Rifts in Rationality

By JAMES D. MILLER

Review of RATIONALITY: From AI to Zombies, by Eliezer Yudkowsky

Machine Intelligence Research Institute, 2015

How do you act rationally from "*inside a human mind*" that evolutionary selection pressures shaped for prehistoric times rather than our modern world? Eliezer Yudkowsky's massive eBook *Rationality: From AI to Zombies* seeks, in part, to answer this question after defining rationality as "*forming true beliefs and making winning decisions.*" His inexpensive eBook mostly comprises a series of posts originally appearing on LessWrong (a community blog founded by the author) and economist Robin Hanson's Overcoming Bias blog.

Our brain continually makes mental maps of the world, using them to plan decisions, to, as Yudkowsky writes, *"search out paths through probability to any desired future."* But we make systematic errors in crafting these maps. Most of us, for example, overestimate the intelligence and honesty of good-looking people, and underestimate how long it will take us to complete tasks. Fortunately, our brain, as Yudkowsky puts it, is a lens that can "*see its flaws.*" After reading about studies of cognitive biases and (this is the hard part) accepting that a bias that afflicts most people probably corrupts your thinking as well, you can correct your mental map.

A sea captain spies an island not on his map. This should cause him to have less trust in his map's accuracy. But instead he interprets a smudge on the map as the island, and uses the island's existence as justification for having more faith in his map. People become attached to their mental models of the world and loathe admitting mistakes, even to themselves, as our status-seeking brains often act as public relations departments for our personalities. Having a high IQ can actually make you more resistant to changing your mind if you show what Yudkowsky calls a "*positive bias*" and therefore deploy your intellect far more to look for evidence that supports your theories than that which might falsify them. Because of hindsight bias we even interpret past events that should be seen as having violated our mental maps as having been nearly inevitable given everything that we now know and believe, and use this false interpretation of these events as further ammunition for convincing others that we are right.

If you don't make an effort to become more willing to change your mind, rationality training, as the author admits, can make you less truth-oriented because it gives you powerful tools to dismiss arguments that don't fit your pre-existing beliefs, especially if you make isolated demands of rigor for views you oppose. (Law school can also do this to you, but in fairness to lawyers they are usually paid to persuade rather than to find the truth.)

You want to run from one place to another. But while checking a map you notice a large mountain between the locations that will slow your travel. So to speed up your journey you…erase the mountain from the map. A silly strategy, since the map is not the territory, but a strategy brains untrained in the art of rationality want to execute. Many of us think that admitting an unpleasant truth makes that truth more real, and in the extreme some of us even deny the existence of truth. For readers who don't want accurate mental maps because they feel that the truth is too horrible to contemplate, Yudkowsky quotes philosopher Eugene Gendlin, advising that we can *“stand what is true”* because we are *“already enduring it.”*

To facilitate changing your mind, the book suggests you “leave a line of retreat” with your beliefs. Chinese military strategist Sun Tzu advised giving your enemy a way to escape so they see an *“alternative to death”* and consequently won’t fight as vigorously. You can give your brain an analogous way out by imagining what the world would be like if you were wrong. Yudkowsky, for example, advised a theists to imagine a world with no souls. You can also fight positive bias by asking yourself what evidence would convince you to change your mind about a belief. If no possible evidence could convince you that your belief is wrong, than your support of this belief is not evidence-based.

In 1982, psychologists Tversky and Kahneman divided a set of experts at a forecasting conference into two groups, and asked the first to estimate the probability of:

> *"A complete suspension of diplomatic relations between the USA and the Soviet Union, sometime in 1983.”*

The second was asked to estimate the likelihood of:

> *"A Russian invasion of Poland, and a complete suspension of diplomatic relations between the USA and the Soviet Union, sometime in 1983."*

The second more detailed prediction got the higher probability estimate, even though this is the equivalent of predicting that a couple is more likely to have a son than a child. Adding a burdensome detail to a prediction should reduce our estimate of its coming true, but when predicting we often act as drama critics. Quality fiction has to tell a compelling story filled with plausible details that nicely fit together and make sense to human brains—reality is under no such obligation.

Ancient calligraphers, probably seeking to make their maps more interesting, sometimes wrote “Here be dragons” on parts of their maps representing uncharted territories. If we don’t know something, that thing (like fire-breathing flying lizards) often seems mysterious, and it feels as if this mystery is a feature of the territory. But as Yudkowsky points out “*[m]ystery exists in the mind, not in reality.”* That a question (such as “what is consciousness,” “how does the quantum wave function collapse,” or “why did the universe come into being”) feels mysterious to you isn’t evidence that the true answer would also be something you would classify as mysterious.

Yudkowsky doesn’t shy away from attacking religion, but there is one topic he and the LessWrong community mostly avoid—except for explaining why they avoid it—politics. Discussing politics often activates the ally or enemy detection subsystem of our brains. Causing your cavemen tribe to doubt your loyalty would have often been a death sentence for your ancient ancestors, so when considering politics our brains naturally tend to worry far more about signaling loyalty (and looking for signs of disloyalty in others) than about seeking truth. Politics, as they say on LessWrong, is the mindkiller. Yudkowsky proposes an interesting test for readers to see if they can rationally contemplate politics: Consider the statement:

> *"Not all conservatives are stupid, but most stupid people are conservatives."*

If, as the author writes *"you cannot place yourself in a state of mind where this statement, true or false, seems completely irrelevant as a critique of conservatism, you are not ready to think rationally about politics."* The reason is that *"reverse stupidity is not intelligence."* By the way, I've used that last phrase many times in response to when one of my students says something like: "But didn't [insert evil or stupid person] believe that?"

Beyond teaching about the brain's flaws, this book also explains how to properly use evidence according to Bayes' rule and Occam's razor. The author succeeds in lucidly explaining the first, but the second, at least for my brain, remains mysterious.

Professors like to pretend that we teach our students how to think. I have no doubt that elementary schools succeed in imparting broad transferable thinking skills when they teach basic reading and math, and professors probably excel at teaching how to think in narrow domains: my introductory microeconomic students can better understand a small set of problems after learning supply and demand analysis. But I'm unaware (which yes I admit is a fact about my mind not the territory outside of my brain) of evidence that the educational system succeeds in imparting general transferable learning skills after covering the elementary school basics. Much of the material in this book could, however, probably help improve generalize thinking by explaining how brains go wrong and how to properly evaluate evidence. This material could even be taught pre-college, although I wouldn't recommend a teacher using Yudkowsky's book for this because it would be too complex for all but the most advanced students, and the anti-religious nature of some sections might get a teacher in trouble in some places.

A few of the book's essays would be helpful for beginning teachers. In prehistoric times adults mostly held the same mental maps. Even if, say, you knew of some oasis I didn't, you would be able to quickly explain to me where it was because we would share a generalized knowledge of our local territory. Back then, adults were almost always within a few inferential steps of each other. And if you couldn't quickly explain something to me, it probably meant that I was an idiot. Our brains still seem to intuitively expect others to quickly understand us, often not taking into account the vast differences in individual knowledge that come about because of the great diversity of experiences and learning (unknown in prehistoric times) that modern humans have. For example, supply and demand analysis feels incredibly obvious to me, and it seems like I should be able to explain it to anyone in under ten minutes, although having taught this material to smart students for many years now, I've repeatedly falsified this emotional expectation.

Expecting short inferential distances combined with positive bias greatly exasperates political conflicts. Imagine that Jeb gets all of his news from red team sources that emphasize stories that support red tribal views, and Hillary gets likewise from the blue media. An opinion that seems obviously correct given Jeb's news stream might seem absurd given Hillary's. If their mental maps mistakenly hold that everyone (as in prehistoric times) has access to the same set of facts they will attribute disagreement to either gross stupidity or willful blindness. The situation will worsen if each lives in an ideological bubble in which most of the people they associate with share their political beliefs, for then the dissenters will feel like stupid and evil others.

This book has great value for anyone interested in what's being called the effective altruism movement. Human emotions cause many people to level-up on happiness through helping others. But many of us can get just as much of that "warm fuzzy feeling of goodness" from giving to a charity that efficiently provides anti-malaria nets to poor African children as to, say, a Society for Curing Rare Diseases in Cute Puppies.

Most of us would garner great satisfaction from saving a human life, and even more from saving a thousand people. But because mankind evolved in small hunter-gatherer bands, our brains have trouble emotionally grasping the real thousand-fold difference between saving one life and a thousand, meaning we suffer from scope insensitivity. The charities hardest hit by this psychological inability to multiply the good done by a charity for a person times the number of people helped are those, such as the one started by Yudkowsky, that attempt to save all of mankind.

Before telling you about Yudkowsky's charity I need to warn you about the absurdity heuristic. As most things that seem absurd are absurd, it's usually reasonable to not bother deeply considering absurd-sounding beliefs. (This must have been especially true of cavemen who rarely encountered truly novel correct beliefs.) Yudkowsky's charity certainly seems crazy on superficial consideration, but if you have read this far into my review, try to prevent your brain from immediately activating its absurdity heuristic and give his premise some deliberate thought. Perhaps take evaluating Yudkowsky's beliefs as a thought experiment testing whether you can carefully consider something that feels absurd to your brain.

Yudkowsky wants to save humanity from being destroyed by an unfriendly artificial super-intelligence. Here's my best attempt at a quick explanation of his hopes and fears: The human brain exists, so if, as atheist Yudkowsky believes, it doesn't have some soul-like supernatural quality then it should be possible to build intelligent, self-aware machines. But the brain consists of meat, an extremely inefficient platform for computing compared to what we think should be possible. Furthermore, evolution faces significant constraints that shouldn't bind human engineers, and it seems extraordinarily unlikely that nature succeeded in creating anything near the smartest possible creature when it birthed Albert Einstein or John von Neumann. Someday, Yudkowsky thinks, if we don't crash our high-technology economy, someone is going to create a computing intelligence significantly smarter than any person who has ever existed. And this program, being more capable than any human programmer or hardware designer, will probably quickly upgrade its own intelligence, becoming exponentially smarter than us. This artificial super-intelligence will then be able to do what it wishes to the world.

Yudkowsky's organization, previously named the Singularity Institute and now called the Machine Intelligence Research Institute, hopes to create a friendly artificial super-intelligence that will save us or somehow preempt the unfriendly artificial super-intelligence. Maybe it can create a utopia for us as well. (Full disclosure: I have given money to this organization and spoke at one of its events.) Yudkowsky strongly believes that any super-intelligence not specifically designed to be friendly towards mankind will destroy us.

Imagine that the computer super-intelligence could rearrange the molecules on earth any way it wished. Only a minuscule percentage of such arrangements would be consistent with human life, or anything we valued, with most of the other arrangements being boring, so if the super-intelligence had a randomly selected goal, we would almost certainly lose everything. Furthermore, the author believes, a super-intelligence merely indifferent to our fate would likely kill us to use the atoms in our bodies to fulfill whatever goals it had. Yudkowsky writes "The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else." So unless the super-intelligence shares our values, it will almost certainly destroy everything we care about.

Yudkowsky thinks that we couldn't count on the super-intelligence being intelligent to cause it to share our values. Drawing on evolutionary psychology, Yudkowsky explains how human values accidently arose because of evolutionary selection pressures. Any artificial super-intelligence would come up with the same value of π as we do (unless the universe is really weird), but there is no reason to think that most intelligences, or even more than a tiny percentage of non-evolved ones, would share our values.

Yudkowsky's views on artificial intelligence have attracted considerable support. For example, self-made technology billionaire Peter Thiel has made substantial donations to Yudkowsky's organizations. Furthermore, Bill Gates, Stephen Hawking, and Elon Musk have expressed fears of artificial super-intelligence that mirror Yudkowsky's own. One could, however, raise numerous objections to the author's claims.

A religious person might argue that our brains are more than mere material machines, and that we will never be able to bestow sentience on a computer, or that since morality is not arbitrary but something capable of being derived through reason, a sufficiently intelligent computer would either share our values or have superior ones that wouldn't allow for genocide. And even if the blind forces of evolution do fully account for our brain's development, evolution created us only after devoting vast amounts of time, space, and resources to the effort, far more than we are likely to put in over the foreseeable future.   One might claim, therefore, that we can't necessarily take the brain's existence as evidence that we could create human level artificial intelligences anytime in the next thousand years.

Robin Hanson, whose blog Yudkowsky once wrote for, believes that even if we do create a smarter-than-human artificial intelligence it won't be able to quickly and exponentially boost its own intelligence. True: if the level of innovation of each large technology company was determined mostly by its single smartest employee, then we would have good reason to believe that a computer with a human equivalent IQ of, say, 250 would be able to out-innovate the rest of the world. But since (outside of science fiction movies) innovation is mostly a slow and collaborative process involving massive amounts of time, space, and resources, a computer intelligence would have to start out being vastly smarter than Einstein before it could hope to surpass the level of innovation of Intel, Apple, and Google. Although, one might counter, if we did create an Einstein-level artificial intelligence and could run it on inexpensive hardware, the program's owners (or the program itself) could run millions of copies at high speed all devoted to figuring out ways of altering the code to make the program brighter, and this might be enough to quickly create a super-intelligence.

Yudkowsky wrote the posts that make up the bulk of his book because he felt the world faces a significant threat, but that mankind was not yet rational enough to see it, so he decided to try to raise the sanity waterline among a small group of potential supporters. Many of his insights are not original but come from economics, psychology, and decision making theory. Even if you don't accept his artificial intelligence arguments, his book contains much useful material that you could use to improve your mental maps.

James D. Miller is an associate professor of economics at Smith College, the author of Singularity Rising and Game Theory at Work, and the creator of the Microeconomics Podcast.