Austin LW 2026-02-14: Concept Day

Useful concepts

Fenceposting: build habits by doing something on a regular, predictable schedule

Einstellung: a state of mind where the things you did before have a negative effect on what comes later; inability to set aside past solutions

Metastable state: in order to make things better you have to locally make them worse

- [related] Activation energy: you might want to be somewhere else, but there's an insurmountable barrier between here and there

Eigenstate: a state you can stay in for arbitrarily long without it changing

Phase transition: when something changes gradually but then suddenly changes qualitatively

Ersatzness: when something has been designed as an inferior substitute for something else because of a need to interface with a surrounding context originally designed for the now-missing thing (e.g. Impossible Burger)

Parsimony (YAGNI): a lot of engineering tries to front-run problems that don't exist, but in doing so creates other problems

Correlated error: wisdom of crowds works, but not if people are wrong in the same way. (E.g. some seemingly-unrelated political stances are tied together by partisanship)

Schelling point: in the absence of coordination, what people tend to converge on

Variance/bias tradeoff: as you get more evidence, bias becomes more significant

Union bound: $P(A \text{ or } B) \leq P(A) + P(B)$. So if someone is worried about a large number of distinct bad eventualities, you can reassure them that the chance of ANY of them happening is no greater than the sum of the chance of each one

Copy-neutral open source: a lot of ideas don't benefit from hoarding. People keep their ideas secret because they don't want to give away their advantage, and neglect ideas shared by others because of ego, but in both cases this inhibits progress

Shooting the moon: when you can solve a problem by making it "worse" to such an extent that the problem goes away (e.g. oil stain on shoe; incriminating photos swamped by fake photos)

Schadenfreude: deriving pleasure from another's misery

Ashby's law: you cannot control a system of complexity X without another system of complexity X+1

Levels of (evolutionary) selection: individual-level, gene-level, meme-level...

Insight porn: something that gives you a misleading impression of learning something by rearranging the way you think of your existing evidence, but doesn't actually provide any new evidence

Rivalrousness/scarcity distinction: Air is rivalrous but not scarce; intellectual property is scarce but not rivalrous

Antimeme: an idea that inhibits its own spread, that kills itself

Mental accounting: when you think of money from different sources as non-interchangeable, this leads to suboptimal use of money (e.g. using a tax refund to splurge on luxuries rather than paying down debt. but, "a dollar is always a dollar"

Moral parliament: rationalists tend to systematize their morality; but what if you don't know which system is right? Imagine a parliament with members representing each system, and what compromise they would reach. Also applies to decisions that aren't about morality (how should I live my own life)

Associative caching: make retrieving memories easier by associating them with other memories

Endowment effect: you tend to value something simply because it's in your possession already. Undermines Coase's theorem

Cash flows: better to think of money as a flow than as an amount

Importance, Tractability, Neglectedness (Effective Altruism cause prioritization framework)

Preference cascade: something was believed to be a consensus, but then someone says they disagree, and then everyone else realizes they didn't agree with the supposed consensus either. (Related: Abilene paradox)

Leaky abstraction: you try to abstract something away, but there are too many edge cases that mess it up

Concepts to be wary of

[Note: Most of these entries are 2nd-order concepts, i.e. concepts about concepts; these 2nd-order concepts are useful because they help you identify bad 1st-order concepts. But a few of these entries are themselves the bad 1st-order concepts; these are marked with a ☆ symbol]

"There are N types of people in the world"-type concepts: people may unconsciously slot themselves into one of the posited categories and alter their behavior accordingly, so the categorization makes itself real even if it wasn't before

- Stereotype threat: the opposite of that. If you know there are negative stereotypes of your category, you'll go out of your way to do the opposite of that

Goodharting: what gets measured gets optimized for

- [related] Looking under the lamp light: focusing on proxy measures

☆ Five blind men and the elephant: This story is used to say everyone is correct in their own way; but in fact they're all wrong

Axelrod tournament for iterated prisoner's dilemma: worst strategy is coinflip, but the second-worst is one that tried to be clever ("a 77-line program by a graduate student of political science whose dissertation is in game theory")

Infinite branching: research that goes down a rabbit hole and you gather lots of "knowledge" but don't accomplish your original goal

☆ Arrow's Impossibility Theorem: there is no social preference function that satisfies the three axioms of transitivity, non-dictatorship, and independence of irrelevant alternatives - but this was a dangerous idea to learn about as an edgy teenager because it made it possible to deny the very concept of social good

"Guru does everything" antipattern: I'm the smartest person in the room, so I should do everything myself, but then when I'm gone, nobody else knows what's going on

☆ Zeno's paradox: obviously false idea that motion is impossible, which you only come up with if you're thinking too much