016, [https://wapo.st/3xKwAL0]; A. Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, in "Big Data", 153(2017), p. 5; J. Kleinberg, S. Mullainathan, M. Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, Leibniz International Proceedings in Informatics, gennaio 2017.

27

Dignità

Poniamo che vi abbiano negato un mutuo non perché qualcuno abbia studiato la vostra situazione, ma perché una banca ha una regola fissa per cui alle persone con la vostra affidabilità creditizia semplicemente un mutuo non viene concesso. Oppure, supponete di avere titoli eccellenti e di aver fatto un'ottima impressione su un selezionatore di una certa azienda, ma la vostra candidatura è stata rifiutata perché quindici anni fa siete stati condannati per un reato minore legato alla droga, e l'azienda ha posto il divieto assoluto di assumere chiunque sia mai incorso in una condanna. O magari siete accusati di un reato e vi viene negata la libertà provvisoria non dopo un'udienza individuale alla presenza di un essere umano, ma perché un algoritmo ha stabilito che le persone con le vostre caratteristiche hanno un rischio di fuga che supera la soglia per la concessione della libertà provvisoria.

In questi casi molti avrebbero da obiettare: vorranno essere trattati come persone, vorranno che le loro circostanze particolari vengano prese in considerazione da un vero essere umano. Forse sapranno anche che un trattamento individuale produrrebbe rumore, ma se è questo il prezzo da pagare, insisteranno che ne vale la pena. Potrebbero avere da ridire ogni volta che le persone vengono trattate, per usare le parole della Corte suprema americana, «non come esseri umani unici e individuali, ma come membri di una massa anonima e indifferenziata da sottoporre al castigo cieco» di una qualche pena (vedi capitolo 26).

Molti insistono sulle udienze individuali, prive di quella che considerano la tirannia delle regole, per dare alle persone l'impressione di essere trattate come individui, quindi con una qualche forma di rispetto. L'idea di un regolare processo, considerata la norma, sembrerebbe richiedere la possibilità di un'interazione faccia a faccia in cui un essere umano, autorizzato a esercitare la propria discrezionalità, tenga in considerazione un'ampia gamma di fattori.

In molte culture questo argomento a favore del giudizio "caso per caso" ha fondamenti morali molto radicati, e si può ritrovare nel pensiero politico, giuridico e teologico, ma anche nelle opere letterarie. *Il mercante di Venezia* di Shakespeare potrebbe essere letto come un'obiezione a regole prive di rumore e una difesa del ruolo della misericordia nella legge e nel giudizio umano in generale. Ecco l'argomentazione conclusiva di Porzia:

> La natura della misericordia non si può forzare,
> Cade come la pioggia gentile dal cielo
> Sulla terra in basso: è due volte benedetta,
> Benedice colui che la esercita e colui
> Che la riceve:
> [...]
> Essa ha il suo trono
> Nei cuori dei re, è un attributo
> Dello stesso Dio; e il potere terreno
> Appare più simile a quello di Dio
> Quando la misericordia tempera la giustizia.

Non essendo vincolata da regole, la misericordia è soggetta a rumore. Nondimeno, l'appello di Porzia può essere avanzato in molte situazioni e in svariate organizzazioni, e spesso ha un forte impatto. Un dipendente potrebbe usarlo per chiedere una promozione; un aspirante proprietario di un immobile per fare una richiesta di prestito; uno studente per candidarsi all'ammissione a un corso di laurea. Chi ha il compito di prendere una

decisione in questi casi potrebbe rifiutare alcune strategie di riduzione del rumore, soprattutto le regole inflessibili, forse perché crede, come Porzia, che la misericordia non si possa forzare. Forse sapranno che il proprio approccio è soggetto a rumore, ma se permetterà alle persone di sentirsi trattate con rispetto e ascoltate da qualcuno, potranno adottarlo ugualmente.[1]

Alcune strategie di riduzione del rumore non incorrono in questa obiezione: se per esempio una decisione viene presa da tre persone invece che da una sola, vi è comunque una forma di udienza individuale per chi viene interessato da tale giudizio; le linee guida, poi, possono lasciare notevole discrezionalità ai decisori. Tuttavia, altre azioni per ridurre il rumore, regole fisse comprese, eliminano tale discrezionalità, e le persone potranno obiettare che costituiscano un'offesa alla propria dignità.

Hanno ragione loro? Certo, per qualcuno spesso è importante ricevere un'udienza individuale, e vi è un indubbio valore umano nell'opportunità di essere ascoltati. Ma se le udienze individuali producono più morti, più ingiustizie e costi molto più elevati, non andrebbero elogiate. Abbiamo sottolineato che in situazioni come selezione del personale, ammissioni scolastiche e diagnosi mediche alcune strategie di riduzione del rumore potrebbero rivelarsi grossolane, arrivando a proibire trattamenti individuali che, seppur soggetti a rumore, nel complesso produrrebbero meno errori, ma se una strategia di riduzione del rumore è mediocre la cosa migliore da fare – come non ci stancheremo di suggerire – è cercare di immaginarne una migliore, che tenga conto di un'ampia gamma di variabili pertinenti. Se tale strategia cancellerà il rumore e produrrà meno errori, avrà ovvi vantaggi rispetto al trattamento individuale, anche se ridurrà o eliminerà l'opportunità di essere ascoltati.

Non stiamo dicendo che il trattamento individuale non conti. Ma il prezzo da pagare è troppo alto, se tale trattamento comporta conseguenze terribili, compresa una palese ingiustizia.

Valori che cambiano

Immaginate che un'istituzione pubblica riesca a eliminare il rumore, per esempio che un'università specifichi cosa intende per *comportamenti illeciti*, in modo che tanto i docenti quanto gli studenti sappiano riconoscerli in maniera inequivocabile. Oppure supponiamo che una grande azienda precisi il significato esatto che dà alla parola *corruzione*, in modo che chiunque ci lavori sappia cosa è permesso fare e cosa no. O ancora poniamo che un'istituzione privata riduca sensibilmente il rumore, magari dichiarando che assumerà solo chi ha una laurea in determinate discipline. Ma cosa accadrebbe nel momento in cui in una di queste organizzazioni i valori cambiassero? Alcune strategie di riduzione del rumore sembrerebbero inadatte ad accoglierne di nuovi, e questa inflessibilità potrebbe essere un problema, peraltro strettamente connesso all'interesse per il trattamento individuale e la dignità.

Ci aiuterà a fare il punto una decisione notoriamente controversa della Corte suprema statunitense, che ha sollevato questioni di diritto costituzionale.[2] Il caso, giudicato nel 1974, riguardava una regola fissa di un sistema scolastico che obbligava ogni insegnante incinta a prendere un'aspettativa non retribuita per i cinque mesi precedenti la data prevista del parto. Jo Carol LaFleur si oppose, sostenendo di essere perfettamente in grado di insegnare, e definendo quella regola discriminatoria e i cinque mesi una misura eccessiva.

La Corte suprema si pronunciò a suo favore, ma non parlò di discriminazione sessuale e non disse che il periodo di aspettativa fosse necessariamente esagerato. Obiettò invece che a LaFleur non era stata data l'opportunità di dimostrare che, nel suo caso particolare, non vi era la necessità fisica di smettere di lavorare. La Corte affermò che «non è stata presa una decisione individuale da parte di un medico o del consiglio scolastico, in merito alla capacità di una particolare insegnante di proseguire la propria attività professionale. Le regole contengono una presunzione inconfutabile di indisposizione fisica, e tale presunzione trova applicazione anche quando è possibile che l'evidenza medica circa lo stato fisico di una specifica donna indichi l'esatto opposto».

Un'aspettativa obbligatoria di cinque mesi sembra effettivamente assurda, ma non è questo il punto su cui ha insistito la Corte. Al contrario, ha avuto da obiettare sulla «presunzione inconfutabile» e sull'assenza di una «decisione individuale», dando l'impressione di sostenere, al pari di Porzia, che la misericordia non si possa forzare e che le circostanze particolari di LaFleur andassero esaminate da una particolare persona.

Ma, in assenza di una procedura di igiene decisionale, così facendo si apre la strada al rumore. Chi decide sul caso di LaFleur? La decisione sarà la stessa per lei come per molte altre donne che si trovano in una situazione simile? In ogni caso, sono tante le regole che «contengono una presunzione inconfutabile»: dobbiamo quindi concludere che stabilire un preciso limite di velocità sia inaccettabile? E l'età minima per votare o consumare alcolici? Il divieto assoluto di guidare in stato di ebbrezza? Tenendo conto di questi esempi, i critici hanno obiettato che un argomento contro le «presunzioni inconfutabili» si dimostrerebbe eccessivo, considerando che la loro finalità e il loro effetto è la riduzione del rumore.

All'epoca, autorevoli commentatori difesero la decisione della Corte ponendo l'accento sul fatto che i valori morali cambiano nel tempo, quindi occorre evitare regole rigide.[3] Costoro sostenevano che le norme sociali circa il ruolo della donna nella società erano in una fase di grande instabilità, e che le decisioni individualizzate fossero particolarmente adatte in tale contesto, perché avrebbero permesso di incorporare queste norme mutevoli. Un sistema vincolato da regole potrebbe eliminare il rumore, il che è un bene, ma anche cristallizzare le norme e i valori esistenti, il che non lo è affatto.

Insomma, alcuni insisteranno che il vantaggio di un sistema affetto da rumore è che permetterà di accogliere i nuovi valori emergenti. Man mano che i valori cambiano, e se ai giudici verrà permesso di esercitare la propria discrezionalità, potrebbero per esempio cominciare a emettere condanne più miti per reati di droga o più alte per stupro. Abbiamo già osservato che, se alcuni giudici sono clementi e altri no, vi sarà un certo livello di ingiustizia: persone che si trovano nella stessa situazione verranno trattate in maniera diversa. Ma questa ingiustizia potrebbe essere tollerata se lascia spazio a valori sociali nuovi o emergenti.

Il problema non è affatto limitato al sistema di giustizia penale o al diritto. In ottemperanza a varie politiche interne, le aziende potrebbero decidere di consentire una certa flessibilità nei giudizi e nelle decisioni, anche se questo crea rumore, perché la flessibilità garantisce che, con l'affermarsi di nuove credenze e valori, nel tempo si possa cambiare politica. Ecco un esempio personale: quando, alcuni anni fa, uno degli autori di questo libro è entrato in una grande società di consulenza, nel pacchetto di benvenuto non proprio recente che gli è stato inviato si specificavano le spese di viaggio per le quali avrebbe potuto richiedere un rimborso («una telefonata alla famiglia per confermare il proprio arrivo; le

spese di stiratura della camicia; le mance per i facchini dell'albergo»). Le regole erano prive di rumore, ma obsolete (e sessiste), e ben presto vennero sostituite da standard modificabili nel tempo. Le spese, per esempio, ora devono essere semplicemente «consone e ragionevoli».

La prima risposta a questa difesa del rumore è semplice: alcune strategie di riduzione del rumore non sono per nulla toccate da questa obiezione. Se si utilizza una scala condivisa fondata su una visione esterna, per esempio, è possibile reagire al cambiamento dei valori nel tempo. In ogni caso, non è detto, né tantomeno auspicabile, che le azioni volte alla riduzione del rumore debbano essere permanenti. Se tali interventi assumono la forma di regole rigide, chi li elabora dovrebbe essere disposto a modificarli nel tempo: potrebbe rivederli con cadenza annuale, oppure decidere che ogni volta che si presenteranno nuovi valori, serviranno nuove regole. Nel sistema penale i legislatori potrebbero ridurre le condanne per certi reati e aumentarle per altri, o depenalizzare del tutto alcune attività e criminalizzarne una che in precedenza era ritenuta del tutto accettabile.

Ma facciamo un passo indietro. I sistemi soggetti a rumore possono accogliere valori morali emergenti, e questo può essere un bene, ma in molti ambiti è assurdo difendere alti livelli di rumore sulla base di tale argomentazione. Come abbiamo evidenziato, alcune delle più importanti strategie di riduzione del rumore, per esempio l'aggregazione dei giudizi, tengono conto dell'emergenza di nuovi valori. E se diversi clienti che sporgono un reclamo per il malfunzionamento di un computer vengono trattati in maniera diversa da una società produttrice, è difficile che questa incoerenza sia dovuta all'emergenza di nuovi valori. Se persone diverse ricevono diagnosi mediche diverse, raramente ciò sarà imputabile a nuovi valori morali. Si può fare molto per ridurre o eliminare il rumore pur

continuando a prevedere procedure che consentano un'evoluzione dei valori.

Manipolare il sistema, eludere le regole

In un sistema soggetto a rumore, chi giudica può adattarsi alla situazione e reagire di fronte a sviluppi inattesi; se si elimina questa capacità di adattamento, alcune strategie di riduzione del rumore possono produrre l'effetto indesiderato di incentivare le persone a manipolare il sistema. Una possibile ragione per tollerare il rumore è che potrebbe rivelarsi un effetto collaterale di approcci adottati da istituzioni pubbliche e private per prevenire questo tipo di manipolazione.

Prendiamo per esempio il sistema fiscale. Da una parte, non dovrebbe essere affetto da rumore, ma, al contrario, chiaro e prevedibile: contribuenti identici non dovrebbero subire un trattamento diverso. Ma se eliminassimo il rumore in quest'ambito, i contribuenti furbi troverebbero inevitabilmente il modo di eludere le regole. Tra i fiscalisti vi è un acceso dibattito su quanto sia auspicabile da una parte avere regole che eliminino il rumore, dall'altra contemplare un certo livello di vaghezza che ammetta l'imprevedibilità ma riduca anche il rischio che delle regole chiare e trasparenti portino a comportamenti opportunistici o egoisti.

Alcune società e università proibiscono al proprio personale di commettere "illeciti" senza specificare a cosa ci si riferisca con questo termine, creando inevitabilmente rumore. D'altro canto, se si stabilisce una casistica completa di comportamenti ritenuti illeciti, si finirà per tollerare anche quelli deprecabili ma non esplicitamente menzionati in tale elenco.

Poiché le regole hanno confini precisi, è possibile eluderle assumendo una condotta tecnicamente non sanzionabile, ma che crea danni identici o

analoghi (come sa bene ogni genitore!). Quando non si riescono a elaborare facilmente delle regole che vietino ogni condotta che andrebbe proibita, esiste una precisa motivazione per tollerare il rumore, o almeno questa è l'obiezione che normalmente viene posta.

In certe circostanze, regole chiare e definite per eliminare il rumore effettivamente possono causare questo rischio di elusione, il che potrebbe essere un buon motivo per adottarne di altre, come l'aggregazione, e forse per tollerare un approccio che ammetta un certo grado di rumore. Potrebbe, ma non sempre. Dobbiamo chiederci quanta elusione – e quanto rumore – si rischia: se la risposta è poco dell'una e tanto dell'altro, allora sarebbe meglio optare per approcci che riducano il rumore. Torneremo su questo tema nel capitolo 28.

Deterrenza e avversione al rischio

Poniamo che l'obiettivo sia scoraggiare comportamenti illeciti da parte di dipendenti, studenti o comuni cittadini. In questo caso, un margine più o meno alto di imprevedibilità potrebbe non essere la cosa peggiore. Un datore di lavoro potrebbe pensare: «Se la sanzione per certi illeciti può essere una semplice ammenda, una sospensione o il licenziamento, i miei dipendenti non commetteranno tali illeciti». Chi amministra il sistema penale potrebbe arrivare alle stesse conclusioni: «Non ci interessa tanto se gli aspiranti criminali non sappiano quale pena li aspetti. Se la prospettiva di una lotteria della pena costituisce un deterrente alle irregolarità, forse il rumore che ne risulta si può tollerare».

In astratto, queste argomentazioni non possono essere respinte, ma non sono neanche tanto convincenti. Sulle prime, ciò che conta è il valore atteso della pena: il 50% di possibilità di prendere una multa da cinquemila

dollari equivale alla certezza di prenderne una da duemilacinquecento. Naturalmente alcuni potranno concentrarsi sullo scenario potenzialmente peggiore: le persone avverse al rischio saranno più scoraggiate dal 50% di possibilità di prendere una multa da cinquemila dollari; tuttavia, quelle propense al rischio lo saranno molto meno. Per sapere se un sistema soggetto a rumore abbia un maggiore effetto deterrente, dovremmo sapere se i potenziali trasgressori sono avversi o propensi al rischio. Ma per aumentare la deterrenza, non sarebbe meglio aumentare la pena ed eliminare il rumore? In questo modo, si eliminerebbe anche l'iniquità.

Creatività, morale e idee nuove

È possibile che alcune misure di riduzione del rumore scoraggino la motivazione e l'impegno? Che incidano sulla creatività e impediscano alle persone di fare grandi progressi? Molte organizzazioni ritengono di sì, e in certi casi potrebbero anche avere ragione. Per scoprirlo, occorre specificare la strategia di riduzione del rumore a cui vengono rivolte queste obiezioni.

Ricordiamo ancora una volta la reazione fortemente negativa di molti giudici alle linee guida sulle condanne.[4] Come affermò un giudice, «dobbiamo tornare a fidarci dell'esercizio della giustizia nelle aule di tribunale». In generale, chi riveste una posizione di autorità non gradisce che gli venga sottratta la propria discrezionalità: potrà sentirsi sminuito oltre che limitato, e perfino umiliato. Quando si prendono delle misure per ridurre la discrezionalità, ci si scontra con le obiezioni di coloro che attribuiscono un grande valore alla propria facoltà di esercitare il giudizio, difendendola a spada tratta. Nel momento in cui vengono privati della loro

discrezionalità e si ritrovano a fare quello che fa chiunque altro, potrebbero sentirsi come le rotelle di un ingranaggio.

Insomma, un sistema affetto da rumore potrebbe avere effetti positivi sul morale non perché sia migliore per principio, ma perché permette ad alcune persone di decidere come ritengono opportuno. Se i dipendenti possono rispondere ai reclami dei clienti a modo loro, valutare i propri sottoposti nel modo che preferiscono o stabilire dei premi assicurativi come pensano sia più appropriato, forse apprezzeranno di più il proprio lavoro. Ma se l'azienda mette in campo delle misure per eliminare il rumore, i dipendenti potrebbero pensare che ciò comprometterà il loro potere decisionale: ora dovranno seguire delle regole invece di esercitare la propria creatività, e il lavoro sembrerà più meccanico, perfino robotico. Chi vorrebbe lavorare in un contesto che limita o annulla la possibilità di prendere decisioni indipendenti?

Le organizzazioni potrebbero mostrarsi sensibili a questo sentimento non solo per rispetto ai propri dipendenti, ma anche per dare loro spazio per concepire nuove idee; una volta stabilita una regola, è possibile che questa riduca l'ingegno e l'inventiva.

Queste obiezioni riguardano molte persone che operano all'interno di organizzazioni, ma, naturalmente, non tutte. Compiti diversi vanno valutati in modo diverso: non è detto che le diagnosi di faringite o di ipertensione affette da rumore richiedano una grande creatività. Ma potremmo essere disposti a tollerare il rumore se porterà a dipendenti più felici e ispirati; la demoralizzazione, infatti, è di per sé un costo, e comporta ulteriori costi in termini di prestazioni mediocri. È certamente possibile ridurre il rumore e, insieme, restare aperti a nuove idee: alcune strategie, come la strutturazione di giudizi complessi, lo consentono. Se vogliamo diminuire il rumore e allo stesso tempo tenere alto il morale, potremmo

scegliere strategie di igiene decisionale che garantiscano questo effetto, e i responsabili potranno chiarire che anche nel caso in cui vengano adottate regole rigide ci sarà sempre la possibilità di contestarle e ripensarle, a patto di non violarle esercitando la propria discrezionalità da caso a caso.

Philip Howard, illustre avvocato e intellettuale, porta avanti in una serie di libri un'appassionata arringa a favore dell'apertura a giudizi più flessibili, auspicando politiche che non assumano la forma di regole prescrittive che eliminano il rumore, ma di princìpi generali: «siate ragionevoli», «agite con prudenza», «non esponetevi a rischi eccessivi».[5]

Secondo Howard, i regolamenti governativi di oggi sono folli, perché troppo rigidi: insegnanti, agricoltori, imprenditori, infermieri, medici e professionisti di molte altre categorie, sono tutti oberati da regole che dicono esattamente cosa fare e in che modo farlo, mentre sarebbe molto meglio permettere alle persone di usare la propria creatività per farsi un'idea di come raggiungere i propri obiettivi, che si tratti di risultati formativi migliori, di una riduzione degli incidenti, di acque più pulite o di pazienti più sani.

Alcune delle argomentazioni portate da Howard sono affascinanti, ma è importante chiedersi quali siano le conseguenze degli approcci da lui promossi, non ultimo il potenziale aumento del rumore e del bias. Quasi nessuno apprezza la rigidità in termini astratti, ma potrebbe essere il modo migliore per ridurre il rumore ed eliminare bias ed errori. Se si introducono solo dei princìpi generali, la loro interpretazione e applicazione sarà intaccata dal rumore, a un livello che potrebbe risultare intollerabile e perfino vergognoso. Occorrerebbe quantomeno considerare con attenzione i costi del rumore, cosa che di solito non viene fatta. Una volta osservato che il rumore produce iniquità diffuse e alti costi, spesso ne

concluderemo che è inaccettabile, e che quindi occorre identificare strategie per ridurlo che non compromettano i valori importanti.

A proposito di dignità

«Le persone apprezzano le interazioni faccia a faccia, e spesso ne hanno proprio bisogno: vogliono che un essere umano ascolti i loro problemi e le loro rimostranze, e abbia il potere di intervenire per migliorare le cose. Certo, queste interazioni inevitabilmente produrranno rumore, ma la dignità umana è preziosa.»

«I valori morali sono in continua evoluzione; se blindiamo tutto, non lasciamo spazio per i valori nuovi. Alcune azioni per ridurre il rumore sono troppo rigide, in quanto impedirebbero il cambiamento morale.»

«Come deterrente agli illeciti, un certo grado di rumore andrebbe tollerato. Se gli studenti non sanno quale sarà la sanzione per il plagio, è un bene, perché si asterranno dal copiare. Un po' di incertezza sotto forma di rumore può aumentare la deterrenza.»

«Se eliminiamo il rumore, potremmo finire per avere regole chiare che i malfattori troveranno il modo di eludere. Il rumore può essere un prezzo che vale la pena pagare se è un modo per evitare comportamenti strategici o opportunistici.»

«Le persone creative hanno bisogno di spazio. Donne e uomini non sono dei robot: qualunque sia il vostro lavoro, meritate un certo spazio di manovra, e se vi sentite ingabbiati, forse non sarete affetti da rumore, ma non prenderete gusto in ciò che state facendo e non riuscirete a esprimere la vostra originalità.»

«In fin dei conti, molte delle argomentazioni a difesa del rumore non convincono: possiamo rispettare la dignità altrui e lasciare molto spazio al progresso della morale e alla creatività umana senza per questo tollerare l'iniquità e il costo del rumore.»

[1] T.R. Tyler, *Why People Obey the Law*, 2ª ed., Princeton University Press, Princeton, NJ 2020.

[2] *Cleveland Bd. of Educ. v. LaFleur*, 414 U.S. 632(1974).

[3] L.H. Tribe, *Structural Due Process*, in "Harvard Civil Rights-Civil Liberties Law Review", 10(1975), n. 2, p. 269.

[4] K. Stith, J.A. Cabranes, *Fear of Judging*, cit., p. 177.

[5] Vedi, per esempio, P.K. Howard, *The Death of Common Sense: How Law Is Suffocating America*, Random House, New York 1995; Id., *Try Common Sense: Replacing the Failed Ideologies of Right and Left*, W.W. Norton & Company, New York 2019.

28

Regole o standard?

Se l'obiettivo è ridurre il rumore, e occorre decidere come, se e fino a che punto perseguirlo, è utile distinguere tra due possibili modalità per disciplinare i comportamenti: le regole e gli standard. Spesso le organizzazioni scelgono le une o gli altri, o una loro combinazione.

In ambito aziendale, una società potrebbe comunicare ai propri dipendenti di arrivare in ufficio in una determinata fascia oraria, di non prendere ferie superiori alle due settimane e di non rivelare informazioni alla stampa, pena il licenziamento. In alternativa, potrebbe dire ai dipendenti che dovranno presentarsi in ufficio «a un orario ragionevole», che le ferie verranno decise «caso per caso, compatibilmente con le esigenze della società» e che chi rivela informazioni alla stampa «incorrerà nelle dovute sanzioni».

In ambito legale, si potrebbe imporre la regola che non si debba superare un preciso limite di velocità, che gli operai non debbano essere esposti a sostanze cancerogene o che tutti i farmaci debbano riportare avvertenze specifiche. Per contro, uno standard potrebbe imporre agli automobilisti di guidare «con prudenza», ai datori di lavoro di garantire ambienti di lavoro sicuri «per quanto possibile» e alle società di decidere «ragionevolmente» se riportare delle avvertenze nei foglietti illustrativi dei farmaci.

Questi esempi illustrano la distinzione fondamentale tra regole e standard: le regole si prefiggono di eliminare ogni discrezionalità da parte di chi le applica, gli standard di garantirla. Ogni volta che vengono imposte

delle regole, il rumore dovrebbe ridursi in maniera netta. Chi deve interpretarle si trova davanti a una questione di fatto: a che velocità andava il conducente? L'operaio è stato esposto a sostanze cancerogene? Il farmaco riportava le avvertenze richieste?

In presenza di regole, l'accertamento dei fatti può comunque prevedere un giudizio, e quindi produrre rumore o essere affetto da bias, come abbiamo visto in vari esempi, ma chi le elabora punta a ridurre tali rischi, e quando una regola contiene un valore numerico («non è ammesso il voto prima del conseguimento dei diciotto anni», oppure «il limite di velocità è di centotrenta chilometri orari»), il rumore dovrebbe ridursi. Le regole hanno una caratteristica importante: *riducono il ruolo del giudizio*. In questo senso, se non altro, si ridurrà anche il lavoro dei giudici, intesi ovviamente, in senso lato, come tutti coloro che si trovano ad applicare delle regole: dovranno soltanto limitarsi a seguirle. Nel bene e nel male, avranno uno spazio di manovra molto più ristretto.

Quando sono in vigore degli standard, invece, avviene il contrario: i giudici devono lavorare molto per specificare il significato di termini indefiniti. Potrebbero dover emettere numerosi giudizi per stabilire cosa si intende, per esempio, per termini come "ragionevole" e "fattibile", e oltre ad accertare i fatti dovranno dare un contenuto a espressioni con un certo grado di vaghezza. Chi concepisce gli standard demanda a tutti gli effetti l'autorità decisionale ad altri: di fatto, delega il potere.

Le linee guida di cui abbiamo parlato nel capitolo 22 potrebbero essere tanto regole quanto standard; nel primo caso vincolano fortemente il giudizio, ma anche nel secondo non è detto che si tratti di standard indefiniti. I punteggi di Apgar sono linee guida, non regole, in quanto non proibiscono l'esercizio della discrezionalità. Solo quando le linee guida sono talmente stringenti da eliminare tale discrezionalità, si trasformano in

regole. Gli algoritmi, per esempio, funzionano come regole, non come standard.

Divisioni e ignoranza

Dovrebbe essere chiaro sin da subito che in presenza di divisioni nette all'interno di aziende, organizzazioni, società o gruppi, sarà molto più semplice generare standard che regole. Il direttivo di una società potrebbe essere d'accordo sul fatto che gli amministratori non debbano commettere atti illeciti, senza sapere di preciso cosa implichi tale divieto; i supervisori potrebbero opporsi alle molestie sessuali sul posto di lavoro ma non saper dire se flirtare sia accettabile o meno; un'università potrebbe proibire agli studenti di commettere atti di plagio senza specificare il significato esatto di questo termine; i cittadini potrebbero essere d'accordo sul fatto che una costituzione debba tutelare la libertà di parola, ma non essere certi se ciò valga anche per le pubblicità, le minacce o le oscenità; le persone potrebbero convenire sul fatto che le autorità competenti in materia ambientale debbano emettere regole prudenti per ridurre le emissioni di gas serra, senza avere un'idea precisa di ciò che intendono per prudenza.

Fissare degli standard senza specificarne i dettagli potrebbe generare rumore, il quale potrebbe essere tenuto sotto controllo attraverso alcune delle strategie qui illustrate, come l'aggregazione dei giudizi e l'uso del protocollo a valutazioni intermedie. È anche possibile che dei leader vogliano stilare delle regole, ma di fatto non riescano a convenire sulle stesse. Perfino le costituzioni contengono molti standard (per esempio quelli a tutela della libertà di culto), e lo stesso vale per la Dichiarazione universale dei diritti umani («Tutti gli esseri umani nascono liberi ed eguali in dignità e diritti»).

La grande difficoltà nell'indurre persone diverse a convenire su precise regole per la riduzione del rumore favorisce l'adozione di standard più vaghi. Se il direttivo di una società non riesce a concordare su una dicitura specifica per normare i rapporti tra dipendenti e clienti, potrebbe accontentarsi di arrivare a definire degli standard. Vi sono situazioni analoghe anche nel settore pubblico: i legislatori potrebbero pervenire a un compromesso su uno standard, e tollerare il conseguente rumore, se è questo il prezzo da pagare per approvare una legge; i medici potrebbero convenire su alcuni standard per la diagnosi delle malattie, mentre eventuali tentativi di formulare regole potrebbero causare disaccordi insolubili.

Ma le divisioni sociali e politiche non sono l'unico motivo alla base del ricorso agli standard invece che alle regole: talvolta il vero problema è che mancano le informazioni che permetterebbero di pervenire a regole sensate. Un'università potrebbe essere impossibilitata a produrne di valide per decidere sulla promozione di un docente; un datore di lavoro potrebbe fare fatica a prevedere tutte le circostanze che potrebbero indurlo a confermare o punire un dipendente; un parlamento nazionale potrebbe non sapere quale sia il livello ammissibile di agenti inquinanti come polveri sottili, ozono, diossido di azoto, piombo: il massimo che potrà fare sarà emanare standard di qualche tipo e affidarsi a esperti di comprovata credibilità per precisarne il significato, anche se la conseguenza è la creazione di rumore.

Le regole possono essere affette da vari tipi di bias: per esempio, una regola potrebbe proibire alle donne di diventare agenti di polizia, o vietare agli stranieri di candidarsi per un posto di lavoro. Anche quando creano elevati livelli di bias, le regole ridurranno nettamente il rumore (se tutti le seguono). Se una regola stabilisce che l'acquisto di bevande alcoliche è

consentito esclusivamente alle persone di età superiore ai ventun anni, probabilmente ci sarà poco rumore, a patto che venga rispettata. Gli standard, al contrario, preparano il terreno al rumore.

Capi che controllano i dipendenti

La distinzione tra regole e standard ha grande importanza per tutte le istituzioni pubbliche e private, comprese le aziende di qualsiasi natura. La scelta sorge ogni volta che un superiore cerca di controllare un sottoposto. Come abbiamo visto nel capitolo 2, i sottoscrittori assicurativi lavorano sodo per arrivare a premi ottimali (cioè né troppo alti né troppo bassi) che favoriscano la compagnia; cosa imporranno loro i capi della società, standard o regole? Ogni leader aziendale potrà dare ai propri dipendenti direttive specifiche, oppure indicazioni generiche come: «usate il buonsenso» o «esercitate al meglio il vostro giudizio». Un medico potrebbe adottare l'uno o l'altro approccio nel dare consigli a un paziente: «Prenda una compressa la mattina e una la sera» è una regola; «Prenda una pillola ogni volta che ne sente il bisogno» è uno standard.

Abbiamo già osservato che una società di social media come Facebook si porrà inevitabilmente il problema del rumore e cercherà di risolverlo. Potrà chiedere ai suoi dipendenti di rimuovere un contenuto quando un post viola una precisa regola (poniamo quella che vieta foto di nudo), o in alternativa imporre di adeguarsi a uno standard (per esempio quello che proibisce materiali aggressivi o palesemente offensivi). I cosiddetti "Standard della community", pubblicati da Facebook nel 2018, sono un affascinante miscuglio di regole e standard, e hanno sollevato molte rimostranze da parte degli utenti di Facebook, convinti che producessero troppo rumore (e che quindi creassero sia errori sia ingiustizie). Molti

temevano che, poiché a effettuare questi giudizi erano migliaia di revisori dei contenuti di Facebook, le loro decisioni potessero essere altamente variabili: nel decidere se rimuovere o no un post, i revisori facevano scelte diverse riguardo a ciò che era permesso o vietato. Per capire perché una tale variabilità fosse inevitabile, esaminiamo uno di questi "Standard della community":

> Definiamo l'incitamento all'odio come un attacco diretto rivolto alle persone sulla base di categorie protette, quali razza, etnia, nazionalità di origine, disabilità, religione, casta, orientamento sessuale, genere, identità di genere e malattie gravi. Definiamo l'attacco come discorsi violenti o disumanizzanti, stereotipi nocivi, dichiarazioni di inferiorità, espressioni di disprezzo, disgusto o rifiuto, imprecazioni e incitazioni all'esclusione o alla segregazione. [...] Proteggiamo inoltre rifugiati, migranti, immigrati e richiedenti asilo dagli attacchi più gravi.[1]

Nell'applicare una definizione di questo tipo, i revisori saranno inevitabilmente soggetti a rumore: cosa, esattamente, rientra nei «discorsi violenti o disumanizzanti»? Facebook era consapevole di questo problema e ha reagito virando nella direzione di regole nette e chiare, proprio per ridurre il rumore. Queste regole sono state catalogate in un documento riservato di circa dodicimila parole intitolato *Standard di implementazione*, che "The New Yorker" è riuscito a ottenere.[2] Negli standard pubblici della community, il testo relativo alla pubblicazione di immagini inizia così: «Rimuoviamo i contenuti che promuovono la violenza» (cosa si intende, di preciso?). Per contro, gli standard di implementazione elencano determinate tipologie di immagini e indicano esplicitamente ai moderatori come comportarsi in presenza, per esempio, di «esseri umani carbonizzati o bruciati» e «rimozione di parti del corpo». In sostanza, le indicazioni pubbliche reperibili sul sito di Facebook sono più simili a standard, mentre quelle contenute nel documento riservato somigliano a vere e proprie regole.

Allo stesso modo, una compagnia aerea potrebbe chiedere ai propri piloti di osservare delle regole o degli standard, per esempio riguardo all'eventualità di tornare al gate dopo essere rimasti fermi sulla pista per novanta minuti, o quale sia l'esatto momento in cui accendere la spia dell'obbligo delle cinture di sicurezza. La compagnia potrebbe preferire delle regole perché limitano la discrezionalità del pilota, riducendo in questo modo l'errore, ma potrà anche ritenere che, in certe circostanze, i piloti dovrebbero usare il proprio giudizio. In queste situazioni gli standard potrebbero essere decisamente più adatti delle regole, pur producendo un certo rumore.

In tutti questi casi e molti altri, chi decide tra regole e standard deve portare l'attenzione sul problema del rumore, del bias o di entrambi. Le aziende, grandi o piccole che siano, si trovano costantemente a dover prendere delle decisioni, e talvolta agiscono intuitivamente, senza un quadro di riferimento.

Gli standard possono assumere diverse forme: possono essere privi di contenuto, come «fate ciò che vi sembra appropriato, date le circostanze», o formulati in maniera simile alle regole, come quando si definisce cosa si intende nello specifico per "appropriato", per limitare la discrezionalità dei decisori. Regole e standard possono anche essere combinati e associati: un ufficio del personale potrebbe adottare una regola («Per presentare una candidatura occorre essere laureati») a cui vincolare l'applicazione dello standard («Fatto salvo tale vincolo, la scelta ricadrà sulla persona che potrà dare i risultati migliori nella posizione offerta»).

Abbiamo detto che le regole dovrebbero ridurre e possibilmente eliminare il rumore, e che, al contrario, spesso gli standard ne producono molto (a meno che non si adotti una strategia finalizzata a ridurlo). Nelle organizzazioni pubbliche e private spesso il rumore deriva dalla mancata

definizione di regole. Quando il rumore si fa sentire, cioè quando diventa evidente a tutti che persone in posizioni analoghe non ricevono un pari trattamento, spesso ci si sposta in direzione delle regole, ma, come abbiamo visto nel caso delle condanne penali, questo spostamento potrebbe trasformarsi in una protesta, che in genere è preceduta da una forma di controllo del rumore.

Il ritorno del rimosso

Prendiamo in considerazione una questione importante: chi va ritenuto disabile, e quindi idoneo a ricevere indennità e benefici riservati alle persone impossibilitate a lavorare? Se la domanda viene formulata in questo modo, i giudici prenderanno decisioni *ad hoc* che saranno soggette a rumore e quindi inique. Negli Stati Uniti, decisioni di questo tipo in passato erano la norma, con risultati vergognosi: persone apparentemente identiche, sulla sedia a rotelle o con una grave depressione o un dolore cronico, ricevevano trattamenti diversi. Per far fronte a questa situazione, i funzionari pubblici si sono indirizzati verso un sistema più simile a una regola: una *matrice di disabilità*, che richiede giudizi relativamente meccanici sulla base di fattori quali istruzione, posizione geografica e abilità fisiche residue, con l'obiettivo di ridurre il rumore nelle decisioni.

Nell'analisi più autorevole del problema, opera del docente di diritto Jerry Mashaw, questa azione tesa all'eliminazione dei giudizi soggetti a rumore viene chiamata *giustizia burocratica*,[3] un termine da tenere a mente. Mashaw elogia l'elaborazione della matrice come un atto di giustizia, proprio per la sua promessa di eliminare il rumore. In certe situazioni, tuttavia, questa promessa potrebbe non realizzarsi; ogni volta che

un'istituzione propende per decisioni vincolate da regole, vi è il rischio che riemerga il rumore.

Poniamo che in casi particolari le regole producano risultati inaccettabili: a quel punto i giudici potrebbero semplicemente ignorarle, ritenendole troppo severe, ed esercitare la propria discrezionalità attraverso una lieve forma di disobbedienza civile, che può essere difficile da tenere sotto controllo o anche solo da notare. Nelle società private, i dipendenti ignorano le regole fisse che sembrano inutili; analogamente, le agenzie governative incaricate di proteggere la sicurezza e la salute pubblica potrebbero rifiutarsi di adeguarsi a statuti troppo rigidi o simili a regole. Nel diritto penale statunitense è previsto l'annullamento della giuria in situazioni in cui i giurati si rifiutano di seguire la legge sostenendo che sia immotivatamente rigida e severa.

Ogniqualvolta un'istituzione pubblica o privata cerca di tenere sotto controllo il rumore attraverso regole ferree, deve sempre essere consapevole della possibilità che tali regole creino una discrezionalità sommersa. Una volta stabilita la politica del "tre errori e sei fuori", per esempio, era difficilissimo controllare o anche solo notare la frequenza con cui un pubblico ministero evitava di formulare un'accusa di reato contro chi aveva già subìto due condanne.

Al verificarsi di eventi simili si creerà rumore, ma nessuno se ne accorgerà. Dobbiamo sempre monitorare le regole per essere certi che funzionino come dovrebbero; in caso contrario, il rumore potrebbe indicare che quelle regole sono da rivedere.

Un quadro di riferimento

In ambito aziendale e governativo, la scelta tra regole e standard risulta spesso intuitiva, ma può essere disciplinata. Con una prima approssimazione, potremmo dire che dipende da soli due fattori: i costi delle decisioni e quelli degli errori.

Con gli standard, i costi delle decisioni possono essere molto elevati per i vari tipi di giudici, che dovranno adoperarsi per dare loro un contenuto. L'esercizio del giudizio può essere oneroso: se ai medici viene chiesto di cavarsela da soli nel formulare una diagnosi, dovranno passare del tempo a riflettere su ogni caso (e i loro giudizi potranno essere molto inclini al rumore), ma se al contrario vengono fornite delle chiare linee guida per stabilire se i pazienti hanno una faringite, per esempio, le loro decisioni saranno rapide e relativamente semplici. Se il limite di velocità è di centotrenta chilometri orari, gli agenti di polizia non saranno costretti a interrogarsi a lungo sulla velocità a cui è consentito viaggiare, ma se lo standard prevede che la gente non debba guidare «a una velocità eccessiva», dovranno rifletterci molto di più (e con ogni probabilità l'applicazione dello standard sarà soggetta a rumore). Con le regole, in genere, i costi delle decisioni sono molto più bassi.

Resta comunque una questione complicata. Una volta fissate, le regole possono essere semplici da applicare, ma prima *qualcuno deve stabilirle*. Elaborare una regola può essere difficile, e talvolta ha dei costi proibitivi; per questo i sistemi legali e le società private impiegano spesso aggettivi come *ragionevole*, *prudente* e *fattibile*, e simili termini assumono grande importanza anche in campi come la medicina e l'ingegneria.

I costi degli errori dipendono da quanto spesso e quanto gravemente si sbaglia. Occorre chiedersi se gli agenti addetti a prendere decisioni siano accorti e affidabili, e se pratichino strategie di igiene decisionale: in caso affermativo, uno standard funzionerà benissimo e creerà poco rumore. I

capi dovranno invece imporre delle regole quando hanno motivo di diffidare dei propri agenti: se sono incompetenti o affetti da bias, e se effettivamente non sono in grado di implementare misure di igiene decisionale, allora dovrebbero essere vincolati da regole. Un'organizzazione ragionevole capirà che il grado di discrezionalità da garantire è strettamente correlato al livello di fiducia nei propri agenti.

Naturalmente vi è un continuum tra la perfetta fiducia e la completa sfiducia. Uno standard potrebbe portare a numerosi errori da parte di agenti non affidabili, ma se si tratta di errori irrilevanti, potranno essere tollerati, mentre una regola potrebbe provocare pochi errori ma catastrofici, e in quel caso sarà meglio optare per uno standard. A questo punto dovrebbe essere chiaro che non esiste un motivo *generale* per ritenere che i costi degli errori siano più elevati con le regole o con gli standard. Se una regola è perfetta, naturalmente, non produrrà errori, ma è raro che le regole lo siano.

Poniamo che la legge stabilisca che è lecito acquistare alcolici solo se si ha un'età pari o superiore ai ventun anni. Questa legge intende proteggere i più giovani dai vari rischi associati al consumo alcolico, ma, se così intesa, produrrà moltissimi errori. Ci sono persone di venti, diciannove, diciotto o anche diciassette anni che sopportano l'alcol senza problemi, mentre altre di ventidue, quarantadue o sessantadue anni che non lo reggono affatto. Uno standard produrrebbe meno errori, se si trovasse una formula adatta e le persone la applicassero correttamente, ma, poiché questa formula è molto difficile da trovare, quasi sempre per la vendita di alcolici vengono imposte regole semplici basate sull'età.

Questo esempio ci porta a una considerazione molto più ampia: quando bisogna prendere tante decisioni è possibile che vi sia tanto rumore, il che costituisce un ottimo argomento a favore di regole chiare. Se i dermatologi

visitano molti pazienti che presentano nei e chiazze pruriginose, probabilmente commetteranno meno errori se i loro giudizi saranno vincolati da regole sensate. Senza tali regole, e con standard indefiniti, i costi delle decisioni tendono a diventare insostenibili, in particolar modo se si tratta di decisioni frequenti, nelle quali vi è un reale vantaggio nell'adozione di regole meccaniche più che di giudizi *ad hoc*. L'esercizio della discrezionalità risulta molto gravoso e i costi del rumore, o della relativa iniquità, potrebbero essere intollerabili.

Le organizzazioni più attente sono estremamente consapevoli degli svantaggi di entrambi questi metodi: adottano regole, o standard simili a regole, con l'intento di ridurre il rumore (e il bias), e per minimizzare i costi degli errori sono disposte a dedicare molto tempo e molta attenzione per garantire regole il più possibile accurate.

Vietare il rumore?

In molte situazioni il rumore dovrebbe fare scandalo. La gente ci convive, ma non è così che dovrebbe essere. Una contromisura semplice sarebbe passare da un'indefinita discrezionalità o un vago standard a una regola o qualcosa di simile, e ormai abbiamo un'idea più chiara dei casi in cui una soluzione facile può rivelarsi anche quella migliore. Ma anche nelle situazioni in cui una regola non sia applicabile o non sia una buona soluzione, abbiamo identificato varie strategie per ridurre il rumore.

Tutto ciò solleva un importante interrogativo: il sistema legale dovrebbe vietare il rumore? Sarebbe fin troppo facile dare una risposta affermativa, ma ciò non toglie che la legge dovrebbe fare molto di più per tenerlo sotto controllo. Questo problema è stato affrontato dal sociologo tedesco Max Weber, il quale si opponeva a quella che chiamava una «giustizia da cadì»,

fatta di giudizi informali formulati *ad hoc* e non soggiacenti ad alcuna regola generale, definendola intollerabile in quanto variava da caso a caso e costituiva una violazione del principio della legalità. Per citare lo stesso Weber, il giudice «si asteneva esplicitamente dal seguire regole formali "senza riguardo alla persona". Al contrario si basava proprio in larga misura sulle qualità concrete della persona e sulla situazione concreta: secondo equità e in conformità al risultato concreto».[4]

In questo approccio, sosteneva il celebre sociologo, «mancava totalmente qualunque sorta di garanzia giuridica certa». È chiaro che Weber obietta all'intollerabile rumore risultante dalla giustizia da cadì, celebrando l'ascesa dei giudizi burocratici e predisciplinati (il che ci riporta al concetto di "giustizia burocratica"). Nella sua ottica gli approcci specializzati, professionali e vincolati da regole erano il punto più alto dello sviluppo della legalità; oggi, a un secolo di distanza, è però chiaro che la giustizia da cadì contestata da Weber, o una sua variante, è ancora pervasiva. Il problema è capire cosa fare per contrastarla.

Non ci spingeremo a dire che la riduzione del rumore dovrebbe essere contemplata dalla Dichiarazione universale dei diritti umani, eppure in certi casi il rumore può essere considerato una violazione dei diritti, e in generale i sistemi legali di tutto il mondo dovrebbero impegnarsi di più per ridurlo. Pensiamo alle condanne penali, alle sanzioni civili per illeciti, alla concessione o negazione del diritto di asilo, alle opportunità formative, ai visti, alle concessioni edilizie o alle licenze professionali. O supponiamo che una grande agenzia governativa intenda assumere centinaia o anche migliaia di persone e le sue decisioni non seguano una logica: vi sarà di certo un rumore diffuso. O ancora, poniamo che i servizi sociali trattino i bambini in maniera molto diversa a seconda dell'assistente a cui viene

assegnato il caso: sarebbe inaccettabile che la vita e il futuro di un bambino dipendessero da una lotteria simile.

In molti casi la variabilità delle decisioni è chiaramente dettata da bias, compresi bias cognitivi identificabili e determinate forme di discriminazione. In tali circostanze, si tenderà a ritenere la situazione intollerabile, e ci sarà chi invocherà l'azione correttiva della legge, chiedendo nuove pratiche diverse da quelle vigenti. In tutto il mondo le organizzazioni danno la colpa ai bias, non a torto, ma non vedono allo stesso modo il rumore, quando invece dovrebbero.

In molti campi il livello attuale di rumore è eccessivamente elevato, impone alti costi e produce terribili ingiustizie. Qui abbiamo illustrato solo la punta dell'iceberg. La legge dovrebbe fare molto di più per ridurne i costi e contrastare questa ingiustizia.

A proposito di regole e standard

«Le regole semplificano la vita e riducono il rumore, ma gli standard permettono alle persone di adattarsi alle circostanze particolari.»

«Regole o standard? Innanzitutto bisogna chiedersi quale delle due misure produca più errori, poi quale sia più semplice o più gravosa da progettare e da attuare.»

«Spesso impieghiamo standard quando dovremmo adottare regole, semplicemente perché non prestiamo attenzione al rumore.»

«Non diciamo che la riduzione del rumore debba essere contemplata dalla Dichiarazione universale dei diritti umani (almeno non per il momento). Ciò non toglie che il rumore possa essere alla base di terribili ingiustizie, e i sistemi legali di tutto il mondo dovrebbero intraprendere azioni più energiche per ridurlo.»

[1] *Contenuti che incitano all'odio*, Standard della community, Facebook, [www.facebook.com/communitystandards/hate_speech].

[2] A. Marantz, *Why Facebook Can't Fix Itself*, in "The New Yorker", 12 ottobre 2020.

[3] J.L. Mashaw, *Bureaucratic Justice: Managing Social Security Disability Claims*, Yale University Press, New Haven, CT 1983.

[4] D.M. Trubek, *Max Weber on Law and the Rise of Capitalism*, in "Wisconsin Law Review", 720(1972), p. 733, nota 22 (citazione da M. Weber, *Confucianesimo e taoismo*, in Id., *Sociologia delle religioni*, vol. 1, a cura di Chiara Sebastiani, UTET, Torino 2008, p. 568).

Sintesi e conclusioni

Prendere sul serio il rumore

Il rumore è la variabilità indesiderata dei giudizi, e la sua presenza è dilagante. In questo libro ci siamo prefissi di spiegarne il motivo e di trovare delle possibili soluzioni. Abbiamo messo in campo molte idee e, a mo' di conclusione, riportiamo qui una breve sintesi dei punti principali per poi inserirli in una prospettiva più ampia.

Giudizi

Nell'accezione che qui abbiamo dato al termine, il *giudizio* non va confuso con l'attività del "pensiero". È un concetto molto più ristretto: il giudizio è una forma di misurazione il cui strumento è la mente umana. Come altre misure, assegna a un oggetto un punteggio, che non deve essere necessariamente un numero. «Il tumore della signora Johnson è con ogni probabilità benigno» è un giudizio, come lo sono affermazioni quali: «L'economia nazionale è molto instabile», «Fred Williams sarebbe la persona migliore da assumere come nuovo manager» e «Il premio per assicurare questo rischio dovrebbe essere di dodicimila dollari». I giudizi integrano in maniera informale diverse informazioni in una valutazione complessiva. Non si tratta di calcoli o dell'applicazione di regole precise: un insegnante usa il giudizio per valutare un tema, ma non per assegnare un voto a un quiz a risposta multipla.

Molte persone esprimono giudizi professionali per lavoro, con forti ripercussioni sulla vita di ognuno di noi. Tra i giudici di professione, come qui li abbiamo chiamati, si annoverano allenatori di calcio e cardiologi, avvocati e ingegneri, produttori cinematografici, sottoscrittori assicurativi e molti altri ancora. I giudizi professionali rappresentano il focus di questo libro, sia perché già ampiamente studiati, sia perché il modo in cui vengono formulati ha un grande impatto su tutti noi. Riteniamo inoltre che le nostre acquisizioni possano applicarsi anche a giudizi che vengono richiesti in altri ambiti della vita.

Esistono giudizi che abbiamo definito *predittivi*, e alcuni di questi sono verificabili: prima o poi scopriremo se erano corretti. Di solito questo è il caso delle previsioni a breve termine su risultati quali gli effetti di un trattamento medico, l'andamento di una pandemia o gli esiti di un'elezione. Ma una gran parte di questi, per esempio le previsioni a lungo termine e le risposte a domande fittizie, non sono verificabili, e la loro qualità si può valutare solo sulla base di quella del processo di pensiero da cui sono emersi. Peraltro, molti giudizi non sono predittivi ma *valutativi*: la condanna emessa da un giudice o la posizione che si aggiudica un dipinto nella classifica di un premio non sono confrontabili con un valore reale oggettivo.

Curiosamente, però, chi esprime un giudizio si comporta sempre come se questo valore reale esistesse, come se vi fosse un bersaglio invisibile a cui puntare, impossibile da mancare di molto. Con l'espressione *giudizio opinabile* si indica sia la possibilità di un disaccordo sia l'aspettativa che tale disaccordo sarà limitato. Le questioni di giudizio sono dunque caratterizzate dall'aspettativa di un *disaccordo limitato* e si collocano a metà tra le questioni di calcolo, in cui non è permesso alcun disaccordo, e quelle di gusto, in cui non ci si aspetta un grande accordo, eccetto in casi estremi.

Errori: bias e rumore

Diciamo che esiste un *bias* quando in un insieme di giudizi la maggior parte degli errori va nella stessa direzione. Il bias è l'*errore medio*, come, per esempio, quello che emerge quando al tiro a segno una squadra colpisce sistematicamente l'area in basso a sinistra del bersaglio; quando i dirigenti sono troppo ottimistici sulle vendite, anno dopo anno; quando una società continua a reinvestire in progetti fallimentari che dovrebbe invece abbandonare.

Eliminando il bias da un insieme di giudizi non si eliminerà del tutto l'errore. Gli errori che rimangono una volta rimosso il bias non sono condivisi: sono la divergenza non voluta nei giudizi, l'inaffidabilità dello strumento di misurazione che applichiamo alla realtà; sono il *rumore*, ovvero la variabilità di giudizi che dovrebbero essere identici. Il *rumore sistemico* è quello osservabile in organizzazioni che si avvalgono di professionisti intercambiabili per prendere decisioni, come i medici del pronto soccorso, i giudici che infliggono sanzioni penali e i sottoscrittori di una compagnia assicurativa. A questa tipologia di rumore è dedicata gran parte di questo libro.

Misurare il bias e il rumore

L'*errore quadratico medio* (o MSE) è da due secoli lo standard di accuratezza nelle misurazioni scientifiche. Le sue principali caratteristiche sono le seguenti: restituisce la media semplice come stima priva di bias della media della popolazione; tratta allo stesso modo errori positivi e negativi; penalizza gli errori gravi in maniera sproporzionata. L'MSE non riflette i costi reali degli errori di giudizio, che sono spesso asimmetrici. Detto ciò, le decisioni professionali richiedono sempre previsioni accurate: per una

città che si prepara ad affrontare un uragano i costi di una sottostima o di una sovrastima della minaccia chiaramente non sono gli stessi, ma sarebbe opportuno che tali costi non influenzassero le previsioni meteorologiche sulla velocità e la traiettoria del temporale. L'MSE è lo standard appropriato per esprimere giudizi predittivi di questo tipo, quando si punta all'accuratezza oggettiva.

Nella misurazione dell'MSE, bias e rumore vengono considerati due fonti indipendenti e cumulative di errore. Ovviamente il bias è sempre dannoso, e la sua riduzione aumenterà sempre l'accuratezza; meno intuitivo è il fatto che il rumore sia altrettanto dannoso, e la sua riduzione porti sempre a un miglioramento. Il miglior grado di dispersione è zero, anche quando i giudizi sono chiaramente affetti da bias. L'obiettivo, naturalmente, è ridurre al minimo sia il bias sia il rumore.

In un insieme di giudizi verificabili, il bias consiste nella differenza tra il giudizio medio su un caso e il corrispondente valore reale. Questo confronto è impossibile per i giudizi non verificabili: per esempio, il valore reale di un premio stabilito da un sottoscrittore per un particolare rischio non si potrà mai conoscere, così come quello di una condanna giusta per un particolare reato. Per questo motivo, spesso e volentieri (per quanto non sempre sia corretto) si parte dal presupposto che i giudizi siano privi di bias e che la media di molti giudizi sia la stima migliore del valore reale.

Il rumore di un sistema può essere valutato attraverso un *controllo del rumore*, un esperimento in cui diversi professionisti esprimono giudizi indipendenti sugli stessi casi (reali o fittizi). È possibile misurare il rumore senza conoscere un valore reale, proprio come è possibile vedere, sul retro di un bersaglio, la dispersione di una serie di colpi. I controlli del rumore sono in grado di misurare la variabilità dei giudizi in molti sistemi, come in un reparto di radiologia o nel sistema della giustizia penale; talvolta

possono richiamare l'attenzione su qualche lacuna nelle competenze o nella formazione, e sono in grado di quantificare il rumore sistemico, come quello che si presenta quando più sottoscrittori all'interno di uno stesso gruppo differiscono nella propria valutazione dei rischi.

Qual è il problema più grave, il bias o il rumore? Dipende dalla situazione, ma potrebbe benissimo essere quest'ultimo. Entrambi contribuiscono all'errore complessivo (MSE) nella stessa misura quando la media degli errori (il bias) è uguale alle deviazioni standard degli stessi (il rumore). Quando la distribuzione dei giudizi è normale (la classica curva a campana), gli effetti di bias e rumore sono uguali se l'84% dei giudizi si colloca al di sopra (o al di sotto) del valore reale: un bias considerevole, che spesso in un contesto professionale sarà facilmente identificabile. Quando il bias è inferiore a una deviazione standard, è il rumore la fonte principale dell'errore complessivo.

Il rumore è un problema

In alcuni giudizi la variabilità non è di per sé problematica, anzi può anche essere gradita. La diversità di opinioni è essenziale per formulare idee e opzioni, e i pareri contrastanti sono cruciali per l'innovazione: la pluralità di vedute tra critici cinematografici è un tratto caratteristico, non un intoppo; il disaccordo tra gli operatori finanziari è alla base dei mercati; le differenze strategiche tra startup rivali permettono agli investitori di scegliere quella che ritengono migliore. Nelle cosiddette questioni di giudizio, tuttavia, il rumore sistemico è sempre un problema: se due medici effettuano due diagnosi diverse, almeno uno avrà sbagliato.

Questo libro è nato dalla nostra sorpresa di fronte all'ubiquità del rumore sistemico e all'entità dei danni che esso produce, entrambe di gran

lunga più consistenti di quanto ci si potrebbe aspettare. Abbiamo presentato esempi da molti campi, come il mondo degli affari, la medicina, la giustizia penale, l'analisi delle impronte digitali, le previsioni finanziarie, le valutazioni del personale e la politica, giungendo alla seguente conclusione: dove c'è giudizio, c'è rumore, e più di quanto non si pensi.

Il grande peso del rumore nell'errore contraddice la credenza condivisa che gli errori casuali non contino, perché "si compensano a vicenda". Non è affatto così: se più tiri si sparpagliano in ogni punto del bersaglio, non aiuta dire che, in media, colpiscono il centro; se un candidato riceve una valutazione più alta di quanto meriti e un altro una più bassa, è possibile che venga assunta la persona sbagliata; se per una polizza assicurativa viene fissato un prezzo troppo alto e per un'altra un prezzo troppo basso, entrambi gli errori costeranno cari alla compagnia, in quanto uno le farà perdere un cliente, l'altro le farà perdere soldi.

Insomma, possiamo essere certi che siamo in presenza di un errore di fronte a giudizi che variano senza un valido motivo. Il rumore è dannoso anche quando i giudizi non sono verificabili e l'errore non è misurabile: non è giusto che persone nella stessa situazione ricevano un trattamento diverso, e un sistema i cui giudizi professionali siano ritenuti incoerenti perde di credibilità.

Tipi di rumore

Il rumore sistemico si può suddividere in *rumore di livello* e *rumore strutturale*. Alcuni giudici sono generalmente più severi, altri più clementi; alcuni analisti finanziari generalmente scommettono al rialzo e altri al ribasso nelle prospettive di mercato; alcuni medici prescrivono antibiotici più spesso di altri: sono tutti esempi di *rumore di livello*, cioè della variabilità

dei giudizi medi formulati da individui diversi. L'ambiguità delle scale di giudizio è una delle fonti di rumore di livello. Parole come "probabile" o anche numeri (per esempio "4 su una scala da 0 a 6") non hanno lo stesso significato per persone differenti. Il rumore di livello è un'importante fonte di errore nei sistemi di giudizio, nonché uno dei principali bersagli a cui devono mirare gli interventi volti alla riduzione del rumore.

Il rumore sistemico comprende un'altra componente, generalmente più ampia. A prescindere dal livello medio dei loro giudizi, due giudici possono avere idee diverse su quali siano i crimini che meritano le condanne più severe, e le loro decisioni produrranno una diversa *classificazione* dei casi. Chiamiamo questa variabilità *rumore strutturale* (il termine tecnico è *interazione giudice × caso*).

La principale fonte di rumore strutturale è stabile, e consiste nella differenza tra le reazioni personali e idiosincratiche dei giudici di fronte a uno stesso caso. Alcune di queste differenze riflettono princìpi o valori a cui gli individui si ispirano, in maniera non sempre conscia. Per esempio, un giudice potrebbe essere particolarmente severo con i taccheggiatori e insolitamente clemente con chi infrange il codice della strada, mentre un altro potrebbe mostrare una tendenza strutturale opposta. Alcuni dei princìpi o dei valori soggiacenti potrebbero essere piuttosto complessi, e il giudice potrebbe non esserne consapevole: per esempio, senza volerlo un giudice potrebbe essere più clemente verso i taccheggiatori anziani. Infine, anche una reazione altamente personale a un caso particolare potrebbe essere stabile: un'imputata che assomigli alla figlia del giudice potrebbe suscitare in lui la stessa compassione, e quindi la stessa clemenza, anche se l'udienza cadesse in un giorno diverso.

Questo *rumore strutturale stabile* riflette l'unicità dei giudici: la loro reazione ai casi è individuale come la loro personalità. Spesso le sottili

differenze tra le persone sono piacevoli e interessanti, ma diventano problematiche quando dei professionisti operano in un sistema che presuppone coerenza. Negli studi che abbiamo esaminato, il rumore strutturale stabile prodotto da queste differenze individuali è in genere la maggior fonte di rumore sistemico.

Detto ciò, gli atteggiamenti distintivi dei giudici in casi particolari non sono perfettamente stabili: il rumore strutturale ha anche una componente temporanea, che abbiamo chiamato *rumore occasionale*. Identifichiamo questo tipo di rumore quando un radiologo effettua diagnosi diverse a partire dalla stessa radiografia in giorni diversi, o se un esaminatore di impronte digitali identifica una corrispondenza tra due impronte in un'occasione ma non in un'altra. Come illustrano questi esempi, il rumore occasionale è più facile da misurare quando i giudici non ricordano di aver già esaminato in precedenza lo stesso caso. Un altro modo per dimostrare la presenza del rumore occasionale è evidenziare l'effetto che può avere sui giudizi un elemento contestuale irrilevante, come quando un giudice è più clemente dopo la vittoria della sua squadra del cuore o un medico prescrive più oppioidi di pomeriggio che di mattina.

La psicologia del giudizio e il rumore

I difetti cognitivi dei giudici non sono l'unica causa di errore nei giudizi predittivi: l'*ignoranza oggettiva* spesso riveste un ruolo ancora più rilevante. Certi fatti sono effettivamente inconoscibili – quanti nipoti avrà tra settant'anni un bambino che è nato ieri o quale sarà il numero del biglietto vincente in un'estrazione che verrà effettuata l'anno prossimo – mentre altri sono forse conoscibili, ma ignoti ai giudici. L'eccesso di fiducia nel

proprio giudizio predittivo porta qualunque persona a sottostimare la sua ignoranza oggettiva oltre che i suoi bias.

Vi è un limite nell'accuratezza delle previsioni, e spesso è piuttosto basso. Ciò nonostante, in genere siamo sereni sui nostri giudizi; a darci questo fiducioso appagamento è un *segnale interno*, un riconoscimento che ci autoattribuiamo per avere inserito fatti e giudizi all'interno di una storia coerente. La fiducia soggettiva nei propri giudizi non è necessariamente legata a un'accuratezza oggettiva.

Molte persone restano sorprese dalla scoperta che l'accuratezza dei propri giudizi predittivi non solo è bassa, ma è anche inferiore a quella delle formule. Perfino dei modelli lineari semplici costruiti su dati limitati o regole elementari e improvvisate arrivano a risultati sistematicamente superiori ai giudizi umani. Il vantaggio cruciale di regole e modelli è che sono esenti da rumore. Nella nostra esperienza soggettiva il giudizio è un processo sofisticato e complesso, e non ci rendiamo conto che quella complessità potrebbe non essere altro che rumore. Ci è difficile immaginare che l'aderenza meccanica a delle semplici regole spesso porterà a un'accuratezza maggiore della nostra, anche se questo è ormai un fatto assodato.

I *bias psicologici*, naturalmente, sono una fonte di errore sistematico, o bias statistico, ma – in maniera meno scontata – sono anche una fonte di rumore. Quando non sono condivisi da tutti i giudici, quando sono presenti in misura diversa e quando i loro effetti dipendono da circostanze esterne, i bias psicologici generano rumore. Se, per esempio, metà dei responsabili delle assunzioni ha un pregiudizio contro le donne e metà ha un pregiudizio in loro favore, nel complesso non vi sarà un bias, ma il rumore sistemico causerà molti errori nelle assunzioni. Un altro esempio è l'effetto spropositato della prima impressione: pur essendo un bias psicologico,

produrrà rumore occasionale se l'ordine in cui vengono presentate le caratteristiche dei candidati varia in maniera casuale.

Abbiamo descritto il processo di giudizio come l'integrazione informale di una serie di segnali per produrre un giudizio all'interno di una scala. Per eliminare il rumore sistemico, quindi, i giudici dovrebbero mantenere un'uniformità nell'impiego di questi segnali, nel peso da assegnare a ciascuno e nell'uso della scala. Anche lasciando da parte gli effetti casuali del rumore occasionale, queste condizioni vengono raramente soddisfatte.

L'accordo spesso è piuttosto elevato nei giudizi espressi su singole dimensioni: non di rado, per esempio, reclutatori diversi converranno su chi è più carismatico o diligente tra due candidati. Il processo intuitivo condiviso del *matching* nelle dimensioni di intensità, come l'associazione tra una media dei voti alta e un'età di lettura precoce, in genere produrrà giudizi simili. Lo stesso vale per i giudizi basati su un numero ridotto di segnali che puntano nella stessa direzione.

Le grandi differenze individuali emergono quando un giudizio richiede la *ponderazione di più segnali contrastanti.* Nel valutare lo stesso candidato, alcuni reclutatori daranno più peso alle caratteristiche dell'intelligenza o del carisma, mentre altri saranno più influenzati da aspetti come la diligenza o la calma in situazioni di stress. Quando i segnali sono contrastanti e non si inseriscono in una narrazione coerente, è inevitabile che persone diverse diano più peso ad alcuni e ne trascurino altri. Ciò produrrà rumore strutturale.

L'oscurità del rumore

Il rumore è un problema a cui non viene dato grande risalto. Se ne parla raramente, e spicca meno del bias; probabilmente voi stessi non ci avevate

mai pensato più di tanto. Vista la sua importanza, l'oscurità del rumore è un fenomeno di per sé interessante.

Per spiegare i giudizi inadeguati, vengono spesso chiamati in causa i bias cognitivi e altre distorsioni del pensiero emotive o motivate. Gli analisti citano l'eccesso di fiducia, l'ancoraggio, l'avversione alla perdita, il bias di disponibilità e altri ancora per spiegare decisioni che si sono rivelate fallimentari. Tali spiegazioni risultano soddisfacenti, perché la mente umana non desidera altro che motivazioni causali: ogni volta che qualcosa va male, cerchiamo una causa, e spesso la troviamo. In molti casi, la identificheremo in un bias.

Il bias ha una sorta di fascino esplicativo che manca al rumore: se cerchiamo di spiegare, col senno di poi, perché una certa decisione era sbagliata, troveremo facilmente l'uno e mai l'altro. Solo una *visione statistica* del mondo ci permette di vedere il rumore, ma questa visione non ci viene naturale, perché preferiamo le narrazioni causali. L'assenza del pensiero statistico nelle nostre intuizioni è uno dei motivi per cui il rumore riceve molta meno attenzione rispetto al bias.

Un altro motivo è che raramente i professionisti sentono il bisogno di affrontare la presenza del rumore nei propri giudizi e in quelli dei loro colleghi. È raro che gli esperti di impronte digitali, i sottoscrittori navigati e i funzionari più anziani degli uffici brevetti si soffermino a immaginare il possibile disaccordo dei propri pari, e ancor più raro che contemplino la possibilità di non trovarsi d'accordo perfino con se stessi.

I professionisti hanno quasi sempre fiducia nei propri giudizi: si aspettano che i colleghi convengano con loro, e non cercano mai di capire se è davvero così. In molti contesti è possibile che un giudizio non venga mai confrontato con un valore reale e che, al massimo, venga sottoposto alla verifica di un altro professionista considerato un *esperto di rispetto*. Solo

in rare occasioni ci si trova di fronte a un disaccordo inatteso e, quando ciò accade, in genere si riesce a trovare dei motivi per considerarlo un caso isolato. Di norma anche le organizzazioni tendono a trascurare o eliminare le prove di eventuali divergenze tra i propri esperti, il che è del tutto comprensibile: dal punto di vista dell'organizzazione, il rumore è una fonte di imbarazzo.

Come ridurre il rumore (e anche il bias)

Vi è motivo di credere che i giudizi di alcuni siano migliori di quelli di altri: le competenze specifiche, l'intelligenza e uno stile cognitivo definito come *apertura mentale attiva* caratterizzano i giudizi più validi, e non sorprende che un buon decisore commetterà pochi errori madornali. Considerando le varie origini delle differenze individuali, tuttavia, non dobbiamo aspettarci che i migliori siano in perfetto accordo su problemi di giudizio complessi. L'infinita varietà delle circostanze individuali, delle personalità e delle esperienze che ci rendono unici è anche alla base dell'inevitabilità del rumore.

Una possibile strategia per la riduzione del rumore consiste nell'eliminazione dei bias. Di solito le persone cercano di rimuoverli dai propri giudizi o correggendo questi ultimi a fatto compiuto o contenendo i bias prima che incidano sul giudizio. Noi proponiamo una terza opzione, applicabile soprattutto nelle decisioni che vengono prese in contesti di gruppo: identificare i bias in tempo reale designando un *osservatore decisionale* per individuarne i primi segni (vedi appendice B).

Per ridurre il rumore nei giudizi consigliamo innanzitutto di praticare l'*igiene decisionale*. Abbiamo scelto questo termine perché la riduzione del rumore, come l'igiene personale, è una forma di prevenzione contro un

nemico sconosciuto: come lavarsi le mani, per esempio, impedisce a patogeni ignoti di entrare nel nostro corpo, così l'igiene decisionale ci aiuterà a prevenire gli errori senza neanche sapere quali siano. Così come il suo nome, l'igiene decisionale non ha il fascino di una lotta vittoriosa contro i bias prevedibili, ed è di certo meno emozionante. Forse prevenire un pericolo non identificato non vi coprirà di gloria, ma ne vale sicuramente la pena.

Un'azione di riduzione del rumore all'interno di un'organizzazione dovrebbe sempre partire da un controllo del rumore (vedi appendice A). Questo passaggio ha l'importante funzione di ottenere un impegno da parte dell'organizzazione a prendere sul serio il problema. Un vantaggio essenziale sta nella valutazione di diversi tipi di rumore.

Abbiamo descritto i successi e i limiti delle azioni di riduzione del rumore in vari campi. Ricapitoleremo ora i sei princìpi di base dell'igiene decisionale, descriveremo come essi tengono conto dei meccanismi psicologici che provocano il rumore e mostreremo il loro legame con le specifiche tecniche di igiene decisionale già discusse.

L'obiettivo del giudizio è l'accuratezza, non l'espressione individuale. Questa affermazione è, a nostro avviso, il primo principio di igiene decisionale nei giudizi, e riflette l'accezione ristretta e specifica con cui abbiamo definito il giudizio in questo libro. Abbiamo mostrato come il rumore strutturale stabile sia un'ampia componente del rumore sistemico e una diretta conseguenza delle differenze individuali, delle personalità di giudizio che portano persone diverse a formarsi opinioni diverse su uno stesso problema. Questa osservazione ci porta a una conclusione forse impopolare ma incontrovertibile: il giudizio non è la sede appropriata per esprimere la propria individualità.

Sia chiaro, i valori personali, la singolarità e la creatività sono necessari, anzi essenziali, in molte fasi della riflessione e del processo decisionale come la scelta degli obiettivi, la formulazione di nuove modalità di approccio a un problema e l'elaborazione delle opzioni, ma quando si tratta di esprimere un giudizio su tali opzioni, le espressioni di individualità sono una fonte di rumore. Quando l'obiettivo è l'accuratezza e vi aspettate che gli altri siano d'accordo con voi, dovreste anche chiedervi cosa penserebbero altri giudici competenti se fossero al posto vostro.

Un'applicazione radicale di questo principio consiste nella sostituzione del giudizio con regole o algoritmi, in quanto la valutazione algoritmica garantisce l'eliminazione del rumore, anzi è l'unico approccio in grado di rimuoverlo del tutto. Gli algoritmi sono già in uso in molti ambiti importanti, e il loro impiego è in crescita, ma è improbabile che sostituiscano il giudizio umano nella fase finale di una decisione importante, e questa a noi sembra una buona notizia. In ogni caso, i giudizi possono essere migliorati, sia con un utilizzo appropriato degli algoritmi, sia con l'adozione di approcci che rendano le decisioni meno dipendenti dalle idiosincrasie di un unico professionista. Abbiamo visto, per esempio, come semplici linee guida decisionali possano contribuire a limitare la discrezionalità dei giudici o favorire l'omogeneità nelle diagnosi mediche, riducendo il rumore e migliorando le decisioni.

Pensare in termini statistici e assumere la visione esterna del caso. Diciamo che chi giudica assume la visione esterna di un caso quando lo considera come un elemento di una classe di riferimento di casi simili, piuttosto che come un problema singolo. Questo approccio diverge dalla modalità di pensiero automatica, che si concentra unicamente sulla situazione in esame e la inserisce all'interno di una narrazione causale. Quando le persone applicano le proprie esperienze peculiari per formarsi

un'opinione peculiare su un caso, producono rumore strutturale. La visione esterna è un rimedio a questo problema: i professionisti che si basano su una stessa classe di riferimento saranno meno inclini al rumore. Inoltre, porta spesso ad avere preziose intuizioni.

Il principio della visione esterna favorisce l'ancoraggio delle previsioni alle statistiche di casi simili, ricordando inoltre l'importanza di effettuare previsioni moderate (il termine tecnico è *regressive*; vedi appendice C). L'attenzione a un'ampia gamma di risultati pregressi e alla loro limitata prevedibilità dovrebbe aiutare i decisori a calibrare la fiducia nei propri giudizi: non si può biasimare qualcuno per non aver previsto l'imprevedibile, ma lo si può criticare per la sua mancanza di umiltà nelle previsioni.

Strutturare i giudizi in diversi compiti indipendenti. Questo principio del *divide et impera* è reso necessario dal meccanismo psicologico che abbiamo definito *eccesso di coerenza*, che porta a distorcere o escludere informazioni che non si inseriscono in una narrazione preesistente o emergente. Quando le impressioni su diversi aspetti di un caso si contaminano a vicenda, l'accuratezza generale ne risente: per trovare un'analogia, pensate a come cambia il valore probatorio di un gruppo di testimoni se viene permesso loro di comunicare.

È possibile ridurre questo eccesso di coerenza scorporando il problema di giudizio in una serie di compiti più piccoli. Questa tecnica è analoga alla pratica dell'intervista strutturata, in cui i selezionatori valutano un tratto alla volta e gli attribuiscono un punteggio prima di passare al successivo. Il principio della strutturazione è alla base di linee guida diagnostiche come il punteggio di Apgar, ed è inoltre il fondamento di un approccio che abbiamo chiamato *protocollo a valutazioni intermedie*, in cui un giudizio complesso viene scomposto in più valutazioni basate sui fatti, per

garantire che ciascuna di esse venga giudicata in maniera indipendente dalle altre. Ove possibile, per garantire tale indipendenza, ciascuna valutazione verrà assegnata a gruppi diversi, riducendo al minimo la comunicazione tra loro.

Resistere alle intuizioni premature. Abbiamo parlato del segnale interno, che dà ai decisori grande fiducia nei propri giudizi. La loro riluttanza ad abbandonare questo segnale gratificante è un motivo essenziale della resistenza all'uso di linee guida, algoritmi e altre regole vincolanti. È chiaro che i decisori hanno bisogno di sentirsi tranquilli sulla propria scelta finale e di arrivare a questa gratificante sensazione di sicurezza intuitiva. Ma non dovrebbero correre troppo: una scelta intuitiva fondata su una considerazione equilibrata e attenta dei dati empirici sarà molto più valida di un giudizio avventato. L'intuizione non va bandita, ma dovrebbe essere informata, disciplinata e differita.

Sulla base di questo principio, raccomandiamo di *sequenziare le informazioni*: i professionisti che esprimono dei giudizi non dovrebbero ricevere dati in più che non servono e che, anche se corretti, potrebbero farli incorrere in un bias. Nelle scienze forensi, per esempio, è buona norma tenere gli esaminatori all'oscuro di eventuali informazioni aggiuntive su un sospetto. In questo discorso rientra anche il controllo dei punti all'ordine del giorno, un elemento chiave del protocollo a valutazioni intermedie: un ordine del giorno efficace porterà a considerare diversi aspetti del problema in momenti separati e a rimandare la formazione di un giudizio olistico finché il profilo della valutazione non sarà completo.

Ottenere giudizi indipendenti da più valutatori, per poi eventualmente aggregarli. Il requisito dell'indipendenza viene normalmente violato nelle procedure attuate dalle organizzazioni, soprattutto nelle riunioni in cui le opinioni dei partecipanti si formano

sulla base di quelle altrui. A causa dell'effetto cascata e della polarizzazione di gruppo, spesso le discussioni non fanno che aumentare il rumore; la semplice procedura di raccogliere i giudizi dei partecipanti *prima* della discussione servirà a rivelare il grado di rumore e a facilitare un'armonizzazione costruttiva delle differenze.

Fare la media dei giudizi indipendenti garantisce la riduzione del rumore sistemico (ma non del bias): un giudizio singolo è un campione della popolazione di tutti i giudizi possibili, e aumentare la dimensione del campione migliorerà la precisione delle stime. Il vantaggio della media sarà ancora maggiore quando i decisori hanno competenze diverse e strutture di giudizio complementari. La media di un gruppo affetto da rumore potrebbe rivelarsi più accurata di un giudizio unanime.

Preferire giudizi e scale relativi. I giudizi relativi sono meno soggetti al rumore di quelli assoluti, perché la nostra capacità di effettuare confronti tra coppie di elementi è molto più alta della nostra abilità nel classificare degli oggetti su una scala. Le scale di giudizio che richiedono dei confronti saranno molto meno affette da rumore di quelle che richiedono dei giudizi assoluti: per esempio, una *scala di casi* richiede che i valutatori collochino un singolo caso su una scala precedentemente definita attraverso esempi noti a tutti.

I princìpi di igiene decisionale qui elencati si possono applicare non solo ai giudizi ricorrenti, ma anche alle grandi decisioni irripetibili, che abbiamo chiamato *decisioni singole*. L'esistenza del rumore nelle decisioni singole potrebbe sembrare controintuitiva: per definizione, non vi è alcuna variabilità da misurare se si decide una volta sola. Eppure il rumore c'è, e produce errori: in una squadra di tiratori, il rumore è invisibile se vediamo solo il primo in azione, ma la dispersione diverrebbe chiara se vedessimo anche gli altri. Analogamente, il modo migliore per pensare ai giudizi

singoli è trattarli come *giudizi ricorrenti formulati una volta sola*. Pertanto, l'igiene decisionale porterà un miglioramento anche in questi ultimi.

Applicare l'igiene decisionale può essere un compito ingrato: il rumore è un nemico invisibile, e una vittoria contro un nemico invisibile non potrà che essere una vittoria invisibile. Ma, al pari dell'igiene personale, anche quella decisionale ha un'enorme importanza. Dopo un intervento riuscito, tendiamo a pensare che siano state le competenze del chirurgo a salvarci la vita – e ovviamente è così – ma se il chirurgo e il personale presente in sala operatoria non si fossero lavati le mani, forse saremmo morti. Forse l'igiene non sarà mai un motivo di gloria, ma darà sempre dei risultati.

Quanto rumore?

Naturalmente, la lotta contro il rumore non è l'unica preoccupazione di decisori e organizzazioni. Ridurre il rumore potrebbe essere troppo costoso: un istituto scolastico potrebbe arrivare a eliminarlo del tutto nelle valutazioni chiedendo a cinque insegnanti di leggere ogni elaborato, ma sarebbe un onere ingiustificato. Un certo livello di rumore può essere di fatto inevitabile, un effetto collaterale necessario all'interno di un sistema basato sul giusto processo, che concede a ogni caso un esame individuale, che non tratta le persone come rotelle di un ingranaggio e assegna ai decisori un senso di responsabilità. Un certo livello di rumore può perfino essere auspicabile, se la variazione così creata permette al sistema di adattarsi ai tempi, come quando il rumore riflette un cambiamento di valori e obiettivi, e accende un dibattito che porta a modifiche nella prassi o nella legge.

È importante ricordare che le strategie di riduzione del rumore potrebbero avere svantaggi inaccettabili. Molti timori sugli algoritmi sono

esagerati, ma alcuni sono legittimi: gli algoritmi potrebbero produrre errori stupidi che un essere umano non commetterebbe mai, e quindi perdere credibilità, anche se allo stesso tempo riescono a prevenire diversi altri errori che gli esseri umani, invece, commettono eccome. Potrebbero essere affetti da bias per un difetto di progettazione o per essere stati allenati su dati inadeguati, e la loro impersonalità potrebbe non ispirare fiducia. Anche le pratiche di igiene decisionale hanno i loro lati negativi: se mal gestite, rischiano di burocratizzare le decisioni e demoralizzare i professionisti che le percepiscono come una minaccia alla propria autonomia.

Tutti questi rischi e limitazioni meritano piena considerazione. Tuttavia, un'obiezione alla riduzione del rumore può avere senso solo se riferita a una particolare strategia attuata per ridurlo. Per esempio, un'obiezione all'aggregazione dei giudizi, motivata magari dai suoi costi eccessivi, potrebbe non riguardare invece l'impiego di linee guida. Sicuramente, qualora i costi della riduzione del rumore superassero i suoi vantaggi, sarebbe meglio rinunciare; una volta effettuato il calcolo di costi e benefici, si potrebbe anche scoprire che il livello di rumore ottimale non è pari a zero. Il problema è che, senza un apposito controllo, non è possibile sapere quanto rumore c'è nei nostri giudizi. Stando così le cose, appellarsi alla difficoltà di ridurre il rumore non è altro che una scusa per non misurarlo.

I bias conducono a errori e ingiustizie, ma lo stesso vale per il rumore, eppure ci sforziamo molto meno per ridurlo. L'errore di giudizio può sembrare più tollerabile quando è casuale che non quando lo attribuiamo a una causa, ma non per questo è meno dannoso. Se vogliamo arrivare a decisioni migliori su ciò che conta, dovremmo prendere sul serio la riduzione del rumore.

Epilogo

Un mondo con meno rumore

Immaginate come sarebbero le organizzazioni se venissero ripensate in modo da ridurre il rumore: ospedali, commissioni addette alle assunzioni, società di consulenza finanziaria, agenzie governative, compagnie assicurative, autorità sanitarie, sistemi di giustizia penale, studi legali e università sarebbero attentissimi al problema e farebbero di tutto per ridurlo. I controlli del rumore sarebbero di routine; forse verrebbero effettuati addirittura una volta all'anno.

I direttori delle organizzazioni impiegherebbero gli algoritmi in sostituzione o a integrazione del giudizio umano in molte più aree di quelle oggi interessate. Le persone scomporrebbero i giudizi complessi in valutazioni intermedie più semplici, sarebbero a conoscenza dell'igiene decisionale e ne seguirebbero le prescrizioni. Verrebbero richiesti giudizi indipendenti, successivamente aggregati. Le riunioni sarebbero molto diverse, le discussioni più strutturate. Nel processo decisionale verrebbe sistematicamente integrata una visione esterna, gli aperti contrasti sarebbero molto più frequenti e verrebbero risolti in maniera più costruttiva.

Ne risulterebbe un mondo con meno rumore, in cui si risparmierebbero molte risorse, migliorerebbero la sicurezza e la salute pubblica, aumenterebbe l'equità e si preverrebbero molti errori evitabili. Il nostro obiettivo in questo libro era di portare l'attenzione su questa opportunità. Ci auguriamo che voi e altri riuscirete a coglierla.

Appendice A

Come condurre un controllo del rumore

Questa appendice costituisce una guida pratica per condurre un controllo del rumore. Dovreste leggerla dal punto di vista di un consulente che viene ingaggiato da un'organizzazione per esaminare la qualità dei giudizi professionali di una delle sue divisioni.

Come si evince dal nome, il controllo si concentra sulla diffusione del rumore; tuttavia, se ben condotto, fornirà preziose informazioni su bias, punti ciechi e difetti specifici nella formazione dei dipendenti e nella supervisione del loro lavoro. Un controllo efficace dovrebbe promuovere delle modifiche nelle attività della divisione, per esempio nei princìpi che guidano i giudizi dei professionisti, nella formazione a loro offerta, negli strumenti che pongono a fondamento dei propri giudizi e nella supervisione ordinaria del loro lavoro. Se il tentativo va a buon fine, potrà essere esteso ad altre divisioni della stessa organizzazione.

Un controllo del rumore richiede un notevole impegno e una grande attenzione al particolare, perché la sua credibilità verrà senz'altro messa in discussione qualora i risultati dovessero far emergere gravi falle. Ogni dettaglio dei casi e della procedura andrebbe quindi vagliato tenendo conto dell'ostilità di questo sguardo indagatore. Il processo qui descritto mira a ridurre i contrasti ingaggiando i critici potenzialmente più severi del controllo affinché risultino tra i suoi stessi fautori.

Oltre al consulente (che potrà essere esterno o interno), dovranno essere coinvolti i seguenti professionisti:

- *Il gruppo di progetto*, responsabile delle varie fasi dello studio. Se i consulenti sono interni, costituiranno il nucleo centrale del gruppo; in caso contrario, saranno affiancati da un gruppo di progetto interno. In questo modo il personale della società considererà il controllo come un *suo* progetto in cui gli esterni avranno un ruolo di secondo piano. Oltre ai consulenti che sovrintendono alla raccolta di dati, analizzano i risultati e preparano la relazione finale, il gruppo di progetto dovrebbe comprendere esperti della materia in grado di elaborare i casi da sottoporre alla valutazione dei soggetti giudicanti. Tutti i membri del gruppo dovranno godere di un'elevata credibilità professionale.
- *Clienti.* Un controllo del rumore sarà utile solo se porterà a un cambiamento significativo, cosa che richiede un coinvolgimento iniziale da parte dei dirigenti dell'organizzazione, i quali rappresentano i "clienti" del progetto. C'è da aspettarsi che sulle prime siano scettici riguardo alla diffusione del rumore, ma questo scetticismo iniziale è in realtà un vantaggio, se accompagnato da un atteggiamento aperto, da una curiosità verso i risultati del controllo e da un impegno a porre rimedio alla situazione se le aspettative pessimistiche del consulente trovano conferma.
- *Giudici.* I clienti designeranno una o più divisioni su cui eseguire il controllo. La divisione selezionata sarà composta da un numero significativo di "giudici", ovvero professionisti che esprimono giudizi e decisioni simili per conto della società. I giudici dovrebbero essere sostanzialmente intercambiabili, cioè se una persona non fosse disponibile per la gestione di un caso, quest'ultimo verrebbe assegnato a qualcun altro che dovrebbe pervenire a un giudizio simile. All'inizio del libro abbiamo riportato come esempi le decisioni dei giudici federali riguardo alle condanne, e le scelte relative all'importo dei premi per i rischi e degli accantonamenti per le richieste di risarcimento in una compagnia assicurativa. Per un controllo del rumore sarebbe meglio selezionare un compito di giudizio che possa essere (1) svolto sulla base di informazioni scritte e (2) espresso in forma numerica (per esempio in euro, in percentuali di probabilità o attraverso dei punteggi).
- *Responsabile del progetto.* Andrebbe nominato responsabile del progetto un manager di alto livello del personale amministrativo. Anche se per questo compito non sono necessarie specifiche competenze, un professionista di alto rilievo dell'organizzazione avrà un importante ruolo pratico nella gestione degli intoppi amministrativi, e la sua presenza sarà simbolica in quanto dimostrazione del valore che la società attribuisce al progetto. Il compito del responsabile è di fornire supporto amministrativo per facilitare tutte le fasi del controllo, comprese la stesura della relazione finale e la comunicazione delle conclusioni ai dirigenti della società.

Elaborazione dei materiali dei casi

Gli esperti della materia coinvolti nel gruppo di progetto dovrebbero avere un'esperienza comprovata nel compito assegnato alla divisione (per esempio, nella definizione dei premi per i rischi o nella valutazione del potenziale di possibili investimenti). A loro verrà affidato lo sviluppo dei casi da utilizzare nel controllo. La definizione di una simulazione credibile dei giudizi espressi dai professionisti sul posto di lavoro è un compito delicato, specialmente se si considera l'esame minuzioso a cui verrà sottoposto lo studio se emergeranno problemi seri. Il gruppo dovrà porsi questa domanda: se i risultati della nostra simulazione indicheranno un alto livello di rumore, il personale della società accetterà che i giudizi reali di questa divisione siano affetti da rumore? Il controllo del rumore avrà un senso solo se la risposta a questa domanda è un "sì" convinto.

Vi sono vari modi per garantire al progetto un'accoglienza positiva. Nel controllo del rumore nelle condanne descritto nel capitolo 1, per esempio, ogni caso è stato sintetizzato attraverso un breve elenco schematico delle relative caratteristiche, ottenendo la valutazione di sedici casi in novanta minuti, mentre in quello effettuato nella compagnia assicurativa descritta nel capitolo 2 sono stati forniti dei riassunti realistici dettagliati di casi complessi: in entrambe le occasioni, il riscontro di elevati tassi di rumore costituiva un'accettabile evidenza empirica, sulla base del criterio che se in esempi semplificati emergeva un tale disaccordo, nelle situazioni reali il rumore non poteva che essere peggiore.

Per ogni caso, occorre preparare un questionario che miri a indagare più a fondo sulle ragioni che hanno indotto ciascun giudice a esprimere un determinato giudizio. Il questionario andrebbe somministrato solo dopo il completamento di tutti i casi e dovrebbe comprendere:

- Domande aperte sui fattori principali che hanno spinto il partecipante a dare una certa risposta.
- Un elenco dei fatti relativi al caso, che il partecipante potrà classificare in ordine di importanza.

- Domande che richiedono una "visione esterna" della categoria a cui appartengono i casi. Per esempio, se è prevista una valutazione in termini monetari, i partecipanti dovrebbero fornire una stima di quanto al di sotto o al di sopra della media questa si collochi rispetto a tutte le valutazioni dei casi appartenenti alla stessa categoria.

Riunione prelancio con i dirigenti

Quando saranno stati raccolti i materiali da utilizzare nel controllo, andrebbe programmata una riunione in cui il gruppo di progetto presenti il controllo ai dirigenti della società. Nel corso dell'incontro andranno presi in considerazione i risultati possibili dello studio, tra i quali la rilevazione di un rumore sistemico inaccettabile. Scopo della riunione è ascoltare le obiezioni allo studio pianificato e ottenere un impegno da parte della dirigenza ad accettarne gli eventuali risultati: è inutile passare alla fase successiva in assenza di un simile impegno. Se vengono sollevate obiezioni serie, il gruppo di progetto potrebbe dover intervenire per migliorare i materiali da presentare e procedere a un nuovo tentativo.

Una volta che i dirigenti avranno accettato la struttura del controllo del rumore, il gruppo di progetto chiederà loro di esplicitare le proprie aspettative sui risultati dello studio. Andranno affrontate questioni come:

- «Quale livello di disaccordo vi aspettate tra una coppia di risposte selezionate in maniera casuale per ciascun caso?»
- «Qual è il livello massimo di disaccordo che ritenete accettabile, da una prospettiva aziendale?»
- «Qual è il costo stimato di una valutazione che si riveli errata di una certa percentuale per eccesso o per difetto (per esempio, del 15%)?»

Le risposte a queste domande andrebbero registrate, perché potrebbero non essere ricordate o perfino credute una volta ricevuti i risultati del controllo.

Somministrazione dello studio

I manager della divisione interessata dovrebbero essere informati sin dall'inizio in termini generici che il loro reparto è stato selezionato per uno studio particolare. Tuttavia, è importante che l'espressione *controllo del rumore* non venga impiegata nel descrivere il progetto: la parola *rumore* andrebbe evitata, specialmente in riferimento a persone. Al suo posto andrà impiegata una denominazione neutra come *studio dei processi decisionali.*

I manager della divisione saranno da subito i responsabili della raccolta dei dati e dell'enunciazione del compito ai partecipanti, con la collaborazione del responsabile del progetto e degli altri membri del gruppo. L'intento dell'esercizio andrebbe descritto ai partecipanti in linea generale, per esempio comunicando che «la società è interessata a capire come i decisori giungano alle proprie conclusioni».

È fondamentale rassicurare i professionisti che partecipano allo studio sul fatto che le risposte individuali non verranno rese note a nessun membro dell'organizzazione, compreso il gruppo di progetto. Se necessario, si può reclutare un'agenzia esterna che abbia il compito di rendere anonimi i dati. È importante anche sottolineare che non vi sarà alcuna ripercussione specifica sulla divisione, che è stata selezionata in quanto rappresentativa di tutti i reparti che svolgono compiti di giudizio per conto dell'organizzazione. Per garantire la massima credibilità dei risultati, dovrebbero partecipare allo studio tutti i professionisti idonei della divisione. Destinare all'esercizio mezza giornata lavorativa contribuirà a convincere i partecipanti della sua importanza.

L'esercizio dovrebbe essere svolto in contemporanea da tutti i partecipanti, che andrebbero distanziati e invitati a non comunicare tra

loro mentre lo studio è in corso. Il gruppo di progetto sarà a disposizione per rispondere a eventuali domande durante lo svolgimento dell'attività.

Analisi e conclusioni

Il gruppo di progetto sarà responsabile dell'analisi statistica dei diversi casi valutati da ogni partecipante, compresa la misurazione del tasso complessivo di rumore e delle sue componenti, ovvero il rumore di livello e il rumore strutturale; se i materiali dei casi lo consentono, identificherà inoltre i bias statistici presenti nelle risposte. Il gruppo avrà anche il compito non meno importante di provare a individuare le fonti di variabilità nei giudizi, esaminando le risposte al questionario in cui i partecipanti spiegano le proprie motivazioni e identificano i fatti che più hanno influenzato le loro decisioni. Concentrandosi soprattutto sulle risposte più estreme su entrambi i lati della distribuzione, il gruppo cercherà di identificare degli schemi nei dati: andrà alla ricerca di indicazioni di possibili lacune nella formazione offerta ai dipendenti, nelle procedure dell'organizzazione e nelle informazioni fornite da quest'ultima alle persone che lavorano per lei.

Il consulente e il gruppo di progetto interno collaboreranno allo sviluppo di strumenti e procedure per l'applicazione dei princìpi di igiene decisionale e l'eliminazione dei bias, al fine di migliorare i giudizi e le decisioni espressi dalla divisione. Questa fase del processo potrebbe estendersi su più mesi. In parallelo, il consulente e il gruppo di professionisti prepareranno anche una relazione sul progetto, che presenteranno ai dirigenti.

A questo punto, l'organizzazione avrà svolto un controllo del rumore a campione su una delle sue divisioni. Se il tentativo sarà considerato utile, il

gruppo dirigente potrebbe decidere di compiere un'azione più ampia per valutare e migliorare la qualità dei giudizi e delle decisioni prodotti in seno all'organizzazione.

Appendice B

Una checklist per l'osservatore decisionale

Quest'appendice offre un esempio generico di checklist che un osservatore decisionale potrebbe utilizzare (vedi capitolo 19). Segue grosso modo la sequenza cronologica del dibattito che conduce a una decisione importante.

Ulteriori chiarimenti sono offerti dalle domande suggerite all'interno di ogni punto della lista, che gli osservatori dovrebbero porsi mentre assistono alla discussione.

Questa checklist non nasce per essere utilizzata così com'è, ma per servire da ispirazione e punto di partenza per gli osservatori decisionali, che elaboreranno una propria lista personalizzata per l'osservazione dei bias.

Checklist per l'osservazione dei bias

1. **Approccio al giudizio**

1a. Sostituzione

– «La scelta dei dati empirici e il focus della discussione del gruppo indicano una sostituzione del quesito difficile che gli era stato assegnato con uno più semplice?»
– «Il gruppo ha trascurato un fattore importante (o sembra aver dato peso a un fattore irrilevante)?»

1b. Visione interna

– «Il gruppo ha adottato la visione esterna per una parte delle sue decisioni e ha cercato seriamente di applicare un giudizio comparativo anziché assoluto?»

1c. Diversità di vedute

– «Avete motivo di sospettare che i membri del gruppo siano accomunati da un bias che potrebbe essere all'origine di una correlazione tra i loro errori? Al contrario, riuscite a pensare a un punto di vista o a una competenza rilevante che non siano rappresentati nel gruppo?»

2. **Pregiudizi e conclusioni premature**

2a. Pregiudizi iniziali

– «Qualcuno dei decisori ha buone possibilità di trarre maggior vantaggio da una conclusione che da un'altra?»
– «Qualcuno si è già pronunciato a favore di una conclusione? Vi sono motivi per sospettare che sia mosso da un pregiudizio?»
– «Chi non è d'accordo ha espresso la propria opinione?»
– «Vi è il rischio che si precipiti verso una direzione fallimentare?»

2b. Conclusioni premature; eccesso di coerenza

– «Si è incorsi in un bias accidentale nella scelta di considerazioni esaminate prima del tempo?»
– «Le alternative sono state pienamente considerate? Si sono ricercati attivamente dei dati a sostegno?»
– «I dati o le opinioni scomode sono stati occultati o trascurati?»

3. Analisi delle informazioni

3a. Disponibilità e salienza

– «I partecipanti esagerano l'importanza di un evento in quanto recente o drammatico, o per motivi personali, anche se non è rilevante ai fini del giudizio?»

3b. Disattenzione alla qualità delle informazioni

– «Il giudizio è fortemente basato su aneddoti, racconti o analogie? Trova conferma nei dati?»

3c. Ancoraggio

– «Nel giudizio finale viene dato molto peso a numeri di dubbia accuratezza o rilevanza?»

3d. Previsione non regressiva

– «I partecipanti hanno effettuato estrapolazioni, stime o previsioni non regressive?»

4. Decisione

4a. Fallacia della pianificazione

– «Quando sono state utilizzate delle previsioni, i partecipanti si sono posti il problema della fonte e della loro validità? Ci si è avvalsi della visione esterna per metterle alla prova?»

– «Sono stati usati degli intervalli di confidenza per i dati incerti? Questi intervalli sono abbastanza ampi?»

4b. Avversione alla perdita

– «La propensione al rischio dei decisori è in linea con quella dell'organizzazione, o il gruppo decisionale è troppo cauto?»

4c. Bias del presente

– «I calcoli (compreso il tasso di attualizzazione impiegato) riflettono il bilanciamento delle priorità a breve e lungo termine effettuato dalla società?»

Appendice C

Correggere le previsioni

Le previsioni basate sul matching (vedi capitolo 14) sono errori in cui incorriamo quando facciamo affidamento sulle informazioni di cui disponiamo per formulare una previsione e agiamo come se queste fossero in grado di prevedere l'esito alla perfezione (o quasi).

Ripensiamo all'esempio di Julie, che era in grado di «leggere speditamente dall'età di quattro anni». Vi è stato chiesto di indovinare quale media dei voti avesse all'università: se ne avete predetta una pari a 3,8, avete giudicato intuitivamente che a quattro anni Julie rientrasse nel 10% dei suoi coetanei più precoci rispetto all'età di lettura (ma non nella fascia più alta, quella del 3-5%); dopodiché, con un ragionamento implicito, avete presupposto sulla base di questa informazione che in termini di voti Julie si sarebbe classificata intorno al novantesimo percentile degli studenti del suo anno, corrispondente a una media di 3,7 o 3,8 (due risposte, proprio per questo, molto gettonate).

A rendere questo ragionamento statisticamente errato è il fatto che sopravvaluta molto il valore diagnostico delle informazioni disponibili su Julie: una bambina precoce a quattro anni non sempre diventa un'eccellenza accademica (e, per fortuna, chi ha difficoltà di lettura da piccolo non resterà per sempre l'ultimo della classe).

Molto spesso, infatti, con il tempo le prestazioni eccezionali perdono di eccezionalità, e quelle scarsissime migliorano. È facile immaginare delle motivazioni sociali, psicologiche o anche politiche alla base di questa

osservazione, ma non è necessario: si tratta di un fenomeno puramente statistico. Le osservazioni estreme di un tipo o dell'altro tenderanno a diventare meno estreme, semplicemente perché non vi è una perfetta correlazione tra prestazioni passate e future. Questa tendenza è chiamata *regressione verso la media* (da cui il termine tecnico "non regressive" per indicare le previsioni basate sul matching, che non ne tengono conto).

In termini quantitativi, il giudizio che avete espresso su Julie sarebbe corretto se l'età di lettura fosse un predittore perfetto della media dei voti, cioè se vi fosse una correlazione pari a 1 tra i due fattori, ma non è questo il caso.

Esiste un metodo statistico che consente di arrivare a un giudizio con ogni probabilità più accurato, ma purtroppo non è intuitivo ed è di difficile impiego anche per chi abbia un'infarinatura di statistica. Di seguito spiegheremo questa procedura, illustrata nella figura 19 attraverso l'esempio di Julie.

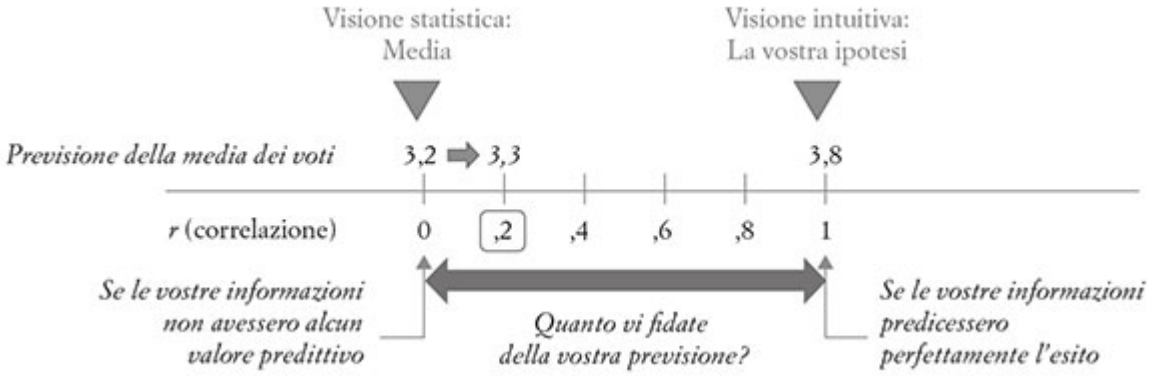


Figura 19. Adeguare una previsione intuitiva alla regressione verso la media

1. Fate un'ipotesi intuitiva.

La vostra intuizione su Julie, o su qualsiasi altro caso su cui avete delle informazioni, non è priva di valore. Il vostro sistema 1 del pensiero veloce porrà facilmente le informazioni in vostro possesso sulla scala della vostra

previsione ed esprimerà un punteggio relativo alla media dei voti di Julie: questa ipotesi è la previsione che effettuereste se le informazioni in vostro possesso fossero perfettamente predittive. Prendetene nota.

2. Individuate la media.

Ora fate un passo indietro e dimenticate per un attimo ciò che sapete su Julie. Come vi esprimereste sulla media dei voti di Julie *se non sapeste assolutamente niente di lei*? La risposta è evidente: in assenza di informazioni, la vostra migliore ipotesi sarebbe la media dei voti degli studenti del suo anno, che probabilmente si attesterebbe intorno a 3,2.

In questo modo avrete applicato il principio generale presentato in precedenza, la *visione esterna*. Quando assumiamo la visione esterna, pensiamo al caso in esame come a un singolo esempio all'interno di una classe considerata in termini statistici. A questo proposito, pensate a come assumere la visione esterna nel caso Gambardi ci abbia indotto a chiederci quale fosse il tasso di base del successo di un amministratore delegato appena nominato (vedi capitolo 4).

3. Stimate il valore diagnostico delle informazioni di cui disponete.

Questo è un passaggio difficile, in cui dovrete chiedervi: «Qual è il valore predittivo delle informazioni di cui dispongo?». Ormai dovrebbe esservi chiaro perché questa domanda è così importante: se di Julie non conosceste altro che il numero di scarpe, giustamente dareste a questa informazione un peso pari a zero e vi atterreste alla previsione basata sulla media dei voti della classe; se, d'altro canto, possedeste l'elenco dei risultati dei suoi esami, questa informazione sarebbe perfettamente predittiva della sua media dei voti (basterebbe un elementare calcolo matematico per ottenerla). Vi sono molte possibilità più sfumate tra questi due estremi. Se

disponeste dei dati sugli eccellenti risultati di Julie alle superiori, questa informazione avrebbe un valore diagnostico molto più elevato rispetto alla sua età di lettura, ma più basso rispetto ai voti dell'università.

Il vostro compito sta nel quantificare il valore diagnostico dei dati che avete a disposizione, espresso come correlazione con i risultati che vorreste prevedere. Salvo rari casi, questo numero sarà una stima approssimativa.

Per effettuarne una sensata, ricordate gli esempi riportati nel capitolo 12. Nelle scienze sociali le correlazioni superiori a 0,50 sono molto rare, e gran parte delle correlazioni che riteniamo significative sono vicine a 0,20. Nel caso di Julie, una correlazione di 0,20 è probabilmente il massimo a cui si possa aspirare.

4. Spostatevi dalla visione esterna verso la vostra ipotesi intuitiva, fino a raggiungere un punto che rifletta il valore diagnostico delle informazioni di cui disponete.

L'ultimo passo da compiere consiste in una semplice combinazione aritmetica dei tre numeri a cui siete arrivati: dovete muovervi dalla media in direzione della vostra ipotesi intuitiva in maniera proporzionale alla correlazione che avete stimato.

Si tratta semplicemente di estendere l'osservazione che abbiamo fatto poche righe fa: se la correlazione fosse 0, dovreste attenervi alla media; se fosse 1, dovreste lasciar perdere la media e procedere felicemente a una previsione basata sul matching. Nel caso di Julie, quindi, la previsione migliore che possiate formulare sulla sua media dei voti non dovrà allontanarsi più del 20% dalla media degli studenti del suo anno, in direzione della stima intuitiva suggerita dalla sua età di lettura. Questo calcolo vi porterà a una previsione di circa 3,3.

Qui abbiamo fatto riferimento all'esempio di Julie, ma questo metodo si può facilmente applicare a molti dei problemi di giudizio affrontati in questo libro. Pensate, per esempio, a un vicedirettore commerciale che debba assumere un nuovo addetto alle vendite e abbia appena avuto un colloquio con un candidato straordinario. Sulla base di questa forte impressione, il dirigente stima che il candidato porterà nelle casse della società un milione di dollari nel suo primo anno di lavoro, ovvero il doppio delle vendite medie dei neoassunti nei primi dodici mesi. Come può il vicedirettore rendere la sua stima regressiva? Il calcolo dipende dal valore diagnostico del colloquio, ovvero: in quale misura i colloqui riescono a predire il successo sul lavoro? Sulla base dei dati empirici in nostro possesso, una correlazione di 0,40 sarebbe una stima molto ottimistica. Di conseguenza, una stima regressiva delle vendite del neoassunto nel primo anno sarà, al massimo, di 500 000 $ + (1 000 000 $ – 500 000 $) × 0,40 = 700 000 $.

Questo processo, lo ripetiamo, non è affatto intuitivo. Nello specifico, come illustrano gli esempi, le previsioni ritoccate saranno sempre più prudenti di quelle intuitive: non saranno mai estreme come queste ultime, ma risulteranno più vicine, e spesso *molto* più vicine, alla media. Se correggete le vostre previsioni, non scommetterete mai che il campione di tennis che ha vinto dieci titoli del Grande Slam ne vincerà altri dieci, né prevedrete che una startup di successo stimata un miliardo di dollari diventerà un colosso di un valore centinaia di volte più alto. Le previsioni ritoccate non scommettono sulle anomalie.

Ciò significa che, col senno di poi, porteranno inevitabilmente ad alcuni fallimenti macroscopici. Ma le previsioni non si fanno col senno di poi. Dovreste ricordare che le anomalie, per definizione, sono rarissime; è molto più frequente l'errore opposto: quando prevediamo che le anomalie

resteranno tali, in genere ciò non avviene, a causa della regressione verso la media. Ecco perché ogni volta che l'obiettivo è la massima accuratezza (cioè il minimo errore quadratico medio), le previsioni ritoccate sono più precise di quelle intuitive basate sul matching.

Ringraziamenti

Abbiamo molte persone da ringraziare. Linnea Gandhi è stata la nostra direttrice d'orchestra, offrendoci la sua assistenza e il suo aiuto concreto, curando l'organizzazione, facendoci ridere e sorridere, e, di fatto, tenendo le fila del nostro lavoro. Oltre a tutto questo, ci ha offerto molti suggerimenti preziosi dopo la lettura del manoscritto: senza di lei non ce l'avremmo fatta. Dan Lovallo ha dato un grande contributo come coautore di uno degli articoli che hanno gettato i semi di questo libro. John Brockman, il nostro agente, è stato entusiasta, speranzoso, acuto e saggio in ogni fase della produzione; gli siamo molto grati. Tracy Behar, la nostra principale editor e guida, ha migliorato il libro in molti modi. Arabella Pike e Ian Straus ci hanno fornito degli ottimi suggerimenti editoriali.

Uno speciale ringraziamento va anche a Oren Bar-Gill, Maya Bar-Hillel, Max Bazerman, Tom Blaser, David Budescu, Jeremy Clifton, Anselm Dannecker, Vera Delaney, Itiel Dror, Angela Duckworth, Annie Duke, Dan Gilbert, Adam Grant, Anupam Jena, Louis Kaplow, Gary Klein, Jon Kleinberg, Nathan Kuncel, Kelly Leonard, Daniel Levin, Sara McLanahan, Barbara Mellers, Josh Miller, Sendhil Mullainathan, Scott Page, Eric Posner, Lucia Reisch, Matthew Salganik, Eldar Shafir, Tali Sharot, Philip Tetlock, Richard Thaler, Barbara Tversky, Peter Ubel, Crystal Wang, Duncan Watts e Caroline Webb, che hanno letto e commentato le bozze di alcuni capitoli e in certi casi dell'intero testo. Siamo grati per la loro generosità e il loro aiuto.

Abbiamo avuto la fortuna di avvalerci dei consigli di molti grandi studiosi. Julian Parris ci ha offerto un aiuto impagabile su molte questioni statistiche. I capitoli sui traguardi raggiunti dall'apprendimento

automatico non sarebbero stati possibili senza Sendhil Mullainathan, Jon Kleinberg, Jens Ludwig, Gregory Stoddard e Hye Chang. La sezione sulla coerenza di giudizio deve molto ad Alex Todorov e ai suoi colleghi di Princeton Joel Martinez, Brandon Labbree e Stefan Uddenberg, e anche a Scott Highhouse e Alison Broadfoot. Questi straordinari gruppi di ricercatori non solo ci hanno pazientemente illustrato le loro idee, ma sono stati così gentili da condurre delle analisi specifiche per noi. Naturalmente eventuali errori o incomprensioni sono imputabili unicamente a noi. Ringraziamo inoltre Laszlo Bock, Bo Cowgill, Jason Dana, Dan Goldstein, Harold Goldstein, Brian Hoffman, Alan Krueger, Michael Mauboussin, Emily Putnam-Horstein, Charles Scherbaum, Anne-Laure Sellier e Yuichi Shoda per aver messo a nostra disposizione le loro competenze.

Siamo inoltre grati a un vero e proprio esercito di ricercatori con cui abbiamo collaborato negli anni, come Shreya Bhardwaj, Josie Fisher, Rohit Goyal, Nicole Grabel, Andrew Heinrich, Meghann Johnson, Sophie Mehta, Eli Nachmany, William Ryan, Evelyn Shu, Matt Summers e Noam Ziv-Crispel. Molti dei ragionamenti qui condotti coinvolgono importanti campi di cui non abbiamo grande esperienza, e grazie al loro eccellente lavoro questo libro è meno soggetto a bias e a rumore di quanto non sarebbe stato altrimenti.

Infine, per tre autori residenti in due continenti diversi non è facile collaborare neanche nelle migliori circostanze, e il 2020 non è stato un periodo facile. Non avremmo completato questo libro senza le meraviglie tecnologiche di Dropbox e Zoom. Siamo grati ai creatori di questi magnifici prodotti.

Indice analitico

Yang, Crystal
Yu, Martin

Zuckerberg, Mark