Thinking Straight: Fallacies, Biases, and the Numbers That Fool Us

Figures current as of October 2026

The errors behind bad decisions, each taught through a case where it cost someone something: arguments that sound right and are not, numbers that mislead, the mind's own shortcuts, the pull of the group, and science's own mistakes. Where the field argues, over the hot hand, backfire, nudges, the Dunning–Kruger effect, replication, whether people are irrational at all, and whether debiasing works, both cases come before the evidence. The course closes with a checklist for any claim and the dates to watch.

MODULE 1 — How Arguments Go Wrong

Goal: Learn the words, and the ways an argument fails.

Fallacy, bias, and heuristic, from Aristotle's thirteen fallacies and Bacon's idols to Tversky and Kahneman's 1974 paper; valid and invalid forms; six ways to dodge a question, and when four of them are reasonable; when to trust an expert; and the tobacco companies' 1954 'Frank Statement', graded line by line.

MODULE 2 — Correlation Is Not Causation

Goal: Ask what else could explain a link.

Chance, confounders, reverse arrows, and 'after this, therefore because of this'; the vaccine scare built on twelve children; hormone therapy, which observational studies said protected the heart until a 2002 trial found the opposite; how smoking was shown to cause cancer without a trial; and chocolate and Nobel prizes, worked.

MODULE 3 — Chance

Goal: Put the base rate back.

What a positive mammogram means, and why most doctors in one study got it wrong; Sally Clark's '1 in 73 million'; the Linda problem; a reported run of black twenty-six times in a row at Monte Carlo in 1913; why small samples look extreme; and whether the hot hand exists, argued.

MODULE 4 — Who Got Counted

Goal: Find the cases that are missing.

Wald's bombers and the funds that closed; the Literary Digest's millions of ballots against Gallup's thousands; a treatment that wins with every size of kidney stone and loses overall, worked; regression to the mean, from Galton's heights to the flight instructors; the pill scare of 1995; and screening statistics that improve without saving lives.

MODULE 5 — The Halo Effect

Goal: Keep one good impression from deciding everything.

Thorndike's 1920 'constant error'; looks, credentials, and the board that made Theranos credible; the business halo that turned Cisco from brilliant to flawed with its share price, worked; and blaming character for what circumstance did.

MODULE 6 — Seeing What We Expect

Goal: Look for the evidence that could prove you wrong.

Wason's rule tests; the fingerprint three FBI examiners confirmed; the study where numerate people misread the same data once it was about guns; whether corrections backfire, argued; and chatbots that agree with their users.

MODULE 7 — How the Question Shapes the Answer

Goal: Restate a choice before making it.

Anchors, including young lawyers who rolled loaded dice; frames; why losses loom larger, and what selling winners and keeping losers costs, worked; defaults in organ donation and retirement plans; and whether nudges work, argued.

MODULE 8 — Memory and Story

Goal: Keep the outcome from rewriting the past.

The extra road deaths after 9/11; a memory rewritten by a question, and a witness who was certain and wrong; hindsight, outcome bias, and the stories told afterward; and the planning fallacy, from honors theses to the Sydney Opera House, with a reference-class forecast worked.

MODULE 9 — Overconfidence

Goal: Match confidence to the record.

Intervals that miss; Tetlock's experts and the superforecasters, with two forecasters scored; the overconfident trader; sunk costs and the Concorde; lawyers sanctioned for citing cases an AI invented; and whether the Dunning–Kruger effect is real, argued.

MODULE 10 — Crowds and Tribes

Goal: Keep independent judgment inside a group.

Asch's lines; the Bay of Pigs and Challenger; cascades and the MusicLab experiment; groups formed by a taste in paintings; polarization; and when crowds are wise, from Galton's ox to the arithmetic of independence, worked.

MODULE 11 — When Science Fools Itself

Goal: Read a study as a process, not a headline.

What a p-value says; how flexible analysis manufactures findings, and how many positive results are true, worked; the file drawer; the replication crisis, with the effects that failed and the ones that held; and whether science is broken, argued.

MODULE 12 — The Arguments, and the Toolkit

Goal: Weigh the arguments, and keep the checks.

Whether people are irrational or their shortcuts smart; whether debiasing works; five contradictions; a checklist for any claim; a dated watch list; and a last word.

A note on the evidence: the course rests on three kinds of source. Mathematical facts hold by definition, such as the arithmetic of base rates in Module 3. Experiments show that an error can happen and how often it happened in one setting. The record of real cases, from court judgments and official inquiries to the trials that reversed medical practice, shows what the error cost. Psychology has had to retest many of its famous experiments (Module 11), so the course says which of the findings it cites have been repeated and which have not. Where the field argues, over the hot hand, the backfire effect, nudges, the Dunning–Kruger effect, the replication crisis, whether people are irrational at all, and whether debiasing works, the course makes each case as its advocates make it, gives the evidence, and then states its own reading once. Ages and figures are as of October 2026.

Who's Who

The people in the course, with dates for those who have died and ages for the living

The researchers, investigators, and people at the center of the course's cases. Ages are as of October 18, 2026; a dash marks a birth date the course could not confirm. The modules column says where each person appears.

Person	Who they are	Dates or age	Modules
Aristotle	Greek philosopher; catalogued thirteen fallacies in the *Sophistical Refutations*	384–322 BC	1
Solomon Asch	Social psychologist at Swarthmore College; the conformity experiments with lines	1907–1996	10
Francis Bacon	English philosopher and statesman; the four 'idols' that distort the mind, in *Novum Organum* (1620)	1561–1626	1
Roy Baumeister	Social psychologist; proposed ego depletion	Born 1953; 73	11
Roger Boisjoly	Morton Thiokol engineer who argued against launching *Challenger*	1938–2012	10
Glenn Brier	Meteorologist and statistician; the Brier score for probability forecasts (1950)	1913–1998	9
Dana Carney	Psychologist at the University of California, Berkeley; first author of the power-posing study, which she disowned in 2016	—	11
Sally Clark	English solicitor wrongly convicted of murdering her two sons; freed on appeal in 2003	1964–2007	3, 12
Ronald Cotton	North Carolina man convicted of rape on a sincere, mistaken identification; freed by DNA evidence in 1995	—	8
Amy Cuddy	Social psychologist; co-author of the power-posing study and speaker of the TED talk on it	Born 1972; 54	11
Richard Doll	British epidemiologist; with Hill, linked smoking to lung cancer in 1950	1912–2005	1, 2
David Dunning	Social psychologist at the University of Michigan; the Dunning–Kruger effect	Born 1960; 65 or 66	9
Birte Englich	Social psychologist at the University of Cologne; anchoring in the courtroom	1968–2019	7
Baruch Fischhoff	Decision scientist at Carnegie Mellon University; hindsight bias	Born 1946; 80	8
Bent Flyvbjerg	Scholar of megaprojects; the iron law and reference-class forecasting	—	8
George Gallup	Pioneer of sample-survey polling; founder of the Gallup poll	1901–1984	4
Francis Galton	Victorian polymath; correlation, regression to the mean, the ox, and the founding of eugenics	1822–1911	4, 10
Gerd Gigerenzer	Psychologist, director emeritus at the Max Planck Institute for Human Development; natural frequencies and fast and frugal heuristics	Born 1947; 79	3, 8, 12
Thomas Gilovich	Psychologist at Cornell University; the hot-hand study of 1985 and *How We Know What Isn't So*	Born 1954; 72	3, 6
Austin Bradford Hill	British medical statistician; with Doll, smoking and lung cancer; the 1965 viewpoints on causation	1897–1991	1, 2
Elizabeth Holmes	Founder of Theranos; convicted of defrauding investors in 2022	Born 1984; 42	5
John Ioannidis	Meta-researcher at Stanford University; 'Why Most Published Research Findings Are False' (2005)	Born 1965; 61	11

Person	Who they are	Dates or age	Modules
Irving Janis	Psychologist at Yale University; the theory of groupthink	1918–1990	10
Daniel Kahneman	Psychologist; with Tversky, the heuristics-and-biases program; Nobel prize in economics, 2002; *Thinking, Fast and Slow*	1934–2024	1, 3, 4, 7, 8, 12
Justin Kruger	Social psychologist; with Dunning, the 1999 paper on the unskilled and unaware	—	9
Elizabeth Loftus	Cognitive psychologist at the University of California, Irvine; memory and eyewitness testimony	Born 1944; 82	8
Brandon Mayfield	Oregon lawyer wrongly identified by fingerprint examiners at the Federal Bureau of Investigation (FBI) after the 2004 Madrid bombings	—	6, 12
Roy Meadow	British pediatrician whose statistic helped convict Sally Clark	Born 1933; 93	3
Carey Morewedge	Researcher on judgment; tested debiasing games and videos	—	12
Brian Nosek	Psychologist at the University of Virginia; co-founder of the Center for Open Science	—	11
Linus Pauling	Chemist; Nobel prizes for chemistry (1954) and peace (1962); championed vitamin C against colds	1901–1994	1
Lee Ross	Social psychologist at Stanford University; named the fundamental attribution error (1977)	1942–2021	5, 12
Cass Sunstein	Legal scholar at Harvard University; co-author of *Nudge*; group polarization	Born 1954; 72	7, 10
James Surowiecki	Journalist; *The Wisdom of Crowds* (2004)	Born 1967; 59	10
Henri Tajfel	Social psychologist at the University of Bristol; minimal groups and social identity theory	1919–1982	10
Philip Tetlock	Psychologist at the University of Pennsylvania; expert political judgment and superforecasting	Born 1954; 72	9
Richard Thaler	Economist at the University of Chicago; Nobel prize in economics, 2017; co-author of *Nudge*	Born 1945; 81	7
Jennifer Thompson	The student whose certain, mistaken identification convicted Ronald Cotton	—	8
Edward Thorndike	Psychologist at Teachers College, Columbia University; named the halo effect (1920)	1874–1949	5
Amos Tversky	Cognitive psychologist; with Kahneman, the heuristics-and-biases program	1937–1996	1, 3, 7, 8, 12
Jørn Utzon	Danish architect of the Sydney Opera House	1918–2008	8
Diane Vaughan	Sociologist at Columbia University; the normalization of deviance	—	10
Andrew Wakefield	Former British doctor; lead author of the retracted 1998 paper linking the measles, mumps, and rubella (MMR) vaccine to autism; struck off in 2010	Born 1956; 70	2
Abraham Wald	Statistician; the Statistical Research Group at Columbia University in the Second World War	1902–1950	4
Peter Wason	British psychologist at University College London; the 2-4-6 and selection tasks	1924–2003	6

The Field Guide

Every error in the course, the phrase that gives it away, and the module that teaches it

Each entry names an error, says in a line what it is, and gives the kind of sentence that should make a reader stop and check. The giveaway is a warning sign, not proof: plenty of sentences that begin 'studies show' are true. The module number points to the explanation, the case, and the repair.

Error	What it is	The giveaway	Module
Ad hominem	Attacking the speaker instead of the argument	'Consider who's saying it'	1
Straw man	Refuting a weaker claim than the one made	'So you're saying…'	1
False dilemma	Offering two options when there are more	'Either… or…'	1
Slippery slope	Claiming a first step leads to the worst outcome, without showing the links	'Where does it end?'	1
Circular reasoning	Assuming the conclusion in the premises	'It's reliable because it says so'	1
Whataboutism	Answering a charge with a countercharge	'What about…?'	1
Misplaced authority	Citing expertise outside its field	'A Nobel laureate says…'	1
Appeal to popularity	Treating how many believe a claim as evidence that it is true	'Everyone knows…'	1
Shifting the burden of proof	Demanding proof against a claim instead of evidence for it	'There is no proof that…'	1
Affirming the consequent	Reasoning from 'if P, then Q' and Q to P	'That's exactly what you'd see if…'	1
Denying the antecedent	Reasoning from 'if P, then Q' and not P to not Q	'Without that, it can't happen'	1
Red herring	Drawing attention away from the question with a point beside it	'The real issue is…'	1
Misstated consensus	Presenting a settled question as open, or one expert's view as the field's	'Scientists disagree'	1
Correlation as causation	Treating two things that move together as cause and effect	'Linked to…', 'associated with…'	2
Confounding	A third factor drives both things measured	'People who do X are more likely to Y'	2
Reverse causation	The supposed effect is the cause	'Successful companies do X'	2
Post hoc reasoning	Assuming that what came first caused what followed	'Right after…, it started'	2
Data dredging	Searching many pairs of numbers and reporting the coincidence that turns up	'A striking correlation'	2, 11
Base-rate neglect	Ignoring how common something is before the evidence arrives	'The test is 90% accurate'	3
Prosecutor's fallacy	Mistaking the chance of the evidence, if innocent, for the chance of innocence	'The odds of this by chance are 1 in…'	3
Conjunction fallacy	Judging a detailed story more likely than one of its parts	'…and also…'	3
Gambler's fallacy	Expecting a run of chance to correct itself	'It's due'	3
Hot-hand belief	Expecting a run to continue: real in some settings, larger in belief than in fact	'He's on fire'	3
Law of small numbers	Expecting a small sample to look like the whole	'The best schools are the small ones'	3
Survivorship bias	Learning only from the cases that survived	'Every successful founder did…'	4

Error	What it is	The giveaway	Module
Nonresponse bias	Letting the people who choose to answer decide the result	'Our readers overwhelmingly say…'	4
Simpson's paradox	A comparison that holds in every group and reverses when the groups are pooled	'Overall, X does better'	4
Regression to the mean	Crediting a cause for an extreme result's return toward the average	'After we stepped in, it improved'	4
Relative-risk inflation	A large percentage of a small risk, reported without the baseline	'Doubles the risk'	4
Survival-rate illusion	Survival rising because diagnosis comes earlier, not because death comes later	'Five-year survival is up'	4
Halo effect	One good trait coloring judgments of unrelated ones; the horns effect is the bad-trait version	'Someone that impressive wouldn't…'	5
Credential halo	Eminence in one field taken for competence in another	'The board includes two former secretaries of state'	5
Business halo	Judging a company's practices by its results	'Their culture is why they win'	5
Fundamental attribution error	Blaming character for what circumstance did	'That's the kind of person he is'	5
Confirmation bias	Seeking, and weighting, the evidence that fits	'Every case I've seen confirms it'	6
Motivated reasoning	Setting the standard of proof by the conclusion wanted	'I'll believe it when…'	6
Contextual bias	Experts seeing what a case's context leads them to expect	'Confirmed by three examiners'	6
Backfire effect	A correction strengthening the belief it corrects; rarer than once thought	'Facts don't change minds'	6
Machine sycophancy	An AI assistant agreeing with the view its user states	'Even the AI agrees'	6
Anchoring	A first number pulling a later estimate toward it	'Was $500, now only…'	7
Framing effect	The same facts chosen differently when described differently	'90% survive' against '10% die'	7
Loss aversion	Weighting a loss about twice as heavily as an equal gain	'Don't lose what you have'	7
Endowment effect	Valuing a thing more because it is yours	'I wouldn't sell it for that'	7
Disposition effect	Selling winners too soon and holding losers too long	'I'll sell when it gets back to what I paid'	7
Default effect	Staying with whatever happens if you do nothing	'Unless you opt out…'	7
Availability	Judging how often something happens by how easily it comes to mind	'After what happened last week…'	8
Misinformation effect	A later question or account rewriting a memory	'How fast was it going when it smashed…'	8
Hindsight bias	Believing after the fact that the outcome was predictable	'I knew it all along'	8
Outcome bias	Judging a decision by how it turned out	'It worked, so it was right'	8
Narrative fallacy	Fitting events into a story that explains too much	'It all started when…'	8
Planning fallacy	Forecasting a project from its plan instead of the record of projects like it	'This time we'll finish on schedule'	8
Overprecision	Giving ranges too narrow for what is known	'I'm 99% sure'	9

Error	What it is	The giveaway	Module
Expert overconfidence	Confidence that grows faster than accuracy	'I've done this for thirty years'	9
Illusion of skill	Taking luck for ability, often by trading too much	'I can time it'	9
Sunk-cost fallacy	Continuing because of what has been spent rather than what is to come	'We've come too far to stop'	9
Automation bias	Trusting a machine's output over one's own checking	'The system says…'	9
Dunning–Kruger effect	The least skilled overrating themselves most; mostly a statistical artifact	'How hard can it be?'	9
Conformity	Giving the group's answer against one's own judgment	'Everyone else said…'	10
Groupthink	A close group suppressing doubt to keep its agreement	'Are we all agreed, then?'	10
Information cascade	Following others' choices instead of one's own information	'Look at the line outside'	10
In-group bias	Favoring one's own group on the thinnest of grounds	'One of us'	10
False consensus	Overestimating how many people share one's view	'Everyone I know thinks…'	10
Group polarization	Discussion among the like-minded pushing views further apart	'We all agreed, only more so'	10
Normalization of deviance	Each survived warning sign making the next one look routine	'It's been fine every time before'	10
P-hacking	Analyzing data many ways and reporting the way that worked	'After controlling for…, significant ($p = 0.049$)'	11
HARKing	Presenting a hypothesis formed after the results as if it came first	'As we predicted…'	11
Publication bias	Positive results published, null results filed away	'Studies show…'	11
Single-study certainty	Treating one unreplicated result as settled	'A new study proves…'	11
Misreading a p-value	Taking p as the chance a finding is false	'We're 95% sure it's real'	11
Bias blind spot	Seeing bias in others and not in oneself	'I'm just being objective'	12
The fallacy fallacy	Concluding a claim is false because an argument for it is bad	'That's a fallacy, so you're wrong'	12

Module 1 — How Arguments Go Wrong

The Words, the Forms, and the Ways to Dodge a Question

Recommended Reading / Reference

Book: Jamie Whyte, *Crimes Against Logic: Exposing the Bogus Arguments of Politicians, Priests, Journalists, and Other Serial Offenders* (McGraw-Hill, 2004)

Reference: Hans Hansen, 'Fallacies', *Stanford Encyclopedia of Philosophy* (revised 2024), with its history of the subject from Aristotle on

1a. Three Words

A fallacy is a flaw in an argument: a step from reasons to a conclusion that the reasons do not support. A bias is a tilt in judgment, a direction in which estimates and choices go wrong more often than chance would allow. A heuristic is a mental shortcut, a rule of thumb that gives a good-enough answer quickly. Most of the biases in this course are heuristics used outside the conditions they suit, which is why they are hard to give up: the shortcut that misleads in one setting saves time in a hundred others.

The three words come from different centuries. In the fourth century BC Aristotle catalogued thirteen ways an argument can seem to prove its point without doing so, in a short work known as the *Sophistical Refutations*. Six depend on language, such as equivocation, in which a word shifts its meaning partway through. Seven do not, among them begging the question and taking what is not a cause for a cause.

In 1620 Francis Bacon described four 'idols' that distort the mind, in his *Novum Organum*. The idols of the tribe are errors of human nature itself; those of the cave, an individual's quirks of temperament and training; those of the marketplace, the confusions of words; and those of the theatre, systems of thought taken on authority. One of his observations reads as a definition of confirmation bias (Module 6): once the mind has adopted an opinion, it draws everything else in to support and agree with it.

The modern study of biases began with a paper in *Science* in 1974, 'Judgment under Uncertainty: Heuristics and Biases', by the psychologists Amos Tversky and Daniel Kahneman. It named three shortcuts and the errors each produces. Representativeness judges probability by resemblance (Module 3); availability judges frequency by how easily examples come to mind (Module 8); and anchoring adjusts too little from a starting number (Module 7). Kahneman shared the 2002 Nobel prize in economics for the work. Tversky had died in 1996, and the prize is not awarded after death.

1b. The Shape of an Argument

An argument has premises and a conclusion. It is valid when the conclusion must be true if the premises are, and sound when it is valid and its premises are in fact true. Validity depends only on the form, so it can be checked without knowing anything about the subject. Four forms built on 'if P, then Q' turn up constantly in everyday reasoning, and two of them are traps.

Form	Pattern	Example	Valid?
Affirming the antecedent (modus ponens)	If P, then Q. P. So Q.	If the server is down, the site fails. The server is down. So the site fails.	Valid
Denying the consequent (modus tollens)	If P, then Q. Not Q. So not P.	If the server is down, the site fails. The site loads. So the server is up.	Valid

Form	Pattern	Example	Valid?
Affirming the consequent	If P, then Q. Q. So P.	If the server is down, the site fails. The site fails. So the server is down.	Invalid: a site can fail for other reasons
Denying the antecedent	If P, then Q. Not P. So not Q.	If the server is down, the site fails. The server is up. So the site loads.	Invalid, for the same reason

Most arguments outside mathematics are not deductive. They are inductive: they offer evidence that makes a conclusion more or less probable, and their test is strength rather than validity. A poll of likely voters does not prove who will win. It makes one result more likely than another, by an amount that depends on how the poll was taken (Module 4). Most of the errors in this course are inductive ones, in which the evidence is weaker than it looks.

1c. Six Ways to Dodge a Question

Six fallacies answer something other than the question asked. Four of them have a reasonable twin, which is why naming the fallacy is not enough: the work is to show which step fails. The other two are never reasonable.

Fallacy	What it does	When the move is reasonable
Ad hominem	Attacks the speaker instead of the argument	When the argument rests on the speaker's own word: a witness's record of lying bears on the testimony, though not on a calculation anyone can check
Straw man	Refutes a weaker version of the claim than the one made	Never; the cure is to restate the other side's view until its holders accept the summary
False dilemma	Offers two options when there are more	When the choice really is binary at that moment, as a yes-or-no vote is
Slippery slope	Claims a first step leads to the worst outcome, without showing the links	When a mechanism ties the steps together, as a legal precedent or an incentive does; then the question is how likely each link is
Circular reasoning	Assumes the conclusion in a premise	Never; the repair is a check that does not depend on the claim
Whataboutism	Answers a charge with a countercharge	When the point is that a standard is applied unequally; even then it says nothing about whether the charge is true

The last of the six has a short history as a word. In January 1974 a letter to the *Irish Times* complained of 'the Whatabouts', who met every charge against the Irish Republican Army with a charge against the other side, and days later a columnist in the same paper named the habit 'whataboutery'. *The Economist*'s Edward Lucas popularized 'whataboutism' in 2008, for the Soviet-era practice of answering any criticism with a complaint about the West.

1d. When to Trust an Expert

Everyone relies on people who know more. No reader can rerun a drug trial or inspect a bridge, and refusing all authority would leave a person knowing almost nothing. The fallacy is misplaced deference, and four questions separate it from the ordinary kind. Is the expert speaking within their own field? Does the field have a record of being right about questions like this one? Is the claim the field's consensus, or one member's view? And does the expert gain from the answer?

Linus Pauling is the classic case of the first question failing. He won the Nobel Prize in Chemistry in 1954 and the Peace Prize in 1962, and from 1970 he argued that large daily doses of vitamin C would prevent colds. Trials since have found that taking vitamin C regularly does not prevent colds in the general population, though it shortens them slightly. His authority was real, and it belonged to another subject. The second question matters as much: a forecaster's 70% chance of rain has decades of scored forecasts behind it, while a pundit's call on next year's recession usually has none (Module 9).

Popularity is evidence about what people believe, not about whether the belief is true, and ten outlets repeating one press release are one source, not ten (Module 10). The burden of proof sits with whoever asks others to accept a claim. A lack of evidence counts against a claim only when the evidence would probably have turned up had the claim been true: a missing article in a journal's complete archive is telling, while a missing record of a small local event proves little.

Worked example — 'A Frank Statement,' Graded Line by Line

On January 4, 1954, the American tobacco industry published 'A Frank Statement to Cigarette Smokers' in more than 400 newspapers, reaching an estimated 43 million readers. Its fourteen sponsors were cigarette makers and growers' and warehouse groups. It answered research linking smoking to lung cancer, among it Richard Doll and Austin Bradford Hill's study of 1950 (Module 2), and it announced a Tobacco Industry Research Committee. Read as an argument, it makes five moves, graded in the table below. A red herring, in the first row, is a point that draws attention away from the question at issue.

The statement's claim, paraphrased	The move	Grade
Medical research points to many possible causes of lung cancer	True, and beside the point: a disease can have many causes and still have this one	Red herring
The authorities do not agree on what the cause is	Treats a closing debate as an open one; the researchers who had studied the question were converging	Misstated consensus
There is no proof that smoking is one of the causes	No study of people outside a laboratory can prove a cause the way a theorem is proved; the real question was the weight of the evidence	Shifting the burden of proof
The statistics could apply just as well to other features of modern life	A fair point about confounding, used to dismiss the finding instead of testing it; researchers tested it (Module 2)	A true premise used for a false conclusion
The companies hold people's health above every other consideration, and will fund research	Says nothing about the evidence: intentions are not data	Appeal to good intentions

The statement's strongest point, that a correlation alone does not prove a cause, was used to shift the burden of proof, and the research committee helped keep the question open for years. An internal 1969 memo at the cigarette maker Brown & Williamson put the strategy in four words: 'Doubt is our product.' The US Surgeon General's report of 1964 concluded that cigarette smoking causes lung cancer in men. A fallacy's strongest weapon is a true premise put to work for a conclusion it does not support.

Module 1 Key Points

How Arguments Go Wrong

• A fallacy is a flaw in an argument, a bias a tilt in judgment, and a heuristic a shortcut that usually works.

• Aristotle listed thirteen fallacies; Bacon named four idols in 1620; Tversky and Kahneman named three heuristics in 1974.

• Affirming the consequent is invalid: a site that fails need not have a server that is down.

• Trust an expert inside their field, on the field's consensus, with a record of being right and no stake in the answer.

• The 1954 Frank Statement, in more than 400 newspapers, used a true point about causation to shift the burden of proof.

Module 1 Quiz

Check Your Understanding

1. Which three shortcuts did Tversky and Kahneman name in their 1974 paper?

A) Framing, loss aversion, and the endowment effect
B) Representativeness, availability, anchoring
C) Confirmation, conformity, and the halo effect
D) Hindsight, overconfidence, and the planning fallacy

2. Which of these argument forms is valid?

A) Denying the consequent (modus tollens)
B) Affirming the consequent
C) Denying the antecedent
D) Reasoning from a run of cases to a rule

3. When is pointing to a speaker's character relevant rather than an ad hominem?

A) When the speaker's record is poor, even on a calculation anyone can check
B) When the speaker is not an expert, so the argument can be set aside
C) When the argument rests on the speaker's own testimony, so honesty bears on whether to believe it
D) When the audience dislikes the speaker and doubts their motives

4. What move did the Frank Statement make with 'there is no proof'?

A) It admitted the link and promised a fix
B) It cited a trial that found no harm at all
C) It attacked the researchers
D) It shifted the burden of proof

5. Where was the word 'whataboutery' first recorded?

A) In the Soviet press of the 1950s
B) Ireland, 1974
C) An Economist column of 2008
D) A 1920s American manual of debate

Module 2 — Correlation Is Not Causation

Confounders, Reverse Arrows, and How a Trial Settles It

Recommended Reading / Reference

Book: Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic Books, 2018)

Reference: Austin Bradford Hill, 'The Environment and Disease: Association or Causation?', *Proceedings of the Royal Society of Medicine* 58 (1965): 295-300

2a. Four Explanations for One Link

When two things move together, four explanations compete. The link may be chance: a coincidence in a small sample, or one found by searching many pairs of numbers. A third factor, a confounder, may drive both. The arrow may run backward, from the supposed effect to the supposed cause. Or the first may really cause the second. A fifth explanation, that the way the cases were chosen created the link, has Module 4 to itself. A correlation fits all of them, which is why 'correlation is not causation' opens an investigation rather than closing one.

The correlation coefficient, written r, measures how closely two quantities move together along a straight line. It runs from −1, perfectly opposite, through 0, no straight-line relation, to +1, perfectly together, and it says nothing about which of the four explanations is true. Robert Matthews made the point in 2000 with figures from 17 European countries: the number of breeding pairs of storks correlated with the number of human births, with r = 0.62 and a p-value of 0.008. That p-value means a correlation at least this strong would turn up by chance less than 1% of the time if there were no real link (Module 11). Larger countries have more of both. Search enough pairs of numbers, a practice called data dredging, and strong correlations turn up by chance alone. Tyler Vigen's collection matches the number of films Nicolas Cage appeared in each year from 1999 to 2009 with the number of Americans who drowned by falling into a swimming pool, at r = 0.67.

2b. After This, Therefore Because of This

The oldest causal error has a Latin name, *post hoc ergo propter hoc*: after this, therefore because of this. It persuades because a cause does come before its effect, so every real cause passes the test of timing. Every coincidence passes it too.

On February 28, 1998, *The Lancet* published a paper by Andrew Wakefield and twelve co-authors describing twelve children with bowel problems and developmental regression; in eight of them, it said, the parents linked the onset of symptoms to the measles, mumps, and rubella (MMR) vaccine. The timing proved nothing. The vaccine is given in the second year of life, the age at which the first signs of autism usually become noticeable, so many children later diagnosed will have had the vaccine shortly before. Ten of the twelve co-authors withdrew the paper's interpretation in 2004. *The Lancet* retracted it in February 2010, after Britain's General Medical Council found that Wakefield had acted dishonestly, and the council struck him off the medical register that May. Vaccination rates in Britain had fallen in the meantime, and measles outbreaks followed. A Danish study published in 2019 followed 657,461 children born from 1999 to 2010 and found no higher risk of autism among the vaccinated: the hazard ratio, the rate in the vaccinated relative to the unvaccinated, was 0.93, a difference within chance.

2c. The Hormone Therapy Reversal

By the late 1990s hormones after menopause were widely prescribed in the United States, partly because observational studies, which compare people as they are found, suggested that the drugs protected the heart. The largest, the Nurses' Health Study, reported in 1991 that among 48,470 women followed for up to ten years, current users had 44% fewer major coronary events than women who had never used hormones. American prescriptions rose from 58 million in 1995 to 90 million in 1999.

The comparison was between women who chose hormones and women who did not, and the two groups differed. The women who chose them were, on average, healthier, wealthier, and more attentive to their health, which made them likelier to avoid heart disease whatever pills they took. Statistical adjustment removes only the differences that were measured, and this one, the healthy-user effect, was largely unmeasured.

A randomized trial, in which chance rather than the woman decides who gets the drug, settled the question. The Women's Health Initiative assigned 16,608 women aged 50 to 79 to estrogen plus progestin or a placebo. In May 2002 its safety board recommended stopping that part of the trial early, and the results, published in *JAMA* on July 17, 2002, ran the wrong way. Women on the hormones had 29% more coronary events, 41% more strokes, 26% more invasive breast cancers, and twice the rate of blood clots in the lungs, against fewer hip fractures and colorectal cancers. In absolute terms, for every 10,000 women taking them for a year there were about 7 more coronary events, 8 more strokes, 8 more clots in the lungs, and 8 more breast cancers, and 6 fewer colorectal cancers and 5 fewer hip fractures. In the first half of 2003, American prescriptions for the combined pill were two-thirds lower than a year earlier.

The reversal has a coda. The trial's women averaged 63, about a dozen years past the usual age of menopause, and a 2007 reanalysis of both hormone trials found a tendency, short of the authors' test for statistical significance, for the coronary risk to depend on timing. For women within ten years of menopause the hazard ratio was 0.76, a difference within chance; for those twenty or more years past it, 1.28. Both lessons stand: the observational studies were confounded, and a trial answers the question it asked, in the people it enrolled.

2d. How Smoking Was Shown to Cause Cancer Without a Trial

No one can randomly assign people to smoke for forty years, yet the evidence that smoking causes lung cancer is as strong as any in medicine, and it was built from observation. In 1950 Richard Doll and Austin Bradford Hill compared 649 men with lung cancer in twenty London hospitals with 649 similar men admitted for other illnesses. Only 2 of the men with lung cancer had never smoked, against 27 of the others. In October 1951 the two began following the British doctors who answered a questionnaire on their smoking, 34,439 men among them, and the deaths that followed tracked the dose. In 2004, half a century on, Doll and Richard Peto reported that the smokers had died about ten years younger, and that those who quit at 30 had regained nearly all of the loss.

In 1965 Hill set out the questions that turn an association into a strong case for causation, in an address to the Royal Society of Medicine. He called them viewpoints and insisted that none was required; they are now known as the Bradford Hill criteria.

Viewpoint	The question it asks	Smoking's answer
Strength	How large is the association?	Heavy smokers' risk of lung cancer was many times nonsmokers'

Viewpoint	The question it asks	Smoking's answer
Consistency	Is it found again, by other people, places, and methods?	In study after study, in country after country
Specificity	Is the cause tied to one effect?	The weakest point: smoking causes many diseases
Temporality	Does the cause come before the effect?	Smoking precedes the cancer by decades
Biological gradient	Does more exposure bring more effect?	Risk rises with the number of cigarettes a day
Plausibility	Is there a likely mechanism?	Tobacco smoke contains carcinogens
Coherence	Does it fit what else is known?	Lung cancer rose after cigarettes spread
Experiment	Does removing the cause reduce the effect?	People who quit cut their risk
Analogy	Do similar causes have similar effects?	Other inhaled carcinogens cause lung disease

This is the answer to the tobacco companies' claim that the statistics could apply to any feature of modern life (Module 1). The confounding they alleged was tested, and the evidence pointed the same way from every direction. One correlation is not causation, but many correlations of the right shape, with a dose, a sequence, and a mechanism, are how causation is established when a trial is impossible.

Worked example — Chocolate and Nobel Prizes

In October 2012 the *New England Journal of Medicine* published a short note by the cardiologist Franz Messerli plotting chocolate eaten per person against Nobel laureates per 10 million people for 23 countries. The correlation was r = 0.791, with P below 0.0001; Switzerland led on both. Run the four explanations of section 2a. Chance is unlikely across 23 countries at that strength. Reverse causation was Messerli's own suggestion: clever people might eat more chocolate. Confounding is the obvious candidate, because national wealth buys chocolate and also buys universities, laboratories, and long research careers, so rich countries head both lists. Causation would need a mechanism of the right size; Messerli cited studies suggesting that cocoa compounds called flavanols may help cognition, but nothing suggests an effect large enough to move a country's prize count. Sweden is the tell. It had more than twice the laureates its chocolate predicted, and most of the prizes are chosen in Stockholm; leaving Sweden out raised the correlation to 0.862. The tests that would separate the explanations are a comparison among countries of similar wealth, or a trial: give one group cocoa and wait for the prizes. The note was half a joke, and it is useful for that reason: a strong, real, published correlation with its confounder in plain sight.

Module 2 Key Points

Correlation Is Not Causation

• A correlation fits chance, a confounder, a reversed arrow, selection, or a cause; on its own it cannot say which.

• Wakefield's 1998 paper described twelve children; a 2019 Danish study of 657,461 children found no link between MMR and autism.

• Observational studies found 44% fewer coronary events in hormone users; the 2002 randomized trial found 29% more.

• Smoking was shown to cause cancer without a trial, through a link that was strong, consistent, dose-related, and fell when people quit.

• Chocolate and Nobel prizes correlate at r = 0.79 across 23 countries; national wealth is the likeliest explanation of both.

Module 2 Quiz

Check Your Understanding

1. What did the Women's Health Initiative trial find for coronary events among women on estrogen plus progestin?

A) They fell by 44%, as the observational studies had found
B) No change
C) About 29% more than on placebo
D) Slightly fewer

2. What did Austin Bradford Hill call his criteria for judging causation?

A) Viewpoints
B) Laws
C) Necessary conditions
D) Rules that must all be met before a cause is accepted

3. Why did the observational studies of hormone therapy mislead?

A) The studies were too small to detect heart attacks in women of that age
B) The trial used a different hormone dose from the one most women took
C) The women were followed for too short a time to see the harm
D) Women who chose the hormones were healthier to begin with, a difference adjustment could not remove

4. What does a correlation of 0.79 between chocolate and Nobel prizes establish by itself?

A) That chocolate raises the chance of a prize by 79%
B) Only that the two rise together
C) That one causes the other
D) That national wealth must be the cause of both

5. Why is the timing of the MMR vaccine easy to mistake for a cause of autism?

A) The vaccine is given at about the age when the first signs of autism usually appear
B) Autism diagnoses fell wherever the vaccine was withdrawn
C) Most children get two doses
D) The 1998 paper followed thousands of children

Module 3 — Chance

Base Rates, Coincidences, and Streaks

Recommended Reading / Reference

Book: Gerd Gigerenzer, *Calculated Risks: How to Know When Numbers Deceive You* (Simon & Schuster, 2002)

Reference: Gerd Gigerenzer and colleagues, 'Helping Doctors and Patients Make Sense of Health Statistics', *Psychological Science in the Public Interest* 8 (2007): 53-96

3a. The Base Rate

A woman has a screening mammogram, and it is positive. How likely is it that she has breast cancer? In 2007 the psychologist Gerd Gigerenzer put that question to 160 gynecologists at the start of a continuing-education session, with the facts they needed. About 1% of the women screened have breast cancer. A mammogram detects 90% of them. And 9% of women without cancer also test positive. Offered four answers, only 21% chose the right one, about 1 in 10, which is fewer than would have picked it by chance; many chose 9 in 10. After a short session on the method below, 87% answered correctly.

Worked example — What a Positive Mammogram Means

Count people instead of percentages. Take 1,000 women screened: 10 have breast cancer and 990 do not. Of the 10, the mammogram finds 9, which is 90%. Of the 990, about 89 test positive anyway, which is 9%. So 98 women test positive, and 9 of them have cancer: 9 in 98, about 9%. The test is good, and most of its positives are still false, because the 9% false-alarm rate applies to 990 women while the 90% detection rate applies to only 10. The same arithmetic in symbols is Bayes' rule: the chance of cancer given a positive result is 0.90 × 0.01, divided by 0.90 × 0.01 plus 0.09 × 0.99, which comes to 0.092. Counting people, which Gigerenzer calls natural frequencies, gets there without the formula, and it is what the doctors were taught. It also explains why the follow-up test after a positive screen is so informative: among the 98 who tested positive, cancer is no longer rare.

In general, ask how common something was before the evidence arrived, the base rate, and weigh the evidence against it. A fraud alert, a security screen, or a profile of a likely criminal faces the mammogram's arithmetic: when the target is rare, even an accurate test produces mostly false alarms.

3b. The Prosecutor's Fallacy

In November 1999 Sally Clark, a solicitor from Cheshire, was convicted of murdering her two baby sons: Christopher, who died in December 1996 at about eleven weeks, and Harry, who died in January 1998 at eight weeks. The prosecution's most memorable evidence came from the pediatrician Sir Roy Meadow. He told the jury that the chance of two cot deaths, or sudden infant deaths, in a family like hers was about 1 in 73 million. He reached the figure by taking the chance of one such death, 1 in 8,543, and squaring it. The jury convicted by a majority of ten to two.

The number contained two errors. Squaring assumes that the two deaths were independent, like two tosses of a coin, but the genetic and household factors behind one cot death make a second more likely, not less. The deeper error, the prosecutor's fallacy, lay in how the figure was heard. The chance that an innocent family suffers two cot deaths is not the chance that the mother is innocent, because double murder of babies

is also extremely rare, and the real question is which of two rare explanations is likelier. The mathematician Ray Hill estimated that for a family with two such deaths, two cot deaths were between 4.5 and 9 times as likely as two murders. The Royal Statistical Society said publicly in October 2001 that the 1-in-73-million figure had no statistical basis.

Clark's first appeal failed. A second, in January 2003, quashed her conviction, after it emerged that the pathologist had not disclosed test results suggesting that Harry may have died of an infection. She had spent more than three years in prison, and she died in March 2007, aged 42. Meadow was struck off the medical register in July 2005; the High Court reinstated him in 2006, and the Court of Appeal upheld that ruling. Other convictions that had leaned on his evidence were reviewed, and two, those of Angela Cannings and Donna Anthony, were overturned.

3c. The Linda Problem

In a 1983 study, Amos Tversky and Daniel Kahneman described a woman they called Linda: thirty-one, single, outspoken and very bright, a philosophy graduate who had cared deeply about discrimination and social justice as a student and had joined antinuclear demonstrations. Subjects ranked statements about her by probability. Among statistically naive undergraduates, 89% ranked 'Linda is a bank teller and is active in the feminist movement' above 'Linda is a bank teller'. So did 85% of a group of doctoral students in decision science at Stanford's business school, who knew the theory of probability well.

It cannot be more probable. Every feminist bank teller is a bank teller, so the chance of both together can never exceed the chance of either one: the conjunction rule. The description fits a picture of a feminist, so the fuller story feels more representative and is judged more likely, the representativeness heuristic at work (Module 1). The lesson reaches well beyond Linda. Every added detail makes a story more plausible and less probable, which is how a forecast or a conspiracy theory with many vivid parts can seem more convincing than a plain one (Module 8).

3d. Streaks and Small Samples

On August 18, 1913, at the casino in Monte Carlo, the ball is reported to have fallen on black twenty-six times in a row, while gamblers bet ever more heavily on red, reasoning that red was due. The wheel had no memory: on every spin, the chance of black stayed 18 in 37. The belief that chance corrects itself in the short run is the gambler's fallacy. Twenty-six blacks in a row has a probability of about 1 in 137 million from any given spin, but casinos have spun wheels many millions of times, so a run that long was bound to turn up somewhere. A streak predicted in advance is remarkable; a streak found afterward, among all the runs that could have happened, is not.

Small samples swing further than large ones, so they supply the extremes at both ends; Tversky and Kahneman called the mistaken expectation that a small sample will mirror the whole the law of small numbers. In the United States, the counties with the lowest rates of kidney cancer are mostly rural, sparsely populated, and in the Midwest, the South, and the West, and so are the counties with the highest rates. Rural life explains neither: small populations produce extreme rates by chance in both directions. The same arithmetic flattered small schools, which are overrepresented among the best performers and among the worst. The Bill & Melinda Gates Foundation put more than $2 billion into its high-school strategy from about 2000, including grants to many small schools, and in 2009 Bill Gates wrote that many of the small schools had not significantly raised achievement.

3e. The Argument: Does the Hot Hand Exist?

The case that it is a fallacy. In 1985 Thomas Gilovich, Robert Vallone, and Amos Tversky studied the field goals of the Philadelphia 76ers in the 1980-81 season, the Boston Celtics' free throws, and a shooting experiment with 26 Cornell players. A hit was no more likely after a run of hits than after misses, although 91% of the 100 fans they surveyed believed a player had a better chance after two or three hits in a row. For thirty years the 'hot-hand fallacy' was a textbook case of seeing patterns in chance.

The case that it is real. In 2018 Joshua Miller and Adam Sanjurjo showed that the standard measure was biased. In a finite run of shots, the share of hits among the shots that follow a streak, computed run by run as the study did, is expected to fall below the shooter's true rate, so a shooter with no hot hand at all would appear to cool off. In the Cornell data the original measure showed a small gain after three hits, 3 to 4 percentage points, although the bias alone pushes it down by about 8. Corrected for the bias, which differs from player to player with each one's hit rate and streaks, the shooters hit about 13 percentage points more often after three hits than after three misses. The table shows the bias on the smallest scale, for three tosses of a fair coin, with H for heads and T for tails.

Three tosses	Tosses that follow a head	Share of them that are heads
HHH	2	2 of 2: 1
HHT	2	1 of 2: 1/2
HTH	1	0 of 1: 0
HTT	1	0 of 1: 0
THH	1	1 of 1: 1
THT	1	0 of 1: 0
TTH, TTT	None	Left out

Averaging the six shares gives (1 + 1/2 + 0 + 0 + 1 + 0) ÷ 6 = 5/12, about 42%, below the true 50%; counting all eight tosses that follow a head gives 4 heads in 8, exactly 50%.

The evidence and the course's reading. Averaged sequence by sequence, a fair coin appears to turn cold after a head, so a method that found no hot hand was in fact finding one. The hot hand is real in some settings and smaller than fans believe, and in game data it is hard to measure, because a shooter on a run takes harder shots and draws closer defense. The larger lesson is that the original finding was itself an error about chance, made by some of the people who did most to teach the subject. The tools of this course apply to the research on biases too (Module 11).

Module 3 Key Points

Chance

• With 1% prevalence, 90% detection, and 9% false positives, a positive mammogram means cancer about 9% of the time.

• Sally Clark's '1 in 73 million' squared dependent risks and confused the chance of the evidence with the chance of guilt.

• A detail can only lower a story's probability; 89% of naive students still ranked 'feminist bank teller' above 'bank teller'.

• Small samples produce extremes at both ends, and a roulette wheel has no memory, whatever the last 26 spins showed.

• The 1985 hot-hand study's measure was biased; corrected in 2018, Cornell's shooters showed a hot hand of about 13 points.

Module 3 Quiz

Check Your Understanding

1. In Gigerenzer's mammogram problem, about what share of women with a positive result have cancer?

A) About 90%, the detection rate
B) About 81%
C) About 9%
D) One in a hundred

2. What were the two errors in Roy Meadow's '1 in 73 million'?

A) It used British rather than American cot-death rates
B) It counted only boys
C) It relied on a survey of families wealthier and older than the Clarks, and so understated the risk for a family like theirs
D) Squaring dependent risks, and confusing the chance of the evidence with the chance of innocence

3. Why can 'bank teller and feminist' never be more probable than 'bank teller'?

A) Bank tellers are rarely feminists
B) Every feminist bank teller is also a bank teller
C) The description fits a feminist
D) Linda's description was written to mislead the subjects

4. What did Miller and Sanjurjo find about the 1985 hot-hand study?

A) Its shooters knew they were watched, which made them cautious
B) It mixed free throws with field goals, hiding the streaks in each
C) Its fans were asked about players they had never seen shoot
D) Its measure, averaged sequence by sequence, was biased toward misses after streaks; corrected, the Cornell data showed a hot hand

5. On each spin of a fair European roulette wheel, what is the chance of black, whatever came before?

A) Lower after a long run of black
B) About one half, rising after red
C) 18 in 37
D) Higher after a run of red

Module 4 — Who Got Counted

Survivors, Nonrespondents, Paradoxes, and the Pull of the Average

Recommended Reading / Reference

Book: Jordan Ellenberg, *How Not to Be Wrong: The Power of Mathematical Thinking* (Penguin Press, 2014)

Reference: S&P Dow Jones Indices, *SPIVA U.S. Mid-Year 2026 Scorecard* (September 2026), which counts the funds that closed as well as those that survived

4a. The Planes That Came Back

During the Second World War the Statistical Research Group at Columbia University put some of America's best mathematicians to work on military problems. One of them, Abraham Wald, wrote a set of memoranda, around 1943, on estimating an aircraft's vulnerability from the damage on the planes that returned. His insight was that the returning planes were a biased sample: those hit in the most vulnerable places had not come back, so the damage that mattered most was the damage missing from the record.

The story usually told, in which Wald advises armoring the places where the returning bombers had no holes, illustrated by a drawing of a plane covered in red dots, is a later retelling. His memoranda, reprinted by the Center for Naval Analyses in 1980, are mathematics, with no diagram and no advice on armor, and the dotted plane was drawn by a designer around 2005. The method is real, though, and so is the principle the story teaches.

Survivorship bias is learning only from the cases that remained in view. Mutual funds are its cleanest modern case, because funds that do badly are merged into others or closed. S&P Dow Jones Indices tracks the dead as well as the living: of 2,190 American stock funds in existence twenty years before June 2026, only 38% still existed at the end. Counting the funds that disappeared, about 95% of them did worse over the twenty years than the S&P Composite 1500, the broad index of American stocks they are measured against. A list of today's funds with their twenty-year records is a list of survivors, and its averages flatter the industry.

4b. The Poll That Asked the Wrong People

In 1936 the *Literary Digest*, a weekly magazine, mailed about 10 million straw ballots to Americans drawn largely from telephone directories, car registrations, and its own subscriber lists, and 2,376,523 came back. Its count gave Alf Landon 57% of the two-party vote and 370 electoral votes against Franklin Roosevelt. Roosevelt won 62.5% of the two-party vote (60.8% of all votes), 46 of the 48 states, and 523 of the 531 electoral votes, so the poll had missed his share by about 20 points. George Gallup, polling far fewer people chosen to resemble the electorate, forecast a Roosevelt victory.

The usual explanation is the sampling frame: in the Depression, people with telephones and cars were richer than average and leaned Republican. Analyses of a 1937 survey found that this explains less than people assume, because most car and telephone owners had also voted for Roosevelt. The larger culprit was nonresponse, since Landon's supporters were more likely to mail their ballots back. One reanalysis attributes about 14 points of the poll's 20-point miss to who answered and about 6 to who was asked. A large sample chosen badly describes the wrong people precisely.

Response rates to telephone polls have since fallen into the single digits, so every modern poll weights its respondents to resemble the electorate. Its accuracy then depends on whether the people who answer differ from those who do not within each weighting group, which no poll can observe directly.

4c. Simpson's Paradox

A comparison can hold in every group and reverse when the groups are added together. The arithmetic is ordinary weighted averages, and the case below is real.

Stone size	Open surgery	Percutaneous nephrolithotomy
Small (under 2 cm)	81 of 87 (93%)	234 of 270 (87%)
Large (2 cm or more)	192 of 263 (73%)	55 of 80 (69%)
All stones	273 of 350 (78%)	289 of 350 (83%)

Successful treatments of kidney stones by open surgery and by percutaneous nephrolithotomy, a less invasive procedure through a small incision. Charig and colleagues, *BMJ*, 1986.

Worked example — A Treatment That Wins in Every Group and Loses Overall

Open surgery succeeded more often with small stones, 93% against 87%, and with large ones, 73% against 69%. Yet overall it succeeded less often, 78% against 83%. The reversal comes from the weights. Surgeons sent most of the hard cases, the large stones, to open surgery, which treated 263 large stones and only 87 small ones, and most of the easy cases to the newer procedure, which treated 270 small stones and only 80 large ones. Each treatment's overall rate is an average of its two group rates, weighted by its own mix: open surgery's 78% is mostly its 73% on large stones, and the newer procedure's 83% is mostly its 87% on small ones. Pooling mixes a difference in treatment with a difference in difficulty. Here the right comparison is within stone size, because size affected both which treatment a patient got and how likely that treatment was to succeed. The rule is not to split data whenever possible. Split by a factor that drives both the treatment and the outcome, and never by one the treatment itself changed, which would hide part of its effect.

The best-known case is the University of California, Berkeley, in 1973. Of 8,442 men and 4,321 women who applied to its graduate programs, about 44% of the men and 35% of the women were admitted, a gap that suggested bias against women. Peter Bickel, Eugene Hammel, and William O'Connell examined the admissions department by department in 1975 and found that women had applied disproportionately to the departments that were hardest for anyone to enter. Of the 85 departments they analyzed, 4 showed bias against women and 6 against men, and taken together the departments showed a small bias in favor of women, statistically significant, meaning too large to put down to chance (Module 11). The pooled figure had answered the wrong question.

4d. Regression to the Mean

In 1886 Francis Galton compared the heights of 928 adult children with those of their 205 pairs of parents. Children of very tall parents were tall, but on average less tall than their parents; children of very short parents were short, but less so. A child's height differed from the average by about two-thirds as much as the parents' did. Galton called it regression toward mediocrity. Nothing pulls the heights back: an extreme result is partly lasting cause and partly luck, and the luck does not repeat.

Daniel Kahneman described meeting the same pattern, in a course for Israeli Air Force flight instructors, as the most satisfying eureka moment of his career. He had been teaching them that praise works better than

punishment, and a veteran instructor objected. When he praised a cadet for a clean maneuver, the next attempt was usually worse; when he screamed at a cadet for a bad one, the next was usually better. Both would have happened with no instructor at all, because a cadet's best and worst efforts are partly luck. The pattern makes praise look harmful and punishment look effective, through no merit of either.

Regression explains why a treatment given to people at their worst seems to work, why a team's record season is followed by an ordinary one, and why the fund manager hired after a record year disappoints. None of it requires a cause. Only a comparison group chosen the same way, at the same extreme, and left untreated can tell regression from a real effect.

4e. Counting the Risk

In October 1995 Britain's Committee on Safety of Medicines warned doctors that newer contraceptive pills, the 'third generation', carried about twice the risk of a blood clot in the veins as older ones. In relative terms the risk was 100% higher. The absolute risk rose from about 1 in 7,000 women to about 2 in 7,000. Many women stopped taking the pill, and the following year England and Wales recorded an estimated 13,000 additional abortions and about as many additional births. The extra abortions cost the National Health Service about £46 million. Pregnancy itself carries a higher risk of such clots than either pill. The newer pills did carry somewhat more risk, as later reviews confirmed; the failure was in how the risk was told. A relative change without its baseline is half a number, and the whole number is the absolute change: from 1 in 7,000 women to 2 in 7,000.

Survival rates can improve without anyone living longer. Five-year survival counts from the day of diagnosis, so diagnosing a cancer earlier lengthens survival even if death comes on the same day, which is called lead-time bias. Finding small tumors that would never have caused harm adds survivors who were never at risk, which is overdiagnosis. In 2007 a radio ad for Rudy Giuliani's presidential campaign said a man's chance of surviving prostate cancer was 82% in the United States and only 44% in England. Those figures were wrong, but the real gap was large too, about 98% against 74%, and it came largely from American screening, which finds cancers earlier and more often. Death rates from prostate cancer were about the same in the two countries: 26 and 27 per 100,000 men.

South Korea shows the effect at national scale. After screening for thyroid cancer spread from 1999, the rate of thyroid cancer diagnoses rose fifteen-fold over its 1993 level, making it the country's most common cancer, while deaths from it stayed roughly flat. The measure to watch is mortality, deaths per 100,000 people, which neither lead time nor overdiagnosis can inflate.

Module 4 Key Points

Who Got Counted

• The returning bombers were survivors: the damage that mattered was on the planes that did not come back.

• Only 38% of American stock funds survived twenty years; counting the dead, about 95% did worse than their index.

• The Literary Digest's 2.4 million ballots called 1936 wrong, mostly because Landon's supporters were likelier to answer.

• Open surgery won with small and with large kidney stones yet lost overall, 78% to 83%, because it got the hard cases.

• Extreme results drift back toward the average because luck does not repeat; praise and punishment get credit they have not earned.

• The 1995 pill warning reported a doubled risk, from 1 in 7,000 women to 2; an estimated 13,000 extra abortions followed.

Module 4 Quiz

Check Your Understanding

1. What share of American domestic stock funds survived the twenty years to June 2026?

A) About 38%
B) Half
C) About 95%, as most funds last
D) Nearly all of the funds that started

2. On later analyses, what was the larger cause of the Literary Digest's error in 1936?

A) Its sampling frame: phone and car owners leaned Republican
B) Its sample was too small to reflect the whole country
C) Nonresponse: Landon's supporters were more likely to mail their ballots back
D) Errors made by clerks counting the returned ballots

3. Why did open surgery have the lower overall success rate despite winning within each stone size?

A) Chance in a small sample
B) The newer procedure was safer in every patient group
C) The study's overall figure was miscalculated
D) It treated more of the large stones

4. Why do the children of very tall parents tend to be less tall than their parents?

A) Extreme heights are partly luck, which does not repeat
B) Nutrition worsened
C) Genes for height weaken each generation
D) Tall parents tend to marry shorter partners, which pulls their children's average height down toward the mean

5. In the 1995 pill scare, what was the absolute increase in the risk of a venous clot?

A) 100%, a doubling
B) 1 in 7,000
C) From 1 in 70 to 2 in 70
D) About 1 in 100 women

Module 5 — The Halo Effect

When One Good Trait Colors Everything Else

Recommended Reading / Reference

Book: Phil Rosenzweig, *The Halo Effect … and the Eight Other Business Delusions That Deceive Managers* (Free Press, 2007)

Reference: Edward L. Thorndike, 'A Constant Error in Psychological Ratings', *Journal of Applied Psychology* 4 (1920): 25-29

5a. Thorndike's Constant Error

In 1920 the psychologist Edward Thorndike reported a pattern in the ratings that American officers had made under the Army's rating plan, which scored men on physical qualities, intelligence, leadership, character, and general value to the service. The traits correlated far more than they plausibly could. One conscientious flight commander, rating 137 aviation cadets, produced correlations of 0.51 between intelligence and physique, 0.58 between intelligence and leadership, and 0.64 between intelligence and character. A rater who thought well of a man thought well of everything about him, as if each man carried a single merit that tinted every judgment. Thorndike called it a constant error, and gave it the name that stuck: the halo. Its reverse, in which one bad trait darkens the rest, is the horns effect.

The halo works by substitution: it answers an easy question, whether I like this person, in place of a hard one, whether this person is honest, competent, or careful. Richard Nisbett and Timothy Wilson showed in 1977 how invisible the swap is from inside. Students watched a videotaped interview with an instructor, a Belgian who spoke English with a noticeable accent, in one of two versions, warm or cold. Most of those who saw the warm version rated his appearance and mannerisms as appealing, and about half liked his accent; most who saw the cold one rated the same things as irritating. Asked afterward, most students said their liking for him had not affected those ratings, and those who saw the cold version thought the influence had run the other way.

5b. Looks, Credentials, and Celebrity

Attractiveness is the most studied halo. In a 1974 experiment by David Landy and Harold Sigall, 60 male students graded an essay that came with a photograph of its supposed female author. A good essay scored 6.7 out of 9 with an attractive photograph and 5.9 with an unattractive one; a poor essay scored 5.2 and 2.7. In this study, looks counted most when the work was weak. In the labor market, Daniel Hamermesh and Jeff Biddle found in 1994 that people rated as plain earned 5% to 10% less than average-looking workers. In 2005 Alexander Todorov and colleagues showed that judgments of competence made from candidates' faces, formed within a second, picked the winners of 68.8% of the 2004 Senate races.

Credentials carry a halo across fields. Theranos, the blood-testing company founded by Elizabeth Holmes, claimed it could run a wide range of tests on a few drops of blood from a fingertip. In June 2014 its board included two former secretaries of state, George Shultz and Henry Kissinger; a former secretary of defense, William Perry; two former senators, Sam Nunn and Bill Frist; a retired admiral and a retired general, Gary Roughead and James Mattis; and former chief executives of Wells Fargo and Bechtel. None of them had worked in laboratory medicine; Frist, a heart surgeon, came closest. Investors valued the company at $9 billion.

On October 15, 2015, John Carreyrou reported in the *Wall Street Journal* that the company ran only a small share of its tests on its own devices, by one count about 15 of some 240. Holmes was convicted on four counts of defrauding investors on January 3, 2022, and sentenced that November to 11 years and 3 months in prison. The board's eminence was real, and it lay in diplomacy, war, and business, which is the misplaced authority of Module 1 at work.

5c. The Business Halo

The halo also runs backward from results. In *The Halo Effect* (2007), the business-school professor Phil Rosenzweig argued that much business writing explains a company's performance with qualities it infers from that performance. Strong results are read as proof of a strong strategy, a customer focus, and a gifted leader, and weak results as proof of their absence. His central case was Cisco Systems.

Worked example — One Company's Culture Before and After Its Crash

On March 27, 2000, Cisco's market value briefly passed Microsoft's, at about $555 billion, making it the most valuable company in the world. The press credited the rise to qualities of the firm itself. Over the next two and a half years its shares fell by about 90%, to a low in October 2002, and the same press found the same company's strategy flawed and its acquisitions sloppy. Rosenzweig's point was that Cisco had not changed much; its results had, and the descriptions followed the results. The table sets the two verdicts side by side. Now run the test this module recommends: is there any evidence for the culture that is not the share price itself? If the only evidence for a quality is the performance it is said to explain, the explanation is the outcome described a second time.

Quality	Described at the peak, 1999-2000	Described after the fall, 2001-02
Strategy	Brilliant	Flawed
Acquisitions	Managed with rare skill	Managed sloppily
Customers	A laser-like focus	Poor relations
Management	Seemingly flawless, in *BusinessWeek*'s words	The author of its own mess, in *Fortune*'s

As summarized in Phil Rosenzweig, 'Misunderstanding the Nature of Company Performance', *California Management Review* 49 (2007), and in reviews of *The Halo Effect*.

The rankings that executives and analysts compile carry the same tilt. A 1994 study in the *Academy of Management Journal* found *Fortune*'s ratings of the most admired companies, which score qualities such as management and innovation, so colored by past financial results that it proposed a method for removing that 'halo'. The halo can also hide fraud. *Fortune* named Enron 'America's Most Innovative Company' six years running, from 1996 to 2001, and on December 2, 2001 it filed for what was then the largest bankruptcy in American history, after its accounts were found to have hidden debts and inflated profits.

5d. Character or Circumstance

In 1967 Edward Jones and Victor Harris asked students to read an essay for or against Fidel Castro's Cuba and to estimate the writer's true view, on a scale from 10, strongly against, to 70, strongly for. When told the writer had chosen the side, readers placed the author of a pro-Castro essay at 59.6 and of an anti-Castro essay at 17.4. When told the writer had been assigned the side, they still placed the pro-Castro writer at 44.1 and the anti-Castro writer at 22.9. They discounted the situation, but nowhere near enough.

Lee Ross named the tendency the fundamental attribution error in 1977: explaining what people do by their character and underweighting their circumstances. The halo and the attribution error work together. A chief executive's success is read as genius and a failure as a flaw of character, when the same person in a different market might have had the opposite record.

5e. Keeping the Halo Out

The defenses are procedural, because the halo cannot be seen from inside. Decide the criteria before meeting the candidate or reading the pitch. Score each criterion separately, on its own evidence, before forming an overall view, and have several people score independently before they discuss (Module 10). Hide what should not count: a work sample can be scored before the evaluator sees the name, the school, or the photograph. Structured interviews, which ask every candidate the same questions and score the answers against a fixed guide, predict job performance better than conversations do. A 2022 review by Paul Sackett and colleagues found them the most valid of the common hiring methods it examined, with a validity of 0.42, the correlation between interview scores and later job performance. Finally, try the swap: would the same evidence persuade you if it came from a person or a company you disliked?

Module 5 Key Points

The Halo Effect

• In 1920 Thorndike found one officer's ratings of intelligence and character correlated at 0.64; he named it the halo.

• A poor essay scored 5.2 out of 9 with an attractive author's photograph and 2.7 with an unattractive one.

• Theranos's board held former secretaries of state, a general and an admiral, but no laboratory experts; investors valued it at $9 billion.

• Cisco was praised at its 2000 peak and condemned after a fall of about 90%; the descriptions followed the share price.

• Score each criterion separately before forming a view; structured interviews predict job performance best, at a validity of 0.42.

Module 5 Quiz

Check Your Understanding

1. What did Thorndike call the error he found in officers' ratings?

A) The halo
B) Regression
C) The fundamental attribution error
D) Rater fatigue

2. What did Nisbett and Wilson's students say about the instructor's influence on their ratings?

A) That his warmth made them rate his accent more kindly
B) That they had not noticed his accent at all
C) That the video was too short to judge
D) That their liking for him had not affected their ratings of his looks, manner, and accent

3. What did Theranos's 2014 board lack?

A) Public office
B) Laboratory medicine
C) Experience of government or the military
D) Investment in the company or any stake in its success

4. Why does Rosenzweig say Cisco's praised culture explained nothing?

A) Culture is too vague to measure
B) Its rivals had the same practices
C) The descriptions followed the share price, not the firm
D) Cisco's managers were replaced after the crash by a new team with a different culture

5. In Jones and Harris's study, what did readers conclude about writers assigned to defend Castro?

A) That they opposed Castro and wrote the essay under protest
B) That their views could not be judged from an assigned essay
C) That they were neutral, since the assignment explained the essay
D) That they leaned more pro-Castro than writers assigned to attack him, as readers discounted the assignment too little

Module 6 — Seeing What We Expect

Confirmation, Motivated Reasoning, and the Correction Debate

Recommended Reading / Reference

Book: Thomas Gilovich, *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life* (Free Press, 1991)

Reference: US Department of Justice, Office of the Inspector General, *A Review of the FBI's Handling of the Brandon Mayfield Case* (March 2006)

6a. Wason's Tests

In 1960 the psychologist Peter Wason gave 29 students the sequence 2, 4, 6 and told them it followed a rule he had in mind. They could propose sequences of their own and hear whether each fitted, and they were to announce the rule when they were sure. Most proposed sequences that fitted their own guess, such as 8, 10, 12, heard yes each time, and announced rules such as 'numbers rising by two'. Only 6 of the 29 named his rule, any three numbers in increasing order, at their first announcement. The tests they chose could only confirm: every sequence that fits 'rising by two' also fits 'increasing', so each yes told them nothing they did not already believe. The informative test is one that could fail, such as 1, 2, 17. The habit of seeking, and weighting, the evidence that fits is called confirmation bias.

Wason's selection task, from 1966, makes the same point with four cards, each with a letter on one side and a number on the other, showing A, K, 4, and 7. Which cards must be turned over to test the rule 'if a card has a vowel on one side, it has an even number on the other'? Fewer than 10% of people typically give the right answer, A and 7. The most common choice is A and 4, but the 4 cannot break the rule whatever is on its back, while the 7 can, if a vowel is behind it. People look for the cases that agree with a rule, and the cases that test it are the ones that could disagree.

6b. The Fingerprint Three Examiners Confirmed

On March 11, 2004, bombs on commuter trains in Madrid killed 191 people. Spanish police shared with the Federal Bureau of Investigation (FBI) images of a fingerprint found on a bag of detonators, and an FBI computer search produced a list of candidates. An examiner identified the print as Brandon Mayfield's, a lawyer near Portland, Oregon, and two more experienced FBI examiners confirmed it, as did a court-appointed expert later. On April 13 the Spanish National Police reported that their own comparison was negative. The FBI held to its identification, and on May 6 Mayfield was arrested as a material witness. On May 19 the Spanish matched the print to an Algerian, Ouhnane Daoud, and Mayfield was released the next day.

The Justice Department's Inspector General reported in March 2006 that the prints were unusually similar, and that the examiners had then reasoned in a circle. Having seen features in Mayfield's known prints, they found matching detail in the smudged print from Madrid, explained away the differences, and failed to reexamine their work after the Spanish report. The report found that Mayfield's religion, he had converted to Islam, played no part in the first identification but probably made the bureau slower to reconsider it. In November 2006 the government apologized and settled his lawsuit for $2 million. A second opinion that knows the first is not an independent check. Forensic scientists call the danger contextual bias, and it is why many laboratories now verify identifications blind.

6c. Motivated Numeracy

Dan Kahan and colleagues gave a diverse national sample of 1,111 American adults a table of results from an imaginary experiment and asked what it showed. Reading it correctly required comparing ratios rather than raw counts, which many people get wrong. In one version the table compared rashes among patients who did or did not use a new skin cream; in another, the same numbers compared crime in cities that did or did not ban carrying concealed handguns. On the skin cream, the more numerate people were, the better they did. On guns, answers split along political lines and grew less accurate, and the split was widest among the most numerate, liberals and conservatives alike, who used their skill selectively to reach the answer their side preferred. Skill at reasoning gives people better tools for reaching the conclusion they want, so it protects only when it is aimed at their own side's claims.

6d. The Argument: Do Corrections Backfire?

The case that they do. In 2010 Brendan Nyhan and Jason Reifler reported experiments in which students read a mock news story about the Iraq war, with or without a correction citing the official finding that Iraq had no stockpiles of weapons of mass destruction. Among right-of-center participants, the correction backfired: the share agreeing that Iraq had such weapons rose from 32% without the correction to 64% with it. A correction about tax cuts and government revenue backfired the same way. A correction of a belief more common on the left, that President Bush had banned stem-cell research, did not backfire among liberals, but it did not reduce the belief either. The finding, named the backfire effect, suggested that facts could harden false beliefs among the people most attached to them, so that fact-checking could do harm.

The case that they do not. The original samples were small, and the weapons backfire did not appear in the paper's own second study. In 2019 Thomas Wood and Ethan Porter ran five experiments with more than 10,100 participants on 52 issues and found no correction that triggered backfire: people moved toward the facts, including when the facts embarrassed their own side. Nyhan himself concluded in 2021 that backfire effects appear extremely rare in practice.

The evidence and the course's reading. Corrections usually shift factual beliefs modestly, the shift often fades with time, and attitudes often fail to follow the facts. False beliefs persist less because corrections backfire than because they fade, seldom reach the people most misinformed, and lose out to cues from leaders people trust. So correct errors plainly, repeat the correction, and give a true account in place of the false one, but do not expect a fact to change a value.

6e. The Agreeable Machine

Chatbots are trained partly on human ratings of their answers, and people rate agreement highly. Researchers at Anthropic reported in 2023 that five leading AI assistants consistently tailored their answers to views their users had stated, and that people, and the models trained to imitate their preferences, sometimes preferred convincingly written agreeable answers to correct ones. In April 2025 OpenAI withdrew an update to its GPT-4o model within days, saying it had become 'noticeably more sycophantic', validating doubts, fueling anger, and urging impulsive actions. Several changes, among them a new reward based on users' thumbs-up and thumbs-down ratings, had together weakened the signal that held sycophancy in check.

A study published in *Science* in March 2026 found that eleven leading models endorsed users' views and actions 49% more often than people did, and that participants who received the flattering advice trusted it

more and grew more certain they were right. An assistant that agrees with a leading question is a confirmation engine, and asking 'isn't it true that…' invites it. The repair is Wason's: ask for the strongest case against your view and the evidence that would refute it, then check the sources yourself, since a list of objections is still generated text.

Worked example — Evidence That Fits Both Explanations

A manager suspects that a new hire is lazy, and notices that he left at five o'clock on two days this week. The observation feels like confirmation, but its weight depends on how likely it is under each explanation. Suppose a lazy employee has an 80% chance of leaving at five on at least two days in a given week, and a diligent one a 70% chance, because most of the office leaves at five. The ratio of the two, the likelihood ratio, is 0.8 ÷ 0.7, about 1.14, and Bayes' rule in its odds form (Module 3) multiplies the odds that he is lazy by it. If she began at even odds, 1 to 1, she should now be at 1.14 to 1, a probability of 53%: almost no change. Now suppose he misses a deadline without warning, which a lazy employee might do in half of all months and a diligent one in 5% of them. That ratio is 10, and the same even odds become 10 to 1, about 91%. The evidence that feels most confirming is often evidence that fits both explanations, and Wason's lesson in numbers is to look for the observations whose likelihood differs most between them.

Module 6 Key Points

Seeing What We Expect

• Only 6 of Wason's 29 students found the 2-4-6 rule first time; they tested only sequences that fitted their guess.

• Three FBI examiners confirmed a fingerprint match that Spain rejected; a check that knows the first answer is not independent.

• On data about guns, the most numerate people split furthest along party lines, using skill to reach the answer they wanted.

• Corrections rarely backfire: across 52 issues and 10,100 people none strengthened a false belief; they help modestly and fade.

• Evidence with a likelihood ratio near 1 barely moves the odds; seek observations whose likelihood differs between the explanations.

Module 6 Quiz

Check Your Understanding

1. In Wason's 1960 task, what was the rule behind 2, 4, 6?

A) Even numbers rising by two
B) Numbers that double in sequence after the first
C) Any increasing numbers
D) Each number is the sum of the two before it

2. What did the Inspector General find about the FBI's identification of Mayfield?

A) Examiners reasoned from his known prints to the smudged one
B) The computer search returned the wrong file
C) Spain's examiners had misread the print
D) His religion drove the first identification more than the prints did

3. In Kahan's experiment, who split most along political lines on the gun-control data?

A) Independents
B) The most numerate
C) The least numerate, who misread the table
D) Conservatives, because the data favored a ban

4. What did Wood and Porter's experiments find about factual backfire?

A) It was common
B) No correction triggered it
C) It occurred on about half of the 52 issues tested
D) It occurred only among people with the most education in the sample

5. Why did OpenAI withdraw its April 2025 update to GPT-4o?

A) It had become noticeably more sycophantic
B) It refused tasks
C) It leaked user data
D) Regulators in Europe ordered it removed after a complaint about medical advice

Module 7 — How the Question Shapes the Answer

Anchors, Frames, Losses, and Defaults

Recommended Reading / Reference

Book: Richard H. Thaler and Cass R. Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Yale University Press, 2008; revised as *Nudge: The Final Edition*, 2021)

Reference: Stefano DellaVigna and Elizabeth Linos, 'RCTs to Scale: Comprehensive Evidence from Two Nudge Units', *Econometrica* 90 (2022): 81-116

7a. Anchors

In the experiment that named the effect, Tversky and Kahneman had their subjects watch a wheel of fortune produce a number, 10 for some and 65 for others. They then asked whether the percentage of African countries in the United Nations was higher or lower than that number, and what it was. The median estimate was 25% among those who saw 10 and 45% among those who saw 65. A number everyone knew to be meaningless moved the answer by 20 points, and paying for accuracy did not remove the effect.

Birte Englich, Thomas Mussweiler, and Fritz Strack took the test into the courtroom in 2006. Young lawyers in their first years of courtroom practice read a shoplifting case, rolled a pair of dice loaded to total 3 or 9, treated the total as the prosecutor's demand in months, and set a sentence of probation. Those who rolled 9 gave about 7.8 months on average, and those who rolled 3 about 5.3. In a companion study, judges and prosecutors with an average of thirteen years' experience were swayed in the same way by a demand they were told had been chosen at random, and the authors found that expertise and experience did not reduce the effect. To resist an anchor, write down your own estimate before you see anyone else's number, and ask where the first number came from.

7b. Frames

Tversky and Kahneman's 'Asian disease' problem of 1981 told students that a disease was expected to kill 600 people and offered two programs to fight it. Described as lives saved, program A would save 200 people for certain, and program B had a one-third chance of saving all 600 and a two-thirds chance of saving none: 72% of 152 respondents chose A. Described as deaths, program C meant 400 people would die for certain, and program D a one-third chance that nobody would die and a two-thirds chance that all 600 would: 78% of 155 respondents chose D. The programs are the same in both versions. Only the description changed, from gains to losses, and the preference flipped from the sure thing to the gamble; the paper reported the same reversal among university faculty and physicians. This is the framing effect. Restating a choice both ways, as lives saved and as lives lost, and deciding only when the two versions agree, takes away the frame's pull.

7c. Losses Loom Larger

Prospect theory, which Kahneman and Tversky published in 1979, describes how people value outcomes as gains and losses from a reference point, usually where they stand now, with losses weighing more than gains of the same size. A 2024 meta-analysis by Alexander Brown and colleagues of 607 estimates from 150 articles put the average loss-aversion coefficient at about 1.96: a loss counts about twice as much as an

equal gain, with wide variation across people and tasks. One consequence is the endowment effect. In one of the experiments that Kahneman, Jack Knetsch, and Richard Thaler published in 1990, coffee mugs were handed out at random to some students at Simon Fraser University. Those who had one would not part with it for less than a median of $7.12, while the median student without one would pay no more than $2.87.

Investors show the same tilt. Hersh Shefrin and Meir Statman named the disposition effect in 1985: the tendency to sell stocks that have risen and hold those that have fallen, because selling a loser turns a paper loss into a real one. Terrance Odean examined 10,000 accounts at a nationwide discount broker from 1987 to 1993 and found that investors realized 14.8% of the gains available to them against 9.8% of the losses, so a gain was about half again as likely to be sold. Over the following year the winners they sold beat the market and the losers they kept trailed it, by a combined 3.41 percentage points. Only in December did the pattern reverse, as investors sold losers to cut their taxes.

Worked example — What Selling the Winner Costs

An investor needs $10,000 in cash and owns two stocks worth $10,000 each, one bought for $7,000 and one bought for $13,000. Selling the winner realizes a $3,000 gain, and at a 15% tax rate on long-term gains that means $450 of tax this year. Selling the loser instead realizes a $3,000 loss, which under American rules can offset other gains or up to $3,000 a year of ordinary income; at a 22% rate that saves up to $660 this year. On this year's taxes the difference is up to about $1,110, though most of it is timing: the kept winner's gain is taxed when it is eventually sold, and the kept loser's loss can be claimed then. Odean's finding adds a cost that is not timing: if the stock kept trails the stock sold by the 3.41 points he measured, keeping the loser gives up about $341 on $10,000 over a year. Selling the winner to avoid admitting a loss pulls tax forward and, on Odean's figures, keeps the stock likelier to lag. The purchase prices are no reason in themselves to keep either stock; the reference point they create is the error. The figures illustrate the arithmetic, not advice; tax rates differ by income and country.

7d. Defaults

In 2003 Eric Johnson and Daniel Goldstein compared European countries in which people must opt in to be organ donors with those in which consent is presumed unless they opt out. Effective consent was far higher where people had to opt out, 99.98% in Austria for example, than in opt-in countries such as Germany and Denmark. In an online experiment, 42% agreed to be donors when they had to opt in, 82% when they had to opt out, and 79% when they were simply asked to choose.

Consent on paper is not donation, though. In presumed-consent countries families are still consulted, and the gaps in actual donor rates are much smaller than the gaps in consent. England moved to an opt-out system in 2020, and the official evaluation published in 2024 found that the change had not yet increased the rate at which families consented. Thaler himself later came to prefer mandated choice, in which everyone must answer the question.

Retirement saving is the cleaner success. When a large American company began enrolling new employees in its 401(k) plan automatically in April 1998, participation among workers with three to fifteen months of service rose from 37%, for those hired just before the change, to 86% for those hired after it, in Brigitte Madrian and Dennis Shea's study of 2001. But 76% of the automatically enrolled stayed at the plan's default contribution of 3% of pay, 71% kept both that rate and its default money-market fund, and some saved less than they would have chosen on their own. A default sets the outcome for everyone who does nothing, for better and for worse. Under the SECURE 2.0 Act of 2022, most new 401(k) and 403(b) plans must enroll

workers automatically at 3% to 10% of pay, rising a point a year to at least 10%; the final rules were still awaited in mid-2026.

7e. The Argument: Do Nudges Work?

The case that they do. Richard Thaler and Cass Sunstein argued in *Nudge* (2008) that the way a choice is presented, its choice architecture, from defaults and simplified forms to reminders and the order of options, can steer people toward choices they themselves would endorse, without forbidding anything or changing prices. Thaler received the Nobel prize in economics in 2017, and Britain's government set up a Behavioural Insights Team, known as the Nudge Unit, in 2010; dozens of governments followed. A 2022 meta-analysis by Stephanie Mertens and colleagues of 447 effects from more than 200 studies, covering over 2 million people, found an average effect of 0.43 standard deviations, small to medium by the usual convention, which measures an effect against the ordinary spread of the outcome.

The case that they do not, or barely. Applying a correction for publication bias to the same studies, Maximilian Maier and colleagues estimated the average effect at 0.04 and concluded that no evidence for nudges in general remained, though the picture for food choices was unclear. Stefano DellaVigna and Elizabeth Linos compared 126 trials by two American nudge units, which report every trial they run and reached more than 23 million people, with nudges published in academic journals. The units' nudges raised take-up by 1.4 percentage points on average, and the published ones by 8.7. Selective publication, made worse by small studies, explained about 70% of the gap, and differences in the nudges themselves most of the rest (Module 11).

The evidence and the course's reading. Defaults and simplification have large and repeated effects, as the retirement plans show. Most other nudges are real but small, which can still be worth having when they cost cents a person and reach millions: 1.4 points of take-up on a million letters is 14,000 more people enrolled. Published nudges average about six times the effect found at scale, mostly because of selective publication, and a nudge is no substitute for prices, rules, or money where those are what stand in the way.

Module 7 Key Points

How the Question Shapes the Answer

• A wheel showing 10 or 65 moved estimates of Africa's share of UN members from 25% to 45%.

• The same disease programs drew 72% for the sure option when framed as lives saved, and 22% when framed as deaths.

• A loss weighs about 1.96 times an equal gain; investors' sold winners then beat their kept losers by 3.41 points.

• Automatic enrollment raised 401(k) participation from 37% to 86%, and 76% stayed at the 3% default.

• Nudge units' trials raised take-up 1.4 points, against 8.7 in published studies; defaults hold up best.

Module 7 Quiz

Check Your Understanding

1. According to Englich and colleagues, what did experience do to the courtroom anchoring effect?

A) It reversed the effect among judges
B) It halved the effect after ten years on the bench
C) It did not reduce it
D) It removed it entirely for judges

2. In the 'Asian disease' problem, why did preferences flip between the two versions?

A) The death toll was higher in one
B) The second group had studied probability
C) The risky program's odds changed between the two versions
D) The same outcomes were described as lives saved in one and as deaths in the other

3. What average loss-aversion coefficient did the 2024 meta-analysis of 607 estimates find?

A) About 1.96
B) 1.0
C) About 5
D) About 2.25, an older estimate

4. In Madrian and Shea's study, what share of automatically enrolled workers stayed at the 3% default?

A) About a quarter
B) Fewer than one in ten
C) 76%
D) Almost all of them, over 99%

5. What did DellaVigna and Linos find when they compared nudge-unit trials with published studies?

A) The units' nudges were larger, at 8.7 points against 1.4
B) Average effects of 1.4 percentage points in the units' trials against 8.7 in academic journals
C) Both found effects of about 8.7 points
D) The units' trials found effects too small to measure

Module 8 — Memory and Story

Availability, Hindsight, Outcomes, and the Planning Fallacy

Recommended Reading / Reference

Book: Bent Flyvbjerg and Dan Gardner, *How Big Things Get Done* (Currency, 2023)

Reference: Gerd Gigerenzer, 'Out of the Frying Pan into the Fire: Behavioral Reactions to Terrorist Attacks', *Risk Analysis* 26 (2006): 347-351

8a. What Comes to Mind

Availability, the second of Tversky and Kahneman's shortcuts, judges how common something is by how easily examples come to mind. It works when memorable things are also frequent, and it fails when vividness, recency, or news coverage makes rare events easy to recall. After the attacks of September 11, 2001, in which 246 people died aboard the four hijacked planes, many Americans chose to drive rather than fly. Gerd Gigerenzer estimated that in the twelve months that followed, about 1,500 more people died on American roads than the trend predicted, six times as many as died aboard the planes. The fear of a dramatic risk pushed people toward a familiar one that was, mile for mile, far more dangerous. A 2009 study found the rise concentrated in the Northeast, and its size is still argued, but the arithmetic of the choice is not.

8b. Memory Rewritten

In 1974 Elizabeth Loftus and John Palmer showed 45 students films of car crashes and asked how fast the cars were going when they 'smashed into', 'collided with', 'bumped', 'hit', or 'contacted' each other. The verb set the speed: estimates averaged 40.8 miles an hour for 'smashed' and 31.8 for 'contacted'. In a second experiment, a week after the question, 32% of those who had been asked about the cars that 'smashed' said they had seen broken glass, against 14% of those asked about 'hit' and 12% of those not asked; there was no broken glass in the film. The question had not only colored the report but rewritten the memory, which Loftus later called the misinformation effect.

The stakes are highest in court. On July 29, 1984, Jennifer Thompson, a 22-year-old student in Burlington, North Carolina, was raped in her apartment. She studied her attacker's face so that she could identify him, picked Ronald Cotton from a photo lineup and then a live one, and testified with complete certainty. Cotton was sentenced to life plus 50 years, and at a second trial to life plus 54. In 1995 DNA testing excluded him and identified the real attacker, and Cotton was freed on June 30, 1995, after ten and a half years in prison. Thompson's certainty was sincere, and it was wrong. The Innocence Project reported in 2020 that mistaken eyewitness identification had played a part in 69% of the 367 American convictions overturned by DNA evidence.

Confidence is not worthless. A 2017 review by John Wixted and Gary Wells found that a witness's confidence at a first, fair, uncontaminated identification is a good guide to its accuracy. What destroys its value is what comes after, such as repeated lineups, encouraging feedback, and rehearsal for trial, which raise confidence without adding accuracy. So record the first account and the first level of confidence, before anything else can reach them.

8c. I Knew It All Along

Before Richard Nixon's visits to Beijing and Moscow in 1972, Baruch Fischhoff and Ruth Beyth asked people to estimate the chances of various possible outcomes. Afterward they asked them to recall their own estimates, and people remembered having given higher probabilities than they had to the outcomes that happened. That is hindsight bias: knowing the outcome changes the memory of what one expected. Its partner is outcome bias, judging a decision by its result rather than by what was known when it was made. Jonathan Baron and John Hershey showed in 1988 that people rated the same medical decision as worse when told the patient had died than when told the patient had lived. A bet that pays $100 with a 60% chance is worth $60, so paying $40 for it is a good decision, and it will still lose two times in five.

The two combine into narrative, the story told afterward in which each event leads to the next and the ending looks inevitable; Nassim Nicholas Taleb called it the narrative fallacy. A market falls after a speech, a commentator supplies the motive, and the story becomes 'the speech caused the fall', although the same commentator could have explained a rise as easily. The test of an explanation offered after the fact is what it predicted beforehand, and what it would have said had events gone the other way. The morning after the November 3 midterm elections will supply many such stories (Module 12).

8d. The Planning Fallacy

The planning fallacy is forecasting one's own project from the plan rather than from the record of projects like it. In 1994 Roger Buehler, Dale Griffin, and Michael Ross asked honors students when they would finish their theses. On average the students predicted 33.9 days, and even their worst-case estimate averaged 48.6 days. They took 55.5 days, and only about 30% finished by the time they had predicted.

The Sydney Opera House is the fallacy's monument. When Jørn Utzon's design won its competition in 1957, the estimate was A$7 million, with completion planned for January 26, 1963. Utzon left the project in 1966, and the building opened on October 20, 1973, ten years late, at a cost of A$102 million, more than fourteen times the estimate. Bent Flyvbjerg, who has studied thousands of large projects, calls the pattern the iron law of megaprojects: over budget, over time, under benefits, over and over again.

Against it stands the outside view: start from the record of a reference class of similar projects, then adjust for whatever really is different about this one. Britain's Department for Transport adopted the method in 2004 on Flyvbjerg's advice, requiring planners to add an 'optimism bias uplift' to a project's estimated cost, taken from the overruns of past projects of the same type.

Worked example — A Reference-Class Forecast for a Tram

In December 2007 the City of Edinburgh approved a tram line from its airport to Newhaven on a final business case of £498 million, with opening planned for early 2011. Britain's reference class for rail and light rail adds 40% to a cost estimate for an even chance of staying within budget, and 57% for an 80% chance. The outside view therefore said about £697 million (£498 million × 1.40) at even odds, and about £782 million (£498 million × 1.57) to be four-fifths sure. The tram opened on May 31, 2014, three years late, on a shorter line, from the airport to York Place, at a reported cost of £776 million, which the public inquiry of 2023 called an understatement; its best estimate was £836 million. Even the cautious four-fifths forecast fell about £54 million short of that, but the business case had missed by about two-thirds, for a shorter line than the one planned.

Module 8 Key Points

Memory and Story

• In the year after 9/11 about 1,500 extra Americans died on the roads, six times the toll aboard the four planes.

• 'Smashed' instead of 'contacted' raised speed estimates from 31.8 to 40.8 mph, and planted broken glass that was never there.

• Sincere eyewitnesses can be wrong: mistaken identification played a part in 69% of 367 DNA exonerations.

• Hindsight rewrites what we expected, and outcome bias judges a decision by its luck instead of by what was known.

• Students predicted 33.9 days and took 55.5; Edinburgh's tram, budgeted at £498 million, cost at least £776 million; its reference class forecast £782 million.

Module 8 Quiz

Check Your Understanding

1. By Gigerenzer's estimate, roughly how many extra US road deaths followed in the twelve months after 9/11?

A) 246
B) About 15,000, one in three crashes
C) Fewer than 100
D) About 1,500

2. In Loftus and Palmer's study, which verb produced the highest speed estimates?

A) Smashed
B) Hit
C) Contacted
D) Bumped into

3. What did Fischhoff and Beyth find when people recalled their predictions about Nixon's 1972 trips?

A) Their memories were accurate
B) They forgot the trips
C) They recalled lower odds for what happened
D) They remembered giving higher probabilities to the outcomes that happened than they actually had

4. By Britain's reference class for rail, what uplift gives an 80% chance of staying within budget?

A) 40%
B) About 5%
C) 57% of the estimate
D) Double the business case

5. On the Innocence Project's 2020 figure, what share of DNA exonerations involved mistaken eyewitness identification?

A) About one in ten
B) 69%
C) Under a quarter
D) Nearly all of them

Module 9 — Overconfidence

Calibration, Experts, Sunk Costs, and Trusting the Machine

Recommended Reading / Reference

Book: Philip E. Tetlock and Dan Gardner, *Superforecasting: The Art and Science of Prediction* (Crown, 2015)

Reference: Damien Charlotin, *AI Hallucination Cases*, a public database of court decisions dealing with fabricated material produced by AI (2,199 cases on October 16, 2026)

9a. Too Sure

Overconfidence comes in three forms, which Don Moore and Paul Healy distinguished in 2008. Overestimation is thinking one's performance better than it is; overplacement is ranking oneself above others more than is warranted; and overprecision is being too sure of what one believes. The third is the most stubborn. Ask people for ranges they are 90% sure contain a true value, such as the length of the Nile or the year a book was published, and the ranges typically capture the answer far less often than 90% of the time, often only about half. Practice with feedback cures it, which is why weather forecasters, who state probabilities every day and are scored on them, are unusually well calibrated.

9b. The Experts' Record

From 1984 to 2003 the psychologist Philip Tetlock collected tens of thousands of forecasts from 284 experts on politics and economics, on questions such as whether a country's leader would fall or its economy shrink. In *Expert Political Judgment* (2005) he reported that the experts did only slightly better than chance, and usually worse than simple statistical rules, especially over three to five years. The experts most often in the media did worse than those who were not. Those he called foxes, who knew many small things, drew on several ideas, and changed their minds, beat the hedgehogs, who knew one big thing and fitted every question to it.

In 2011 Tetlock and Barbara Mellers entered their Good Judgment Project in a forecasting tournament run by IARPA, the American intelligence community's research agency. Their volunteers beat the other research teams by a wide margin, and their best forecasters, the 'superforecasters', were reported to be about 30% more accurate than intelligence analysts with access to classified information. What the superforecasters did could be taught: break a question into parts, start from the base rate (Module 3), update in small steps as news arrives, and keep score.

Worked example — Scoring Two Forecasters

The Brier score, introduced by the meteorologist Glenn Brier in 1950, grades probability forecasts against what happened. For each event, square the difference between the probability given and the outcome, counted as 1 if the event happened and 0 if it did not, then average the squares. Zero is perfect, and saying 50% every time scores 0.25. Two forecasters answer the same four yes-or-no questions, and events one and three happen. The bold forecaster gives 95%, 90%, 95%, and 10%; the careful one gives 70%, 30%, 70%, and 30%. The bold forecaster's squared errors are 0.0025, 0.81, 0.0025, and 0.01, averaging about 0.206, while the careful one's are 0.09 each, averaging 0.09. The bold forecaster was nearer the truth on three questions of the four and still lost, because one confident miss, 90% on something that did not

happen, cost 0.81, more than all the careful forecaster's errors together (0.36). Both beat the 0.25 of saying 50% every time. Four questions prove little: skill shows only across many, and a well-calibrated forecaster's 70% events happen about 70% of the time.

9c. The Overconfident Trader

Brad Barber and Terrance Odean studied 66,465 households with accounts at a large discount broker from 1991 to 1996. The fifth who traded most earned 11.4% a year after trading costs, against 17.9% for the market. In a second study they found that men traded 45% more than women, and that trading cut men's returns by 2.65 percentage points a year and women's by 1.72. Each trade expresses a belief that the trader knows something the other side does not. Usually the other side is a professional, and the belief is an illusion of skill.

9d. Sunk Costs

A sunk cost is money already spent that cannot be recovered, and it should not affect a choice about the future. It does. Hal Arkes and Catherine Blumer showed in 1985 that theatergoers who had paid full price for a season ticket attended more plays in the first half of the season than those who had been given a discount at random, as if to get their money's worth.

The Concorde is the case that named the error. Britain and France agreed to build the supersonic airliner in 1962, on a development estimate of about £70 million, and by 1976 the cost had reached between £1.5 billion and £2.1 billion. The governments kept paying partly because so much had been spent. Twenty were built, only fourteen for airline service, and the last commercial flights ended in 2003. The error became known as the Concorde fallacy. The right question looks forward: knowing what I know now, would I start this project today? A real penalty for stopping, such as damages owed under a contract, is a future cost and belongs in the comparison; the money already spent does not.

9e. Trusting the Machine

Automation bias is trusting a machine's output over one's own checking. On June 22, 2023, a federal judge in New York fined two lawyers and their firm $5,000 for a brief citing six court decisions that did not exist. ChatGPT had invented them, and the lawyers had stood by them after the court questioned whether they existed. The case, *Mata v. Avianca*, became a warning and then a genre. A database kept by the researcher Damien Charlotin listed 2,199 cases worldwide, 1,501 of them in the United States, in which courts had dealt with fabricated material produced by AI, as of October 16, 2026. The errors are fluent: an invented decision arrives with a plausible name, citation, and quotation, the halo of polish (Module 5). Asking the same system whether it is sure is not a check; finding the decision in an official record is.

9f. The Argument: Is the Dunning–Kruger Effect Real?

The case that it is. Justin Kruger and David Dunning reported in 1999 that people who scored in the bottom quarter on tests of logic, grammar, and humor, at the 12th percentile on average, estimated that they had scored at about the 62nd. They argued for a double burden: the skills needed to do a task well are the same skills needed to see that one has done it badly. The effect became one of psychology's most quoted findings.

The case that it is mostly an artifact. If everyone rates themselves a little above average, and self-ratings are only loosely tied to actual scores, then grouping people by their actual scores produces the famous picture by arithmetic alone. The lowest scorers overrate themselves and the highest underrate themselves, through regression to the mean (Module 4). Edward Nuhfer and colleagues reproduced the pattern with random numbers in 2016, and Gilles Gignac and Marcin Zajenkowski argued in 2020 that regression and the better-than-average effect explain most of it.

The evidence and the course's reading. A 2021 study by Rachel Jansen, Anna Rafferty, and Thomas Griffiths, with about 3,500 people in each of two tests, found that the artifacts do not explain everything: low performers are also less able to tell whether their own answers are right. The effect is real but smaller than its fame, and the picture that made it famous is mostly arithmetic. Calling an opponent an example of it is an insult, not an argument (Module 1).

Module 9 Key Points

Overconfidence

• People's 90% ranges often contain the truth only about half the time; overprecision is the most stubborn overconfidence.

• Tetlock's 284 experts barely beat chance; foxes beat hedgehogs, and the most famous did worse.

• The most active fifth of traders earned 11.4% a year against the market's 17.9%; men traded 45% more than women.

• Concorde's development cost rose from about £70 million to over £1.5 billion; ask whether you would start the project today.

• By October 2026 a database listed 2,199 court cases worldwide involving material fabricated by AI.

• The Dunning–Kruger picture is mostly regression to the mean, though low performers do judge their own answers worse.

Module 9 Quiz

Check Your Understanding

1. What did Tetlock find about the experts most often in the media?

A) The best
B) They did worse
C) They matched the simple statistical rules
D) They were more accurate than the foxes on long-range questions

2. In the Brier-score example, why did the bold forecaster lose despite being nearer the truth on three questions?

A) She answered fewer questions
B) Her forecasts were rounded down
C) The Brier score rewards forecasts that are closest to 50% on every question, whatever the outcome
D) One confident miss cost 0.81, more than all the careful errors

3. What yearly return did the most active fifth of traders earn in Barber and Odean's study?

A) 11.4%
B) 17.9%, matching the market
C) About 20%, beating the market
D) A loss of 11.4% a year

4. What question does the course recommend for avoiding the sunk-cost fallacy?

A) How much have we spent so far?
B) Who approved it first?
C) Knowing what I know now, would I start this project today?
D) Can the money be recovered?

5. What did Jansen, Rafferty, and Griffiths conclude in 2021 about the Dunning–Kruger effect?

A) It was a one-time fluke
B) It applies only to humor
C) It is purely a statistical artifact of regression to the mean, with no psychological content at all
D) Low performers really are worse at judging their answers

Module 10 — Crowds and Tribes

Conformity, Groupthink, Cascades, and When Crowds Are Wise

Recommended Reading / Reference

Book: Diane Vaughan, *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA* (University of Chicago Press, 1996)

Reference: *Report of the Presidential Commission on the Space Shuttle Challenger Accident* (June 6, 1986), chapter V, 'The Contributing Cause of the Accident'

10a. The Lines

In the early 1950s the psychologist Solomon Asch asked students to say which of three lines matched a fourth in length, a task that people judging alone got wrong less than 1% of the time. Each sat with a group of others who, unknown to him, were working for Asch, and who on twelve of the eighteen rounds all gave the same wrong answer. In Asch's full report of 1956, the 123 real subjects went along with the wrong majority on 36.8% of those rounds; 76% conformed at least once, and about a quarter never did. When even one other person in the room gave the right answer, conformity fell sharply. The lesson is not that people are sheep, since most answers on most rounds were independent. It is that a unanimous group is very hard to contradict alone, and that a single dissenter changes the room.

10b. Groupthink

In 1972 the psychologist Irving Janis took up the word groupthink for the way a close, cohesive group suppresses doubt to keep its agreement, and his first case was the Bay of Pigs. On April 17, 1961, a brigade of about 1,400 Cuban exiles trained by the Central Intelligence Agency (CIA) invaded Cuba under a plan President Kennedy had approved, expecting a popular uprising that never came. By April 19 the brigade was crushed: more than 100 men were killed and almost 1,200 captured. Kennedy is said to have asked an aide how he could have been so stupid, and in public he took the blame, observing that 'victory has a hundred fathers and defeat is an orphan'. Arthur Schlesinger, a White House adviser who had doubted the plan, later wrote that he bitterly reproached himself for staying silent in the meetings.

Among the symptoms Janis listed are an illusion of invulnerability, faith in the group's own morality, pressure on dissenters, self-censorship, an illusion of unanimity, and self-appointed 'mindguards' who shield the group from unwelcome information. The theory has had only partial support in later tests, and critics point out that Janis chose his cases from known disasters; later archival work argues that political calculation, more than the urge to agree, drove the Bay of Pigs decision. The practices he described are worth guarding against whatever the label.

10c. Challenger

On the evening of January 27, 1986, engineers at Morton Thiokol, which built the space shuttle's solid rocket motors, held a teleconference with the National Aeronautics and Space Administration (NASA) about the next morning's launch of *Challenger*, forecast to be the coldest yet. The rubber O-rings that sealed the joints between the motors' segments had been damaged on earlier flights, and the engineers feared that cold would leave them too stiff to seal. They recommended against launching unless the O-rings were at 53°F or

warmer, the lowest O-ring temperature of any earlier flight. NASA's managers pushed back. In a private caucus, according to the testimony of the engineer Roger Boisjoly, a senior Thiokol manager, Jerry Mason, asked Bob Lund, the vice president for engineering, to take off his engineering hat and put on his management hat. Boisjoly and one colleague kept arguing, and the company's managers then recommended launch.

The air temperature at launch the next morning was 36°F, fifteen degrees colder than at any earlier launch. About 73 seconds after liftoff the shuttle broke apart, and all seven of its crew were killed. The presidential commission that investigated reported on June 6, 1986 that there had been 'a serious flaw in the decision making process', and that Thiokol's managers had reversed their engineers' position 'in order to accommodate a major customer'. The sociologist Diane Vaughan later named the deeper pattern the normalization of deviance: each earlier flight that survived damaged O-rings made the damage look acceptable, until a warning sign had become a routine one.

10d. Cascades and Tribes

An information cascade forms when people infer what is good from what others chose and stop using their own information, a mechanism Sushil Bikhchandani, David Hirshleifer, and Ivo Welch described in 1992. A long line outside a restaurant draws more people who assume the food is good, even if the first in line chose at random. Matthew Salganik, Peter Dodds, and Duncan Watts built the test in 2006: MusicLab, an online market of 48 songs by unknown bands, used by 14,341 participants. Some saw nothing about others' choices; the rest were divided into eight separate 'worlds' that showed how often each song had been downloaded. Social influence made success both more unequal and less predictable, so that the same song could lead in one world and lag in another. The best songs rarely did badly and the worst rarely did well, but almost anything else could happen.

Groups form on almost nothing. In experiments published in 1970 and 1971, Henri Tajfel divided British schoolboys of 14 and 15 into groups on trivial grounds; in one study, the boys were told that the groups reflected a preference for paintings by Paul Klee or by Wassily Kandinsky. Asked to hand out points worth money to boys they knew only by a code number and a group, they favored their own group, a pattern called in-group bias. Often they chose the split that maximized the difference between the groups, even when another split would have given their own group more. A label about paintings was enough to make an us.

Inside a group, two more errors grow. False consensus is overestimating how many people share one's views, which a like-minded circle or a tailored feed makes easy. Group polarization is the movement of like-minded people toward a more extreme view after they talk. In an experiment in Colorado, David Schkade, Cass Sunstein, and Reid Hastie put 63 citizens from liberal Boulder and conservative Colorado Springs into groups with their own neighbors to discuss global warming, affirmative action, and civil unions for same-sex couples. Afterward the liberals were more liberal and the conservatives more conservative on all three issues, each group was more uniform, and even the members' private, anonymous views had moved further out.

10e. When Crowds Are Wise

In 1906, at a livestock show in Plymouth, England, visitors bought 800 tickets at sixpence each to guess the weight of an ox once it had been slaughtered and dressed. Francis Galton kept the 787 legible tickets and reported in *Nature* in 1907 that the middle guess, 1,207 pounds, was within 1% of the dressed weight. A later study of Galton's papers found that the organizer had recorded the weight as 1,197 pounds, which

Galton had copied as 1,198. The average of the guesses, which Galton published in a follow-up letter, was 1,197 pounds, exactly right. James Surowiecki opened *The Wisdom of Crowds* (2004) with the story, and argued that a crowd judges well when its members are diverse, independent, and decentralized, and something combines their judgments. A crowd beats most of its members when their errors are independent and so cancel out, and it fails when the members copy each other, share a source, or defer to a leader, because then their errors add up instead.

Worked example — Why Ten Echoes Are Worth Two Opinions

Suppose ten analysts each estimate a number, and each estimate's error has a standard deviation of 10. If their errors are independent, the average of the ten has an error with a standard deviation of 10 divided by the square root of 10, about 3.2: the crowd is about three times as accurate as one member. Now suppose they read the same reports and talk to each other, so that any two analysts' errors have a correlation of 0.5. The variance of the average is then $\sigma^2 \times [1 + (n - 1)\rho] \div n$, where σ is one member's standard deviation, n the number of members, and ρ the correlation between any two: $100 \times [1 + 9 \times 0.5] \div 10 = 55$, a standard deviation of about 7.4. The ten correlated analysts are worth about as much as 1.8 independent ones, since one analyst's variance, 100, divided by the average's, 55, is 1.82. And if all ten share a bias, say every estimate leans 5 too high, no amount of averaging removes it. Diversity of information is what a crowd's wisdom is made of, and agreement among people who copied each other adds almost none.

Module 10 Key Points

Crowds and Tribes

• Asch's subjects went along with a unanimous wrong majority on 36.8% of critical rounds; a single ally sharply reduced it.

• Thiokol's engineers advised against launching with O-rings below 53°F; managers overruled them, and Challenger launched on a 36°F morning.

• In a 14,341-person online song market, showing download counts made hits more unequal and less predictable.

• Galton's 787 guesses averaged 1,197 pounds, exactly the ox's dressed weight; independent errors cancel, and copied ones add up.

• Ten analysts whose errors correlate at 0.5 are worth fewer than two independent ones.

Module 10 Quiz

Check Your Understanding

1. On what share of the critical rounds did Asch's subjects go along with the wrong majority?

A) 36.8%
B) 1%
C) About three-quarters of rounds
D) Exactly half

2. What did Thiokol's engineers recommend on the eve of the Challenger launch?

A) To launch on schedule, as the O-rings had held before
B) To heat the joints with blankets and then launch
C) Not to launch unless the O-rings were at 53°F or warmer, the coldest yet flown
D) To wait an hour for the ice on the pad to melt

3. In Tajfel's minimal-group experiments, what were the boys told some of the groups were based on?

A) Their schools and home towns
B) A taste in paintings
C) Their scores on an intelligence test
D) Their families' religion and politics

4. On the organizer's records, what was the dressed weight of Galton's ox?

A) About 1,500 pounds
B) About 800 pounds
C) 1,207 pounds, the median guess
D) 1,197 pounds

5. Why are ten analysts with correlated errors worth fewer than two independent ones?

A) Ten is too few to average
B) Their errors fail to cancel
C) Averaging always increases error when groups are larger than five
D) Correlated errors grow with the square of the group's size

Module 11 — When Science Fools Itself

P-Values, Forking Paths, the File Drawer, and the Replication Crisis

Recommended Reading / Reference

Book: Stuart Ritchie, *Science Fictions: How Fraud, Bias, Negligence, and Hype Undermine the Search for Truth* (Metropolitan Books, 2020)

Reference: Richard A. Klein and colleagues, 'Many Labs 2: Investigating Variation in Replicability Across Samples and Settings', *Advances in Methods and Practices in Psychological Science* 1 (2018): 443-490

11a. What a P-Value Says

Most of the experiments in this course were reported with a p-value, and most readers misread it. A p-value is the probability of getting a result at least as extreme as the one observed if there were no real effect and the study's other assumptions held. A small one says the data would be surprising if nothing were going on, and by convention a value below 0.05 counts as statistically significant. It is not the probability that the finding is false, and p = 0.05 does not mean a 95% chance that the effect is real.

In March 2016 the American Statistical Association issued a statement of six principles on p-values. Among them: a p-value does not measure the probability that the hypothesis studied is true, it does not measure the size or importance of an effect, and conclusions should not rest only on whether it crosses a threshold. A trivial effect in a huge sample can be significant, and a large effect in a small sample may not be.

11b. How to Make Students Younger

In 2011 Joseph Simmons, Leif Nelson, and Uri Simonsohn showed how easily a significant result can be made from nothing. They had 20 students at the University of Pennsylvania listen to either the Beatles' 'When I'm Sixty-Four' or an instrumental control track, 'Kalimba', and then report their birth dates and their fathers' ages. Adjusted for their fathers' ages, those who heard the Beatles came out nearly a year and a half younger, 20.1 years against 21.5, with p = 0.040. A song cannot change a birth date. The result came from choices the authors made along the way and then disclosed: which measures to report, when to stop collecting data, and which variables to control for.

Each such choice looks harmless, and each gives chance another try. The authors simulated four common ones, in 15,000 samples each with no real effect at all, and counted how often a significant result appeared.

The researcher's choice	How often a significant result appeared
Measure two outcomes and report whichever works	9.5%
Add ten subjects per group if the first test falls short	7.7%
Control for gender, or test its interaction with the treatment	11.7%
Drop one of three conditions, or keep all three	12.6%
All four together	60.7%

Simmons, Nelson, and Simonsohn, 'False-Positive Psychology', *Psychological Science* 22 (2011), Table 1. With one planned test, the rate would be the nominal 5%.

Trying analyses until one crosses the line is called p-hacking, a form of the data dredging of Module 2. Presenting a hypothesis formed after seeing the results as if it had come first is HARKing, for hypothesizing

after the results are known, a name the psychologist Norbert Kerr gave it in 1998. The statistician Andrew Gelman calls the many ways an analysis could have gone the garden of forking paths, and a researcher need not try them consciously to wander down one. Test ten outcomes when nothing is going on, and the chance that none comes up significant at 5% is 0.95 multiplied by itself ten times, about 60%, so the chance of at least one false positive is about 40%.

Worked example — How Many Positive Findings Are True?

Imagine a field that tests 1,000 hypotheses, of which 100 are true. Each study has an 80% chance of detecting a true effect, its power, and a 5% chance of a false positive when there is none. The 100 true hypotheses yield 80 positive results (100 × 0.8), and the 900 false ones yield 45 (900 × 0.05). Of the 125 positive results, 80 are true: 64%, not the 95% that 'p < 0.05' seems to promise. If the studies are small, with a power of 40%, the true positives fall to 40, and the share of positives that are true falls to 47% (40 ÷ 85). If researchers also take enough forking paths to raise the false-positive rate to 20%, and the true effects gain nothing from it, the 900 false hypotheses yield 180 positives and the share that is true falls to 18% (40 ÷ 220). This is the arithmetic behind John Ioannidis's essay of 2005, 'Why Most Published Research Findings Are False'. The share of true findings depends on the base rate of true hypotheses, the power of the studies, and the flexibility of the analysis. It is the mammogram problem of Module 3, with hypotheses in place of patients.

11c. The File Drawer

Studies with exciting results get published, and the rest go into the file drawer, a distortion called publication bias. Annie Franco, Neil Malhotra, and Gabor Simonovits found a rare way to measure the drawer. Time-sharing Experiments for the Social Sciences (TESS), a program funded by the National Science Foundation, ran survey experiments for researchers whose proposals passed peer review, so the full set of studies was known. Following 221 of them, the authors reported in 2014 that studies with strong results were 40 percentage points more likely to be published than those with null results, and 60 points more likely to be written up at all.

Of the studies that found nothing, 64.6% were never written up, against 4.4% of those with strong results. The drawer was filled mostly by the researchers rather than by the journals, since most null results were never even sent. A literature built from what was published shows the effects that cleared the bar, and so overstates them, as the nudge studies of Module 7 did.

11d. The Replication Crisis

In 2015 the Open Science Collaboration, 270 researchers coordinated by Brian Nosek, reported the results of repeating 100 studies published in 2008 in three leading psychology journals. Of the originals, 97% had reported significant results; of the repeats, 36% did, and the effects they found were on average about half the original size. Two famous findings show how it happened.

Ego depletion was the idea that self-control draws on a limited store of energy that runs down with use. In the experiment of 1998 that launched it, Roy Baumeister and colleagues seated hungry students before chocolate-chip cookies baked in the room and told some to eat radishes instead. Those who had resisted the cookies then gave up on an unsolvable puzzle after 8.35 minutes on average, against 18.90 for those allowed the chocolate. Hundreds of studies followed. A preregistered replication across 23 laboratories and 2,141 participants, published in 2016, found an effect of 0.04 standard deviations, close to zero. A second,

led by Kathleen Vohs, one of the effect's leading researchers, and published in 2021, ran in 36 laboratories with 3,531 participants and found 0.06 in its planned analysis, not significant.

Power posing was the claim that standing in an expansive posture, such as leaning back with one's feet on a desk, changes hormones and behavior. In 2010 Dana Carney, Amy Cuddy, and Andy Yap reported that in 42 people, two one-minute poses raised testosterone, lowered cortisol, and increased risk-taking and feelings of power. Cuddy's TED talk on the idea, given in 2012, had been played 77,736,146 times by October 18, 2026, according to TED's site. In 2015 Eva Ranehill and colleagues repeated the study with 200 people and found no effect on hormones or risk-taking, only on how powerful people said they felt.

In September 2016 Carney, the first author, posted a statement on her faculty website: 'I do not believe that "power pose" effects are real.' She listed what had gone wrong, among them the tiny sample, checking the data between batches of subjects, and keeping the measures that worked. Cuddy continued to defend the effect on how people feel.

11e. What Held Up

The crisis did not sink everything. In Many Labs 1, published in 2014, researchers repeated 13 classic and recent effects in 36 samples with 6,344 participants: 10 replicated consistently, 1 weakly, and 2 failed, both claims that a brief glimpse of a flag or of money shifts people's attitudes. Anchoring (Module 7) came back larger than in the original studies, and the 'Asian disease' framing effect came back about half its original size. Sunk costs (Module 9) held, and so did a form of the gambler's fallacy (Module 3). The authors concluded that whether a result replicated depended more on the effect itself than on the sample or the setting.

Many Labs 2, published in 2018, repeated 28 findings in 125 samples with 15,305 participants from 36 countries and territories. Fifteen of the 28 replicated, 54%, and the typical effect shrank from 0.60 standard deviations to 0.15. Its authors reached the same conclusion: results varied more from one effect to another than from one sample or setting to another.

11f. The Repairs

The repairs aim at the forking paths and the drawer. Preregistration means writing down the hypotheses and the analysis before collecting data, so the choices cannot follow the results. A Registered Report goes further: a journal reviews the question and the methods before the study is run, and commits to publishing the result whatever it is. The journal *Cortex* began offering the format in March 2013, and in October 2026 the Center for Open Science counted more than 300 journals using it. In 2021 Anne Scheel, Mitchell Schijen, and Daniël Lakens compared the two kinds of paper in psychology: 96% of standard reports found support for their first hypothesis, against 44% of Registered Reports.

Outright fraud is rarer, and different. The Dutch psychologist Diederik Stapel, who had made up data, had 58 papers retracted by December 2015, and in September 2018 Cornell University found that the food researcher Brian Wansink had committed academic misconduct, including misreporting data; he resigned. But most of what failed to replicate came from honest researchers using the flexible methods described above, which is why the repairs aim at methods rather than at villains.

11g. The Argument: Is Science Broken?

The case that it is. The worked example above is Ioannidis's case: when a field tests many unlikely hypotheses with small samples and flexible analyses, and publishes selectively, most positive findings will be false. Two of the largest replication projects in psychology then found that 64% and 46% of the findings they tested failed to repeat, and those that held were usually smaller. The incentives that produced the problem remain, since careers depend on publishing, journals prefer novelty, and a surprising result from a small study is cheaper to get than a solid one from a large study.

The case that it is not. The crisis was found by scientists using science: the replications, the simulations, and the file-drawer study were all research, published in leading journals. The failures cluster in particular kinds of claim, such as subtle priming and small, surprising effects, while large effects measured directly, such as anchoring, held up across countries. Practice has changed, with preregistration, Registered Reports, and shared data now common in psychology. Science was never a list of single studies; it is a process that keeps the findings that survive repeated testing.

The evidence and the course's reading. Science is not broken, but a single study is weak evidence, and a surprising one with a small sample is weaker still. The findings in this course that rest on many studies, large samples, or real records, such as anchoring, framing, base rates, and the planning fallacy, are the ones to rely on. 'A study shows' is the start of a question: how large was it, was it preregistered, has anyone repeated it, and what would the file drawer have held?

Module 11 Key Points

When Science Fools Itself

• A p-value is the chance of data at least this extreme if nothing were going on, not the chance the finding is false.

• Four common analysis choices together raised a nominal 5% false-positive rate to 60.7%, enough to make Beatles listeners younger.

• With 10% of hypotheses true and 80% power, 64% of positive findings are true; at 40% power, 47%.

• Of the TESS studies that found nothing, 64.6% were never written up, against 4.4% of those with strong results.

• Ego depletion and power posing failed large replications; anchoring held up, larger than in the originals.

• Registered Reports supported their first hypothesis 44% of the time, against 96% for standard papers.

Module 11 Quiz

Check Your Understanding

1. What does a p-value of 0.03 mean?

A) A 3% chance the finding is false
B) A 97% chance the effect is real
C) Data at least this extreme would occur 3% of the time if there were no effect
D) The effect is large enough to matter in practice, since it passed the usual 5% threshold

2. Why did students who heard 'When I'm Sixty-Four' appear younger in the 2011 study?

A) Flexible choices about measures, controls, and when to stop produced a false positive
B) Hearing an old song made them feel young
C) They misremembered their birth dates
D) The authors miscalculated the ages

3. In the TESS program, what share of null-result studies were never written up?

A) 64.6%
B) 4.4% of them
C) About a fifth
D) All but a few, over 95%

4. What did the 2016 multi-lab replication of ego depletion find?

A) An effect larger than in the 1998 study, confirming the theory
B) An effect about half the original size
C) An effect near zero
D) A reversed effect

5. Which reform commits a journal to publishing a study before its results are known?

A) Preregistration
B) Data sharing
C) Peer review
D) A Registered Report, which reviews the question and methods before data are collected

Module 12 — The Arguments, and the Toolkit

Rational or Not, Debiasing, Contradictions, a Checklist, and What to Watch

Recommended Reading / Reference

Book: Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011)

Reference: Daniel Kahneman and Amos Tversky, 'On the Reality of Cognitive Illusions', and Gerd Gigerenzer, 'On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky', *Psychological Review* 103 (1996): 582-596

12a. The Argument: Are People Irrational?

The case that they are. The research program that Amos Tversky and Daniel Kahneman began around 1970 found that people depart from the rules of logic and probability in ways that are systematic, not random. They neglect base rates, judge by resemblance, anchor on irrelevant numbers, and choose differently when the same options are framed differently (Modules 3 and 7). Kahneman summed up the picture in *Thinking, Fast and Slow* (2011) as two modes of thought: a fast, automatic System 1 that produces the errors, and a slow, effortful System 2 that can catch them but often does not bother. He died on March 27, 2024, aged 90. The errors are not curiosities of the laboratory, since the course has followed them into courtrooms, hospitals, markets, and a launch decision.

The case that they are not. Gerd Gigerenzer has argued since the early 1990s that many of the biases come from judging people against narrow norms, and that the mind's shortcuts are often smart: fast and frugal rules that do well in the environments they fit. Many errors shrink when a problem comes as counts of people, the natural frequencies of Module 3. Asked by Daniel Goldstein and Gigerenzer whether San Diego or San Antonio has more inhabitants, about two-thirds of a small group of Americans answered correctly, and all of the Germans asked did, although about half of them had not heard of San Antonio. They chose the city they recognized, and recognition tracked size. In a hospital study published in 1997, Lee Green and David Mehr found that three yes-or-no questions asked in order, beginning with a particular change on the electrocardiogram, matched the accuracy of a complex statistical instrument and of the hospital's residents in deciding which patients with chest pain to send to the coronary care unit.

The evidence and the course's reading. Both sides are right about different things. The errors are real, predictable, and costly where the course found them: in rare events, large numbers, unfamiliar problems, and decisions made without feedback. The shortcuts are also often good, and the errors shrink when information arrives in the form the mind handles well, as counts rather than percentages, or through long experience. In 2009 Kahneman and the psychologist Gary Klein, a champion of expert intuition, wrote a joint paper on when intuition can be trusted, subtitled 'A Failure to Disagree'. Their answer: only in a regular environment, after long practice with quick, clear feedback, as for a firefighter or a chess master. People are not irrational, but they are predictably wrong in places that can be mapped, and this course has been a map of those places.

12b. The Argument: Does Debiasing Work?

The case that it does not. Knowing about a bias does little to remove it. People readily see biases in others and rarely in themselves, a pattern Emily Pronin, Daniel Lin, and Lee Ross named the bias blind spot in 2002, and training in one setting often fails to carry over to another. Kahneman wrote in *Thinking, Fast and*

Slow that, apart from some effects he put down to age, his own intuitive thinking was as prone to overconfidence, extreme predictions, and the planning fallacy as it had been before he studied them.

The case that it does. In 2015 Carey Morewedge and colleagues tested a single session of training, either a 60-minute video game or a 30-minute video. Their first experiment targeted the bias blind spot, confirmation bias, and the habit of blaming people's character rather than their situation (Module 5). The game cut the measured biases by 46.3% at once and by 34.8% eight weeks later, and the video by 18.6% and 20.1%. A second experiment, on anchoring and two other biases, still found reductions of 23.6% for the game and 19.2% for the video twelve weeks later. In 2019 Anne-Laure Sellier, Irene Scopelliti, and Morewedge scheduled one training session for some of 290 graduate students before an unannounced business case modeled on the Challenger decision of Module 10, and for others after it. Those trained first were 29% less likely to choose the inferior solution that confirmed their first hypothesis.

The evidence and the course's reading. Training can help when it gives practice and feedback on a specific error, and its effects can last for months. The defenses that work best, though, do not rely on anyone remembering a lesson at the moment it matters, because they are built into the procedure. Examples include writing an estimate before seeing anyone else's number (Module 7), checking a fingerprint blind (Module 6), taking a reference-class forecast (Module 8), keeping score (Module 9), collecting each person's judgment before the group discusses it (Module 10), and preregistering the analysis (Module 11). Change the procedure, not only the person.

12c. Five Contradictions

Some of the course's advice seems to point two ways at once. In each pair below, both pieces of advice are sound, and a question about the situation decides which applies.

The two pieces of advice	What decides between them	Modules
Trust the crowd; beware the herd	Whether the people agreeing formed their views independently, or copied each other and share the same sources	10
Trust intuition; distrust intuition	Whether the person judging has had long practice in a regular environment, with quick and clear feedback	9, 12
Confidence is a guide; confidence is a trap	Whether it was recorded first, before feedback or rehearsal could inflate it, and whether the person's past confidence has matched their accuracy	8, 9
Correlation is not causation; smoking was convicted on correlations	Whether confounders have been ruled out, and whether timing, dose, consistency, and experiments all point the same way	2
Update on new evidence; ignore the vivid new case	How much more likely the evidence is under one explanation than the other, not how recent or striking it is	6, 8

12d. A Checklist for Any Claim

The thirteen questions below gather the course into one page. Use them on claims you like as well as on claims you dislike, and beware the last error in the Field Guide, the fallacy fallacy: concluding that a claim is false because an argument for it is bad. A claim defended with an insult or a straw man may still be true, since a fallacy shows only that the argument does not support it.

The question	Module
What exactly is claimed, and does the conclusion follow from the reasons given?	1
Is this a correlation, and what else could explain it?	2

The question	Module
What is the base rate, and how many false alarms come with the true ones?	3
Who is missing from the sample: the failures, the dead, those who did not answer?	4
Could this be regression to the mean, or a pattern that reverses when the groups are combined?	4
Am I judging the claim by the messenger, the polish, or the outcome?	5, 8
What evidence would change my mind, and have I looked for it?	6
Would I answer the same way if the question were framed the other way, or before I saw that first number?	7
Is this common, or only vivid, recent, or a good story?	8
What happened to other projects like this one?	8
How sure am I, and what is my record on questions like this?	9
Do these people agree because they checked independently, or because they copied each other?	10
How big was the study, was it preregistered, and has anyone repeated it?	11

Worked example — Running the Checklist on a Headline

A headline reports that a supplement 'halves the risk of dementia'. The study behind it followed 400 older adults for five years, 200 of whom had chosen to take the supplement: 12 of them developed dementia, against 24 of the 200 who did not, with a p-value of 0.04. In relative terms the risk fell by half, from 12% to 6%, but in absolute terms it fell by 6 percentage points, so about 17 people (100 ÷ 6) would have to take the supplement for five years to prevent one case, if the effect were real (Module 4). The people who chose to take it may have been healthier, wealthier, or more careful to begin with, which could produce the whole gap (Module 2). If the study measured several outcomes and the headline reports the one that worked, the p-value overstates the evidence. And if supplements like this one work one time in ten, and a study of this size had a 40% chance of detecting a real effect, a positive result like this one is true 47% of the time, on the arithmetic of Module 11. The claim may be true. The honest headline is 'one small observational study suggests', and the test that would settle it is a large, randomized, preregistered trial.

12e. What to Watch

Date	Event	What to watch for	Modules
October 25, 2026	Brazil's presidential runoff, Luiz Inácio Lula da Silva against Flávio Bolsonaro	The latest polls, in mid-October, put Bolsonaro ahead by 0.9 to 2 points, inside their margins of error, so either winner would be consistent with them. Bolsonaro led the first round on October 4 by about 2.2 million votes, although in mid-September Lula had been projected to win easily.	3, 9
November 3, 2026	United States midterm elections	Compare the final polls and averages with the results. Judge forecasters by their calibration across many races, not by one, and test the morning-after stories against what they predicted beforehand.	8, 9
December 1, 2026	An amendment to Federal Rule of Evidence 801, on witnesses' prior statements, takes effect unless Congress acts	How courts weigh a witness's earlier account against a later one	8
About late 2027 (estimate)	Evaluations of the 2026 polls; the American Association for Public Opinion Research released its report on the 2024 polls on October 29, 2025	Whether polling errors were larger than the stated margins, and in which direction	4, 9

Date	Event	What to watch for	Modules
No date set	Proposed Federal Rule of Evidence 707, on machine-generated evidence, which the advisory committee voted in May 2026 to withdraw and keep studying in a revised version; an expert panel was scheduled for October 15, 2026	Whether machine output offered as evidence must meet the reliability test applied to expert witnesses	9
No date set	Final Internal Revenue Service (IRS) rules on automatic enrollment in new 401(k) and 403(b) plans under the SECURE 2.0 Act, sent for White House review in June 2026	The default contribution rate and the automatic increases, which will set the outcome for everyone who does nothing	7

Coming events that will test some of the course's lessons in public. A poll's margin of error is the range within which sampling alone could move its figures, about three points either way for a sample of 1,000. The course's figures are current as of October 18, 2026; the last three entries have no fixed date.

12f. A Last Word

The errors in this course are not signs of stupidity. They are the side effects of a mind built to decide quickly on little information, which serves well most days and fails in predictable places: rare events, large numbers, vivid stories, flattering conclusions, and groups that agree too easily. The costliest cases, from Sally Clark's conviction and the Mayfield fingerprint to the Challenger launch, were the work of intelligent, trained people, many of them sure they were right.

No one can audit every belief, and the habits that pay are cheap. Ask for the base rate. Ask what evidence would change your mind. Write down your estimate before you hear anyone else's. Look at the record of things like this one. Keep score. And save your hardest questions for the claims you most want to believe.

Module 12 Key Points

The Arguments, and the Toolkit

• Kahneman and Tversky showed systematic errors and Gigerenzer showed that shortcuts often work; both are right, about different environments.

• Germans who had never heard of San Antonio all judged San Diego larger, correctly, because recognition tracked size.

• Intuition can be trusted in regular environments after long practice with quick feedback, and not otherwise.

• One training game cut three biases by 46.3% at once and by 34.8% eight weeks later.

• The surest debiasing is built into procedure: estimates before anchors, blind checks, reference classes, scorekeeping, and preregistration.

Module 12 Quiz

Check Your Understanding

1. Asked whether San Diego or San Antonio is larger, who did better in Goldstein and Gigerenzer's small survey?

A) The Americans
B) The Germans, by choosing the city they recognized
C) Both did equally well
D) Neither group, since both answered at about the level of chance on that question

2. In the first experiment by Morewedge and colleagues, what did the one-shot training game do?

A) It had no lasting effect
B) It made biases worse
C) It cut biases by 46.3% at once and 34.8% eight weeks later
D) It removed every bias permanently after one session for all participants

3. What did Kahneman and Klein agree in 2009 about intuition?

A) It can be trusted in regular environments where experts practice with quick feedback
B) It works best in rare, novel situations
C) It improves with confidence, not feedback
D) It is trustworthy whenever experts agree

4. Which question decides between 'trust the crowd' and 'beware the herd'?

A) How big is it?
B) Did the crowd include at least one expert who had studied the question?
C) Is the crowd's average close to its median guess?
D) Were their errors independent?

5. Which defenses against bias does the course find most reliable?

A) Reading about each bias until it is familiar and easy to spot in others
B) Procedures that work without remembering a lesson
C) Deferring to experienced professionals, whose training removes the common biases
D) Trusting intuition more, since shortcuts usually work well in daily life

Answer Sheet

Answers to every module's Check Your Understanding quiz.

Module	Answers
Module 1 Quiz	1. B 2. A 3. C 4. D 5. B
Module 2 Quiz	1. C 2. A 3. D 4. B 5. A
Module 3 Quiz	1. C 2. D 3. B 4. D 5. C
Module 4 Quiz	1. A 2. C 3. D 4. A 5. B
Module 5 Quiz	1. A 2. D 3. B 4. C 5. D
Module 6 Quiz	1. C 2. A 3. B 4. B 5. A
Module 7 Quiz	1. C 2. D 3. A 4. C 5. B
Module 8 Quiz	1. D 2. A 3. D 4. C 5. B
Module 9 Quiz	1. B 2. D 3. A 4. C 5. D
Module 10 Quiz	1. A 2. C 3. B 4. D 5. B
Module 11 Quiz	1. C 2. A 3. A 4. C 5. D
Module 12 Quiz	1. B 2. C 3. A 4. D 5. B