BLOCK I · FOUNDATIONS OF KNOWLEDGE & REASONING

Heuristics, Biases & Rationality

Your mind takes shortcuts. The question that started a fifty-year war: are they bugs, or are they brilliant?

Meet Linda. She is 31, single, outspoken, and very bright. She majored in philosophy. As a student she cared deeply about discrimination and social justice, and she joined anti-nuclear demonstrations. Now, quickly, which is **more probable**? **A.** Linda is a bank teller. **B.** Linda is a bank teller *and* is active in the feminist movement.

Most people feel the pull of B. It fits. It tells a story. When Amos Tversky and Daniel Kahneman ran this in 1983, **more than 80%** of respondents chose B, including doctoral students in decision science who had studied probability formally. But look again. Every feminist bank teller *is* a bank teller. The group in B sits entirely *inside* the group in A. A conjunction can never be more probable than one of its parts. Answer B is not just wrong; it violates a law of logic you already know. You may have just watched your own mind break a rule it agrees with. That crack has a name, the *conjunction fallacy*[1], and prying it open is today's work.

CORE BIASES

RESOURCE-RATIONAL SYNTHESIS

WILLPOWER FUEL MODEL COLLAPSED

WHERE WE ARE

For ten days we have built the machinery of good reasoning: what counts as knowing (**Day 1**), how science filters claims (**Day 2**), the three engines of inference (**Day 3**), Bayes as the law of belief-updating and the base-rate trap (**Day 4**), and why every model is a useful lie (**Day 10**). All of that was *normative*: how a mind *ought* to reason. Today we turn the microscope on the actual instrument. Do human beings live up to the standard? And when we fall short, is that a defect to fix, or the fingerprint of a mind doing something cleverer than logic?

[1] The conjunction fallacy is judging a combined claim as more probable than one of the claims it contains, even though the combined claim is a subset.

THE PROGRAM

The shortcuts we think with

In 1974, two Israeli psychologists published a paper in *Science* that quietly rerouted a field. Tversky and Kahneman argued that people do not estimate probabilities by doing probability. Instead we reach for a handful of *heuristics*[2]: fast mental shortcuts that usually work and occasionally fail in *predictable* ways. Three did most of the heavy lifting.

Representativeness. We judge how likely something is by how much it *resembles* our mental prototype. This is exactly what snares Linda: she resembles the stereotype of a feminist so strongly that "feminist bank teller" feels like a better fit than plain "bank teller," and the feeling of fit quietly overrides the arithmetic of sets. Representativeness also makes us ignore base rates, the trap from Day 4: told that "Steve is meek and tidy and likes order," people guess *librarian* over *farmer*, forgetting that there are vastly more farmers than librarians to begin with.

Availability. We judge how common something is by how easily examples *come to mind*. After a plane crash leads the news, flying feels lethal, even though the very newsworthiness of the crash is evidence of how rare it is. Vivid, recent, emotional events are overweighted because they are easy to recall.

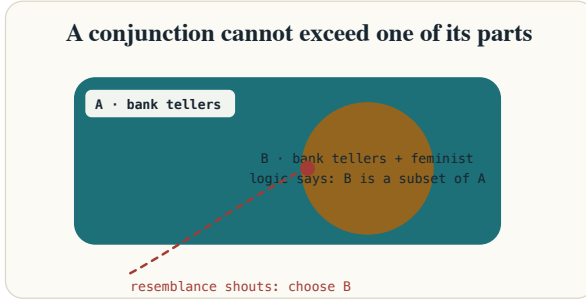
Anchoring and adjustment. When we estimate an unknown number, we latch onto whatever value is nearby and adjust from it, usually not far enough. Tversky and Kahneman spun a rigged wheel of fortune, let it stop on 10 or 65, then asked people what percentage of African nations are in the UN. Those who saw 65 guessed far higher than those who saw 10. The wheel was obviously random and had nothing to do with Africa. It anchored them anyway.

The unsettling part is not that we err. It is that these errors are as stubborn as **optical illusions**. In the Muller-Lyer illusion, two lines of equal length look unequal because of little arrowheads on their ends, and knowing the trick does not make them look equal. Kahneman's claim was that many cognitive biases are like that: not ignorance, but the automatic output of machinery you cannot simply switch off. Understanding the fallacy does not dissolve the feeling. Kahneman won the 2002 Nobel in Economics for this line of work; Tversky, who died in 1996, would surely have shared it, but the prize is not given posthumously.

Static alternate · Linda Machine

The live panel lets readers switch between the resemblance lens and the set-logic lens. The static version below shows the essential result: the vivid description makes the smaller set feel like a better match, but "bank teller and feminist activist" remains a subset of "bank teller."

[2] A heuristic is a fast rule of thumb that reduces cognitive work, often by using one cue or pattern instead of a full calculation.



The Linda description makes B feel more story-shaped, but the set relation does not change: P(A and B) can never exceed P(A).

LENS	WHAT IT SEES	VERDICT
Resemblance	Linda fits the feminist-activist story better than the plain teller story.	B feels more plausible.
Set logic	Every feminist bank teller is already inside the set of bank tellers.	B cannot be more probable than A.

THE MODEL — AND ITS CRACKS

Fast and slow, handled carefully

To organize all this, Kahneman popularized a two-character drama in his 2011 bestseller *Thinking, Fast and Slow*. *System 1* is fast, automatic, effortless, intuitive: it reads the word on a billboard, senses hostility in a voice, and blurts "feminist" at Linda before you have noticed. *System 2* is slow, deliberate, effortful: it is what you use to fill out a tax form or check whether B is really a subset of A. Biases, on this telling, happen when busy System 2 fails to audit the quick answer.

It is a wonderful teaching device. It is also, taken literally, probably wrong, and the field knows it. Three cracks are worth naming, because a good scientist keeps the ladder and kicks away the scaffolding.

There may not be two of anything. The "two systems" were always more metaphor than anatomy. Even leading proponents retreated: Jonathan Evans and Keith Stanovich, in a careful 2013 paper, abandoned the idea of two *systems* in favor of two *types of processing*, a weaker and fuzzier claim. In 2018 David Melnikoff and John Bargh went further, calling the dual-process typology "a convenient and seductive myth" that "lacks empirical support" and "systematically

thwart[s] scientific progress." The tidy binary, conscious/unconscious and effortful/automatic, does not cluster into two neat bundles.

The mental-fuel tank was a mirage. A famous companion idea, *ego depletion*[3], held that effortful self-control runs on a limited energy reserve: resist the cookies now, and you will cave to the next temptation because your willpower muscle is tired. It launched a thousand studies. Then, in 2016, a preregistered replication across 23 laboratories found the effect was essentially **zero**: a standardized effect size of about 0.04, statistically indistinguishable from nothing. This is the Day 2 replication crisis reaching into today's topic. One of the marquee findings behind effortful System 2 evaporated when tested rigorously.

So hold the model loosely. Fast versus slow remains a useful shorthand, and we will keep using it, but treat it the way Day 10 taught us to treat any model: a lossy map, not the territory. The deeper action has moved elsewhere.

THE DEBATE

The Great Rationality War

Here is where the field split into two camps that have argued, productively and pointedly, for four decades.

On one side stands the **heuristics-and-biases** tradition of Kahneman and Tversky. Its message, roughly: human intuition is riddled with systematic error. Biases are cognitive illusions, real measurable failures against the gold standard of logic and probability. The practical upshot is *meliorist*: since our minds mislead us, we should build corrections such as training, checklists, expert systems, and later, nudges.

On the other side stands Gerd Gigerenzer and the ABC Research Group in Berlin, who reply: not so fast. The "errors," they argue, are often not errors at all, and the experiments are frequently rigged, unintentionally, against the humans. They make two distinct moves, and it is worth keeping them apart.

Move one: change the format, and the fallacy fades

Gigerenzer's first argument is about **information format**, a direct callback to **Day 7**. The same content can carry very different amounts of usable structure depending on how it is encoded. Take the base-rate problems that make people, and doctors, look hopeless. In probability format:

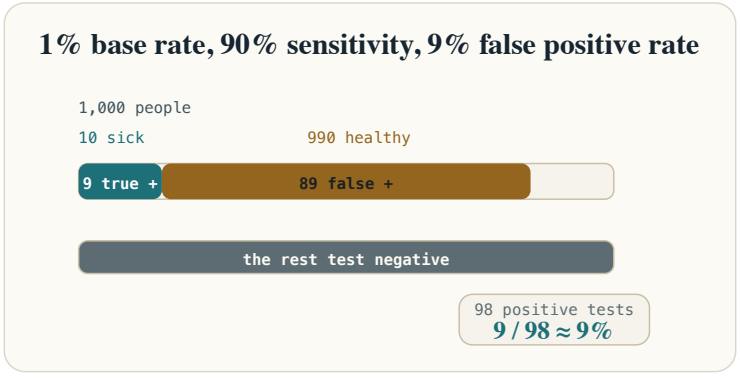
> *A disease affects 1% of women. A test detects it 90% of the time, but also gives a false positive 9% of the time. A woman tests positive. How likely is it that she is actually sick?*

[3] Ego depletion was the claim that self-control draws down a limited willpower resource, making later self-control harder.

Most people, including many physicians in study after study, answer around 80-90%. The correct answer is about **9%**. But watch what happens when you hand the mind the same facts as *natural frequencies*[4] instead of abstract percentages: *Of 1,000 women, 10 have the disease and 9 of them test positive. Of the 990 healthy women, about 89 also test positive. So of roughly 98 women who test positive, only 9 are sick.* Suddenly the answer, 9 out of 98, about 1 in 11, is almost visible. You can count it. A 2017 meta-analysis found that switching to natural frequencies roughly **quintupled** correct Bayesian answers, from about 4% to about 24% of people. Gigerenzer's point is that the base-rate fallacy is not just a fixed flaw in human wiring. It is partly an artifact of feeding the mind single-event probabilities, a format it is poorly built to digest. Change the representation and much of the irrationality dissolves.

Static alternate · Natural-Frequency Grid

The live panel changes the disease base rate and switches between probability and frequency formats. The static figure fixes the classic 1% case so the Bayesian answer can be counted.



Natural frequencies turn the Bayesian answer into a counting problem: among positive tests, the true cases are only a small part of all positives.

[4] Natural frequencies express probability information as counts in a reference class, such as 9 true positives out of 98 positive tests.

GROUP	COUNT OUT OF 1,000	POSITIVE TESTS
Sick women	10	9 true positives, 1 missed case
Healthy women	990	89 false positives, 901 true negatives
All positive tests	98	Only 9 are true cases, so the posterior chance is about 9%.

Move two: less can be more

Gigerenzer's second, bolder argument is that simple heuristics are not merely excusable. They are sometimes *better* than the fancy, information-hungry methods statisticians prefer. The banner is *fast-and-frugal heuristics*[5]: little rules that ignore most of the available information and still make excellent decisions.

The showpiece is the **gaze heuristic**. How does an outfielder catch a fly ball? Not by measuring velocity, computing a parabola, and sprinting to where the ball will land. That is intractable in real time with wind and spin. Instead the fielder fixes their gaze on the ball and runs so that the angle of gaze stays constant. Follow that one rule and you arrive where the ball comes down, no calculus required. The heuristic works not *despite* ignoring information but *because* it does: it throws away everything except the single cue that matters.

Or the **recognition heuristic**: asked which of two cities is larger, if you recognize one and not the other, bet on the one you recognize. Absurdly crude, and yet, because recognition tends to track size, it can beat elaborate models. In a celebrated finding, this even produced a "less-is-more effect": people who recognized fewer of the cities sometimes scored better, because they could use the heuristic while those who recognized everything had to fall back on shakier knowledge.

Why would ignoring information ever help? The deep answer, which Gigerenzer and Henry Brighton laid out in a 2009 paper titled *Homo Heuristicus*, is the *bias-variance tradeoff*[6], an idea we will meet again in machine learning on **Day 136**. A complex model with many parameters fits the data it trained on beautifully, but it also fits the noise, so it lurches around on new data. A simple heuristic cannot fit noise because it barely fits anything. It may be biased, but it is stable, and in a noisy, small-sample world, stability often wins.

[5] Fast-and-frugal heuristics are simple decision rules that deliberately ignore most available information and rely on a small number of useful cues.

[6] The bias-variance tradeoff is the tension between a simple model that is stably wrong in one direction and a complex model that fits noise and varies too much on new data.

THE REFRAME

Simon's scissors

Underneath the whole fight is a question about the **yardstick**. When Kahneman says people are irrational, he means: measured against the laws of logic and probability. When Gigerenzer says they are smart, he means: measured against success in the actual environment they live in. These are different rulers, and much of the war is a disagreement about which ruler is legitimate.

The person who saw this first was Herbert Simon: economist, cognitive scientist, AI pioneer, and winner of the 1978 Nobel. Simon spent the 1950s arguing that the fantasy of a perfectly rational agent who optimizes over all options is a nonstarter for any real creature. Real minds have limited time, memory, and attention. So instead of *optimizing*, we *satisfice*[7]: we search until we find an option that is good enough and stop. Simon called this *bounded rationality*[8], and he gave it an image that quietly settles the debate:

> *"Human rational behavior is shaped by a scissors whose two blades are the structure of task environments and the computational capabilities of the actor."* Simon, 1990.

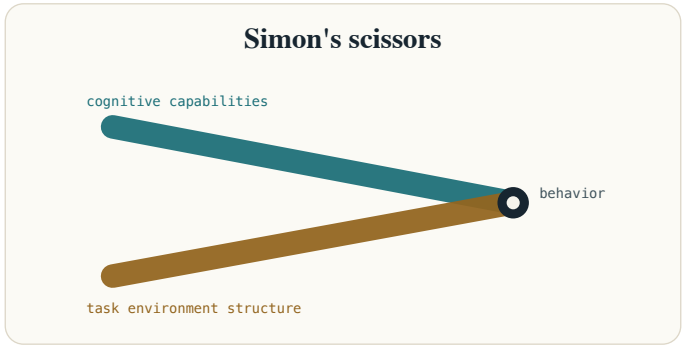
A pair of scissors cuts only when **both** blades engage. You cannot judge a blade on its own. Kahneman and Tversky studied mostly the mind blade, and, holding it up against the ruler of logic, found it wanting. Gigerenzer insists you must look at both blades together: the mind and the world it is cutting through. Once you do, yesterday's "bias" often reveals itself as a tool matched to its environment. Neither camp is simply wrong. They are inspecting different blades of the same scissors.

Static alternate · Simon's Scissors

The live diagram toggles between judging a heuristic against the mind blade alone and judging it against mind plus environment. The static figure keeps both blades visible: behavior is produced by their contact.

[7] To satisfice is to search until an option is good enough for the task, then stop instead of searching for the theoretical optimum.

[8] Bounded rationality studies rational action under limits on time, memory, attention, information, and computation.



one blade alone: the rule looks crude
both blades together: the rule may fit the world

Bounded rationality is not just the study of mental defects; it studies how limited minds and structured environments produce workable behavior together.

JUDGED AGAINST	WHAT A HEURISTIC LOOKS LIKE	EXAMPLE VERDICT
Logic alone	It ignores most of the information.	Irrational: the rule is too crude.
Mind plus environment	It exploits the structure of the task world.	Smart: the gaze heuristic solves the real-time catching problem cheaply.

THE FRONTIER · 2026

Three live edges, with the hype filter on

Every day in this course ends at the research frontier, each claim tagged for how much weight it can bear. The rationality wars did not end in a treaty; they matured into sharper empirical questions.

EDGE 01 ● ROBUST BIASES ESTABLISHED ● WILLPOWER FUEL MODEL COLLAPSED

Which biases survived the replication crisis?

The heuristics-and-biases literature was not spared the reckoning of Day 2. But the wreckage sorted itself into a revealing pattern. The perceptual-style, one-shot cognitive illusions held up; the fragile, many-moving-parts social effects often did not.

FINDING	EVIDENCE STATUS	WHAT SURVIVED
Conjunction fallacy	Established	The Linda effect replicates readily across decades, cultures, and framings, and large language models can fall for it too.
Anchoring	Established	Among the sturdier effects in psychology; now paired with computational explanations.
Ego depletion	Collapsed	The 23-lab preregistered replication found d about 0.04, effectively no effect.
Many social-priming effects	Contested	Some concept-priming results failed to replicate; Kahneman publicly warned the field.

The lesson is not "biases are not real." It is that the robust ones tend to be the ones with a mechanism, which sets up the most interesting development of all.

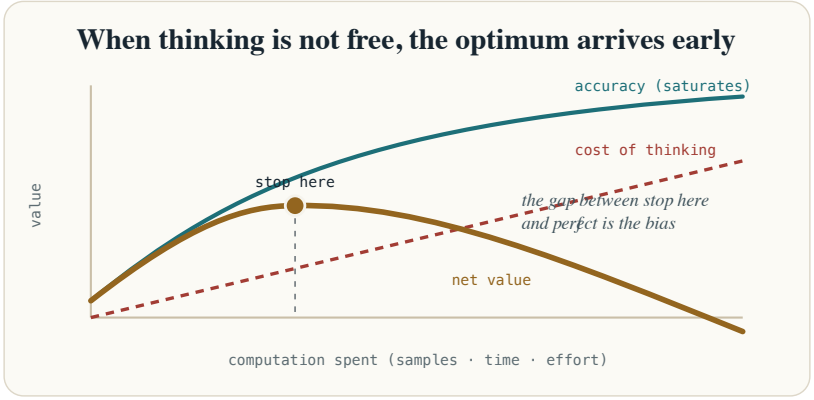
EDGE 02 ● RESOURCE-RATIONAL SYNTHESIS PROMISING ● UNIVERSAL EXPLANATION CONTESTED

The peace treaty: biases as an optimally lazy mind

The most important recent idea in this area is a genuine attempt to dissolve the war rather than win it, and it lives squarely on our computation thread. It is called *resource-rational analysis*[9], developed by Falk Lieder and Tom Griffiths in a 2020 target article in *Behavioral and Brain Sciences*. Take Simon's insight seriously and make it mathematical: assume the mind is trying to be accurate, but thinking costs a resource: time, memory, computation. Then ask what strategy an agent optimally uses given that budget. Many classic biases begin to look like cheap approximations that trade a little accuracy for a lot of speed.

The cleanest worked example is anchoring. Suppose the mind estimates an uncertain quantity the way a statistician might: by drawing samples from a probability distribution and averaging them. But each sample costs effort, so an optimal agent takes only a few and stops. If it starts sampling from a nearby value, an anchor, and stops before the samples have wandered far, its estimate stays near the anchor. That is under-adjustment: the anchoring bias, derived as the correct behavior for a mind that values its own time. The flaw is what optimal can look like when thinking is not free.

[9] Resource-rational analysis balances accuracy against costs such as time, memory, and computation.



Resource-rational analysis treats bias as a tradeoff under finite computation: thinking longer may be more accurate, but not always more valuable.

How much weight can this bear? The framework is promising and increasingly productive. It has re-derived anchoring, certain memory effects, and even a rational reinterpretation of why a mind might look like it has two systems. But keep the hype filter on. Because resource-rationality has free parameters, such as the cost, prior, and algorithm, critics warn that it risks becoming unfalsifiable, able to rationalize any behavior after the fact. That is the Day 2 demarcation worry wearing new clothes: a theory that can explain everything explains nothing.

EDGE 03 ● LLMS MIRROR SOME HUMAN BIASES ● SAME-REASONING INTERPRETATION HYPE

The machines inherited our shortcuts

Here is a twist that would have delighted Simon. When researchers gave large language models classic cognitive-psychology tasks, the models looked unnervingly like us. In a 2023 *PNAS* paper, Marcel Binz and Eric Schulz gave GPT-3 the Linda problem, and it committed the conjunction fallacy. It also showed anchoring and framing effects. Later work found that stronger models make fewer intuitive slips, but the biases do not vanish. A system trained to predict the next word in a vast ocean of human text apparently soaks up not just our knowledge but some of our reasoning reflexes.

Resist the overclaim. That an LLM reproduces the conjunction fallacy does **not** establish that it reasons like a human. Whether these systems perform genuine reasoning or sophisticated pattern-completion is exactly the debate reserved for **Day 139**. For now: the mirroring is a real, replicated, fascinating hint; the interpretation is wide open. And it hands the AI blocks, **Days 138-145**, a sharp

question: when a machine gives a fluent, confident, wrong answer, is that a flaw in the machine, or a mirror held up to us?

OPEN QUESTIONS

What's genuinely unsettled

Fifty years into the study of how minds actually reason, the honest ledger is still long:

- **Is there one correct standard of rationality at all?** Or is "rational" always relative to a goal and an environment?
- **One process or many?** Is the mind a single engine doing approximate inference under a budget, or are there genuinely distinct modes? The two-systems picture is wounded, but no agreed replacement has won.
- **Are biases to be corrected or respected?** If a bias is the optimal output of a bounded mind, debiasing it might make you worse in the real world while better on a logic test.
- **Does resource-rationality explain, or merely redescribe?** Can it be made falsifiable, or will it always find a cost function that excuses whatever people did?
- **And the question the AI block inherits:** when a system reproduces our fallacies, has it learned our reasoning, or only our residue?

The day in three sentences

Big idea

Heuristics can be useful shortcuts and predictable traps; the yardstick decides which story we tell.

Best analogy

Simon's scissors: mind and environment cut together.

Live controversy

Resource-rational analysis tries to reconcile bias research with ecological rationality.
computation · evolution · information formats · failed willpower energy · light emergence.

TOMORROW → DAY 12

Networks

Today we studied a single mind and the environment it is tuned to. Tomorrow we zoom out to the connections between minds and things: six degrees of separation, hubs and power laws, how ideas and epidemics spread, and the controversy over whether real-world networks are truly scale-free.

Sources & further reading

SOURCES

1. Tversky, A. & Kahneman, D. (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science* 185(4157): 1124-1131.
2. Tversky, A. & Kahneman, D. (1983). "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment." *Psychological Review* 90(4): 293-315. doi:10.1037/0033-295X.90.4.293. doi.org/10.1037/0033-295X.90.4.293
3. Simon, H. A. (1955). "A Behavioral Model of Rational Choice." *Quarterly Journal of Economics* 69(1): 99-118.
4. Simon, H. A. (1956). "Rational Choice and the Structure of the Environment." *Psychological Review* 63(2): 129-138.
5. Simon, H. A. (1990). "Invariants of Human Behavior." *Annual Review of Psychology* 41: 1-19.
6. Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux.
7. Evans, J. St. B. T. & Stanovich, K. E. (2013). "Dual-Process Theories of Higher Cognition: Advancing the Debate." *Perspectives on Psychological Science* 8(3): 223-241.
8. Melnikoff, D. E. & Bargh, J. A. (2018). "The Mythical Number Two." *Trends in Cognitive Sciences* 22(4): 280-293.
9. Hagger, M. S. et al. (2016). "A Multilab Preregistered Replication of the Ego-Depletion Effect." *Perspectives on Psychological Science* 11(4): 546-573.
10. Gigerenzer, G. & Hoffrage, U. (1995). "How to Improve Bayesian Reasoning Without Instruction: Frequency Formats." *Psychological Review* 102(4): 684-704.
11. McDowell, M. & Jacobs, P. (2017). "Meta-analysis of the Effect of Natural Frequencies on Bayesian Reasoning." *Psychological Bulletin* 143(12): 1273-1312.
12. Gigerenzer, G. & Goldstein, D. G. (1996). "Reasoning the Fast and Frugal Way: Models of Bounded Rationality." *Psychological Review* 103(4): 650-669.
13. Gigerenzer, G. & Brighton, H. (2009). "Homo Heuristicus: Why Biased Minds Make Better Inferences." *Topics in Cognitive Science* 1(1): 107-143.
14. Lieder, F. & Griffiths, T. L. (2020). "Resource-rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources." *Behavioral and Brain Sciences* 43: e1. doi:10.1017/S0140525X1900061X. doi.org/10.1017/S0140525X1900061X
15. Lieder, F., Griffiths, T. L., Huys, Q. J. M. & Goodman, N. D. (2018). "The Anchoring Bias Reflects Rational Use of Cognitive Resources." *Psychonomic Bulletin & Review* 25(1): 322-349.
16. Hertwig, R. & Grune-Yanoff, T. (2017). "Nudging and Boosting: Steering or Empowering Good Decisions." *Perspectives on Psychological Science* 12(6): 973-986.
17. Mertens, S., Herberz, M., Hahnel, U. J. J. & Brosch, T. (2022). "The Effectiveness of Nudging: A Meta-analysis of Choice Architecture Interventions." *PNAS* 119(1): e2107346118.

18. Maier, M., Bartos, F., Stanley, T. D., Shanks, D. R., Harris, A. J. L. & Wagenmakers, E.-J. (2022). "No Evidence for Nudging After Adjusting for Publication Bias." *PNAS* 119(31): e2200300119. With Szaszi et al. (2022), e2200732119. doi.org/10.1073/pnas.2200300119

19. Binz, M. & Schulz, E. (2023). "Using Cognitive Psychology to Understand GPT-3." *PNAS* 120(6): e2218523120. doi.org/10.1073/pnas.2218523120