Review of[1]

The Engines of Cognition:
Essays by the Less Wrong Community
Author: Less Wrong
Publisher: Less Wrong Press
https://www.Lesswrong.com/books/2019
720 pages, Year: 2019
$30.00

Reviewer: William Gasarch gasarch@umd.edu

1 Introduction

For those who read my review of the first Lesswrong collection of essays, *A Map that Reflects the Territory* (see
https://www.cs.umd.edu/~gasarch/bookrev/FRED/lesswrong.pdf
for my review), this intro will give you a sense of what the Klingons call nlb'poH, the French call Déjá vu, and the English call Déjá vu.

Less Wrong is a forum founded by Artificial Intelligence Theorist Eliezer Yudkowsky in 2009. The stated philosophy is:

> **We are a community dedicated to improving our reasoning and decision-making. We seek to hold true beliefs and to be effective at accomplishing our goals. More generally, we work to develop and practice the art of human rationality.**

That seems to cover a lot of ground! The actual topics seem to be (1) how does one find the truth in science and in life, (2) AGI (Artificial General Intelligence), and (3) probability. The most common non-trivial word in this book might be *Bayes*. Another common non-trivial word is *Goodhart*. (Goodhart's law is that when a measure becomes a target, it stops being a measure. It is often referred to when an AI system performs well but for the wrong reasons.) A trivial word would be something like *the* which is likely more common but less interesting. (Or is it trivial? The SIGACT News book review editor Fred Green pointed out that Ohio State has trademarked *the*. See
https://www.cnn.com/2022/06/23/us/ohio-state-university-trademarks-the/index.html

I do not know if that is more or less absurd than Donald Trump's failed attempt to trademark *you're fired*. See
https://www.cobizmag.com/who-owns-the-trademark-to-youre-fired/
to see who really owns the trademark to *you're fired*.)

The Engines of Cognition are actually a set of four books, titled *Trust*, *Incentives*, *Modularity*, and *Failure*. Each book is small—about 9 inches long and 5 inches wide. They can be read in any order. This set of book is a best-of-2019 collection as decided by the readers in some fashion.

[1] ©2022 William Gasarch

2 General Comments

PROS: Many of the essays bring up a topic point that I had not thought of before, or have interesting thoughts about a topic I had thought of before.
CONS: Some of the essays go on and on about some point and either don't have much to say, or take too long saying it. This is most notable in the essays on AI where I want to yell at the author *try it out and see what happens rather than yakking about it.* When I posted a review of another Lesswrong collection, *A Map that Reflects the Territory* here
`https://www.lesswrong.com/posts/JXTEDFCC5r4dW2tta/review-of-a-map-that-reflects-the-territory`
I had the same complaint. Some comments said that building AI systems is dang hard. Okay. Even so, stop yakking about it. It's getting boring.
CAVEAT (both a PRO and a CON): Some of the essays use words or phrases as though I am supposed to already know them. If I was a regular member of the forum then perhaps I *would* know them. In the modern electronic age I can try to look them up. This is a PRO in that I learn new words and phrases. For me this is a really big PRO since *I collect new words and phrases as a hobby.* This is a CON in that going to look things up disrupts the flow of the essays.

In the *third to last* section of this review I will have a list of all of the words and phrases I learned by reading these books and either their meaning or that I could not find their meaning. Why *third to last?* Because the second to last section is my summary opinion, and the reader of this review should be able to find it quickly (the last section is acknowledgments).
CAVEAT: As an extension of the last caveat, the essays tend to be written for other Lesswrongers. Now that I've read 9 Lesswrong books (5 from *Map*, 4 from *Engines*, and a few other occasional essays) I have may have become a Lesswronger; hence, this is no longer a problem for me. However, I sometimes read a paragraph and think "*A mundane*[2] *would not understand this.*"

In the spirit of the Lesswrong's quest for objective truth I will, in each section (and at the end), tabulate how many of the essays were Excellent (E), Good (G), and Meh (M) (none were bad). This will be an objective record of my subjective opinion.

3 Trust

I quote the first paragraph:

The first book is about trust, the belief in something in the absence of understanding.

There are 16 essays of which 7 are excellent, 6 are good, and 3 are meh. I will describe two that are excellent and linked, and one that is meh.

Excellent
Book Review: The Secret to our Success by Scott Alexander and **Reason isn't Magic by Ben Hoffman**

Lesswrong is devoted to reason. Yes indeed, reason is how humans succeeded and is a valuable tool today. Hence it was great to see an article in Lesswrong that uses reason to challenge the notion that reason is so great.

Meta time: I am reviewing Scott Alexander's review of a book. I wonder if when I post this on Lesswrong someone will review my review of Scott's review.

[2] slang term for people who are not Lesswrongers

I will only discuss one aspect of the review; however, the review is fascinating and I assume the book is also.

The book that Scott reviews is *The Secret to our Success*. How did human beings survive? Did you ever try hunting and gathering—it's really hard! One common answer is that humans survived because they are smarter. The book Scott reviews challenges this notion. The book contends that the biggest advantage was cultural learning. Over time techniques that worked were learned *and passed down to the next generation*.

We discuss one of their examples: Manioc. This is a plant that some peoples used as a staple. The time and effort they used to prepare it was very intense. Was this just a tradition (and hence perhaps a waste of time) or was it beneficial? The answers are Yes and Yes. Manioc has a lot of cyanide in it and the process they used got rid of the poison. Of course, they didn't know this. But that's not quite enough— how would they know the *long term affects* of eating the plant? They didn't; however, the process removed the bitter taste and got rid of some short term affects.

Someone could have tried to make the process less time consuming and still get rid of the bitter taste. This would have seemed reasonable but lead to cyanide deaths in the long term. But NO – nobody did this. So their *lack of reason*, their adhering to tradition for no good reason, was beneficial.

The second essay, *Reason isn't Magic* challenges this view. Hoffman points out that the time spent processing the food is also time lost— and perhaps some people starved since the process also made the supply less. This reminds me of the joke:

Vegans don't live longer, it just seems that way.

Meh
Chris Olaf's Views on AI Safety by Evan Hubinger

This is typical of the essays both in these books, in the last set of books I reviewed, on the Lesswrong blog, and other blogs that discuss AI Safety, AI alignment, and other AI issues. They seem to talk a lot but not really say anything. Or, more to the point, they have some ideas. Fine. TRY THEM OUT, then come back with what you found.

Of course there is a caveat: If I rate all of the AI article as *Meh* then does that mean there is something wrong with them (too long, not enough info) or with me (to impatient, not in the area of AI)?

4 Modularity

I quote the first paragraph:

The second book is about modularly. Well-designed or evolved structures are often not just made of parts, but made of parts with simple interfaces. These interfaces allow the parts to be reused in alternative contexts, and thus recombined in different ways.

There are 14 essays of which 6 are excellent, 2 are good, and 6 are meh. I will describe one from each category, which is not a fair sample.

Excellent
Gears-Level Models are Capital Investments by John S. Wentworth

When doing research should you strive to understand *why* things are the way they are (Gears-Level) or just *what* is happening?

An Example from Marketing:

Gears-Level: With massive data find correlations like "People who earn over $100,000 prefer to buy brand name chocolate," and use these to guide your ad campaigns.

Black-Box: Run lots of ad campaigns and see which one works.

The article discusses the pros and cons of these two approaches and gives lots of examples.

Good
Forum Participation as a Research Strategy by Wei Dai

If I read two articles and find a novel way to combine them, do an experiment to verify that my insight is correct, and publish the result, that's clearly research. If I write a blog on Lesswrong (or some other forum) or write a comment on someone else's blog, is that research? Probably not, but it can *lead to* research. This essay discusses the PROS of participating on a forum and how it can contribute to research.

Giving this a *good* instead of an *excellent* might not be fair since I've had a blog on theoretical computer science, shared with Lance Fortnow since 2007, so there was nothing new in it for me. The blog is at
`https://blog.computationalcomplexity.org/`

Meh
The Credit Assignment Problem by Abram Demski

This essay begins with examples of how to assign credit to success or failure and asserts correctly that this is an important problem. They then have some good ideas about the problem. But then the article goes off topic and is too long.

5 Incentives

I quote the first paragraph:

The third book is about incentives, which are patterns of what is rewarded and what is punished.

There are 16 essays of which 10 are excellent (wow!), 2 are good, and 4 are meh. I will review 3 of the excellent essays (2 of which are tied together) and none of the others. This is asymmetric, which is the topic of the first essay I describe. More to the point, it seems like all of my critiques of the Meh essays are the same: talk too much, don't say much. Hence my critiques of them also talk too much and don't say much, so I omitted them.

Excellent
Asymmetric Justice by Zvi Mowshowitz and **The Copenhagen Interpretation of Ethics by Jai Dhyani**

The Copenhagen Interpretation of Ethics is that when you observe or interact with a problem you can be blamed for it. Or perhaps you will be blamed for not doing enough. I give two examples, one from the second essay, and one from neither essay.

1. (This is from the second essay.) At one time Detroit was having a hard time with high water bills. People for the Ethical Treatment of Animals (PETA) told families that they would pay their water bills for a month, if the family went vegan for that month. This article

   `https://www.nbcnews.com/news/us-news/peta-detroit-go-vegan-month-well-pay-your-water-bill` says that PETA was criticized for this. One quote:

Water is a human right. Period. Holding it out like a prize proves PETA doesn't value human life.

2. (This is not from either essay.) During the Flint Michigan Water crisis Ted Cruz donated water to crisis pregnancy centers, which are really places women go to thinking they will get help, but instead they are lectured about why they should NOT get an abortion. He was criticized for this. And I also thought badly of him (more than usual).

BUT WAIT A MINUTE! Did the guy who blasted PETA give any water or money to Detroit? Did I do anything for Flint? It is unfair to criticize them for doing *something* as opposed to doing *nothing.*

The two essays are about issues of justice. One is the issue above, that there may be a disincentive to help. Another issue is asymmetry: bad actions are punished but good actions are not rewarded. How to fix this? The two essays give you a lot to think about.

Moloch Hasn't Won by Zvi Mowshowitz

This essay is one in a sequence of essays that go back before Lesswrong was a forum. In *Hierarchy for Philosophers*[3], C.S. Lewis writes:

Who does it? Earth could be fair, all men glad and wise. Instead we have prisons, smokestacks, asylums. What sphinx of cement and aluminum breaks open their skulls and eats up their imagination?

Alan Ginsberg answers the question: *Moloch does it* He gave a much longer answer, where he is howling at Moloch, but that's the drift. And even with that, I had a hard time figuring out what he meant.

Scott Alexander's Slate article *Meditations on Moloch*, which you can find here
`https://slatestarcodex.com/2014/07/30/meditations-on-moloch/`,
takes the question of why humankind is in such bad shape seriously. He is particularly interested in why, if nobody likes the current system, it persists

Scott's essay has a list of 14 real world phenomena which any rational person would want to change and yet nothing changes. They are mostly Prisoner's Dilemma, Tragedy of the Commons, Malthusian scenarios, but they are not abstract. They are real. He then proposes some ways out of these traps.

Zvi's follow-up essay says what Scott got right and what Scott got wrong. Hmmm, that sounds too shallow. Zvi's essay is an intelligent comment on Scott's essay. Read them both.

6 Failure

I quote the first paragraph:

The fourth book is about failure. It's what happens when a system behaves differently from how we expect it to, with adverse consequences for those who were relying on the success of that system. Failure is often as much about misunderstanding how a system works, as it is about the lack of effort or plan to bring the system into a successful configuration.

[3] I have not been able to find the book or article *Hierarchy for Philosophers by C.S. Lewis.* The only references to it are on the Lesswrong Forum. Conspiracy?

There were 13 essays of which 5 are excellent, 2 are good, and 6 are meh. I will discuss 2 excellent, 1 very good, and 1 meh.

Excellent
Blackmail by Zvi Mowshowitz

This essay discusses why blackmail should be illegal. You might think *of course it should be.* This essay gives good arguments for why it is illegal but also raises questions about the entire endeavor.

Why wasn't science invented in China? by Ruben Bloom

The title is not quite right: some science was done in China at about the same time as in Europe. But far less. This essay gives cogent reasons for this. I quote one here: *Unlike Europe, China's political, religious, legal, and educational systems did not afford the neutral spaces where novel ideas could be advanced and old ideas questioned.*

Good
AI Success Stories Wei Dai

This article discusses various AI success stories and gives criteria to tell if they really were successes. This is interesting; however, it was only 5 pages – I would have wanted more examples.

Meh
The Strategy Stealing Assumption by Paul Christiano

The strategy stealing assumption is that for any strategy an unaligned AI can use to influence the long-run future, there is an analogous strategy that a similarly-sized group of humans can use in order to capture a similar amount of flexible influence over the future. The article is speculative about this. I would prefer it to give concrete examples.

7 Newords that I Learned From These Books

The word *Newords* is not a misspelling. The best neologisms do not need to be explained. Oh well.

7.1 From the Book Trust

1. **Kaggle Competition**: From the website of Kaggle,

   `https://www.kaggle.com/docs/competitions`

   Kaggle Competitions are designed to provide challenges for competitors at all different stages of their machine learning careers. As a result, they are very diverse, with a range of broad types.

2. **The model contains no gears**: Machine learning models often work great but nobody knows why. But it's worse than that. There is no why, it's "just" pattern matching.

3. **Chesterson's Fence**: The theologian G.K. Chesterson said that if you see a fence that you want to knock down, *don't!* You must first understand why it is there. More generally, any proposed reform to a system must understand why the system is there in the first place. Economist-philosopher Edmond Burke had similar ideas and is considered one of the founders

of a certain school of conservative thought. While this is often a wise policy, it can also be an excuse for doing nothing.

4. **The Toxoplasma of Rage**: Memes that are controversial and incite rage, even negative rage, are more effective at getting the message out. Scott Alexander has a great article about this that probably coined the term (for this context–it also has a medical meaning) here:

   `https://slatestarcodex.com/2014/12/17/the-toxoplasma-of-rage/`

5. **Zombie Theories**: A theory appears in a paper that is likely false. However, nobody bothers debunking it, so it keeps getting revived. This can even happen to theories that are debunked, like that vaccines cause autism.

6. **FOOM Debate**: Robin Hanson and Eliezer Yudlowsky had a debate about the future of AI, called *AI-FOOM Debate*. You can read about it, and download it, here

   `https://intelligence.org/ai-foom-debate/`

   Why is it called FOOM? Because **FOOM** is a sudden increase in artificial intelligence such that an AI systems becomes extremely powerful. This may be the same or close to **The Singularity**.

7.2 From the Book Modularity

1. **Gear-level Models** vs **Black-Box Models**: A Gears-Level Model strives to understand what's really going on. A Black-Box Model is only concerned with input-output.

2. **metis and Metis**: The book says that metis is knowledge handed down for generations that might not make sense. Wikipedia says that the Metis are a group of indigenous peoples who inhabit parts of Canada. It is likely that the the Metis have metis. (lalaithion informs me that *metis* is an ancient Greek word which originally meant *magical cunning* but drifted to mean *wisdom* or *prudence* or *the je ne sais quoi of being able to solve practical problems*. He also points to James C Scott's book *Seeing Like a State* where it was used to mean implicit knowledge passed down through a culture.

3. **Disputant**: The article *Coherent Decisions Imply Consistent Utilities*, in the section *Why not circular preferences?* begins as follows

   > *De gustibus non est disputandum* goes the proverb, matters of taste cannot be disputed.

   Okay, that's fine. But later in that section it says,

   > That (circular preferences) sound wrong. But can we disputandum that.

   Clearly the author just meant dispute and is either attempting to be funny (I don't think it is, but of course, *De gustibus non est disputandum*) or made a mistake. Even so, I am happy to know the quote.

4. **Utilon**: A unit of pleasure. You need to define how much it is for yourself as there is no standard. From *Map* I learned the word **Hedon** which is a unit of pleasure. I got 2 utilons and 3 hedons when I read gjm' comment reminding me that I had **Hedon** in my word list from my review of *Map*.

5. **The Allais Paradox**: Wikipedia does a good job on this one, so see

   `https://en.wikipedia.org/wiki/Allais_paradox`

6. **MIRI**: Machine Intelligent Research Institute. I quote their website:

   A non-profit research organization devoted to reduce the existential risk from unfriendly AI, and understanding problems related to friendly AI.

7.3 From the Book Incentives

1. **The Copenhagen Interpretation of Ethics**: See my description of the Incentives book for the definition.

2. **Trained in the Way**: I assumed this meant thinking rationally and objectively and all the good things that Lesswrong values. They reference Eliezar Yudkowsky's post *Twelve Virtues of Rationality.* I read that and he never quite defines the term but it seems to mean what I thought.

3. **Schelling Point**: Two people (or companies) want to communicate but for some reason can't. Even so, if they have the same social-cultural background they may be able to, without communication, get an agreement. The agreement is called the Schelling point. This was first introduced by economist Thomas Schelling in his book *The Strategy of Conflict (1960).* I give three examples.

   (a) Two Americans know they need to meet but can't communicate where and when. They might both end up under the clock in Grand Central Station on New Years Eve at 12:01PM.

   (b) Alice is writing a review of Bob's book, and Bob is writing a review of Alice's book. They *want* to say *I'll give you a good review if you give me a good review* but that would be unethical. Even so, they end up doing just that.

   (c) Two companies sell the same product and the price fluctuates between 8 and 12 dollars. Eventually both will sell it for 10 dollars.

4. **Simulacrum**: A representation of something. Often thought to be inferior. In the book it's used for things like meaningless titles (e.g., Vice President in charge of Sorting). See next phrase. (The plural is *Simulacra.*)

5. **Baudrillard's Theory**: Baudrillard was a French philosopher who thought that society had become so saturated with simulacra and lives so saturated with the constructs of society that all meaning had become meaningless by being infinitely mutable. This theory applied to the Trump presidency explains the epidemic of akrasia in high office as well as our citizenry.

7.4 From the Book Failure

1. **Akrasia**: The article uses the term without saying what it means. A search on the Lesswrong website yielded other articles that use the term without saying what it means. I think they are using the following which I got from Wikipedia:

   Akrasia: A lack of self-control or the state of acting against one's better judgment.

2. **CFAR**: Center for Applied Rationality. Here is their website:

   `https://rationality.org/`

3. **Connectionism**: The belief that we can explain intellectual ability using artificial neural networks.

4. **The Curse of the Counterfactual**: When you compare reality to what could have been (or what your rose-colored hindsight glasses see) you may get depressed. I've also heard this called **buyer's remorse**.

5. **Internal Family Systems (IFS) Therapy**: A therapy used to cure traumas. See the website

   `https://ifs-institute.com/`

6. **Litany of Gendlin**: This is a method to combat the curse of the counterfactual. This is a quote from Eugene Gendlin:

   > "What is true is already so.
   > Owning up to it doesn't make it worse
   > Not being open about it doesn't make it go away
   > And because it's true, it is what is there to be interacted with
   > Anything untrue isn't there to be lived
   > People can stand what is true
   > for they are already enduring it."

   (I got this from the Less Wrong Website. I don't think it's used beyond that.)

7. **Litany of Tarski**: This is a method to combat the curse of the counterfactual. A template to remind oneself that beliefs stem from reality, from what actually is as opposed to what we want or what would be convenient. Logically

   If X then I desire to believe X.

   (I got this from the Less Wrong Website. I don't think it's used beyond that.)

8. **The Unilateralist's Curse**: We give an example from the article *The Unilaterist's Curse and the Case for a Principle of Conformity* by Bostrom, Douglas & Sandberg.

   A group of scientists working on HIV accidentally create an air-transmissible variant. 19 out of the 20 scientists agree that this should not be published. But one disagrees since he thinks the world should know about the danger and prepare. He announces the result at a conference.

   The Unilaterist's Curse is that a small number of people can act against what the vast majority wants.

9. **The Great Divergence**: This refers to when Europe began to dominate other countries culturally, economically, and scientifically. This occurred in the 19th century.

8 Should You Read This Book?

Yes.

Okay, I will elaborate on that.

8.1 Should You Read This Book? The Numbers

I review my ratings E for Excellent, G for Good, or M for Meh (none were B for Bad):

1. *Trust* E-7, G-6, M-3.
2. *Modularity* E-6, G-2, M-1.
3. *Incentives* E-10, G-2, M-4.
4. *Failure* E-5, G-2, M-6.

What to do with this information?

1. There are 28 excellent articles! That's... excellent!
2. There are 12 good articles! That's... good?
3. There are 14 meh articles! That's... meh.

The ratio 28-12-14 is excellent and is better than that for *A Map that Reflects the Territory* which got 15-15-15. Since I don't really think *Map* was worse than *Engines* I may have been harsh on *Map*. Or perhaps I've drank the Kool Aid. In any case, yes, you should buy this book.

8.2 Should You Read This Book? Not the Numbers

Let's look at the extremes: the best and worst thing about the book.

Best things I got out of the book:

1. Many of the essays discuss how the world got the way it is, what's wrong with it, and possible solutions. While these are mostly in the books *Trust* and *Incentives* there are other articles that touch on these points. The discussions are refreshingly honest, objective, and do not have a bias ahead of time.
2. Some of the essays discuss gears-level vs black box. This is one of those concepts which you sort-of know ahead of time but is great to see written down and explained and explored at much greater length. While these were mostly in the book *Modularity* other articles touch on it.
3. This will sound like a back-handed compliment. Or a left-handed complimented. The book gives pointers to OTHER really good books and essays.

Worst thing about the book:

1. As noted earlier, the AI essays often had too much talk-talk-talk and not enough walk-walk-walk. I am reminded of how early (in the 1960s) people would talk about how great AI was going to be and then babble some incoherent philosophy about machines thinking. The current discussion is *not* how great it will be. It's about (a) is AI dangerous? and (b) if so how to tame it? Those are good questions, but the essays about it were talking in a vacuum. This reminds me of an old joke and a new joke. Both begin the same.

   How does someone in AI make love to their spouse?

   - Old Joke: They sit on the bed and tell them how great its going to be.
   - New Joke: They sit on the bed and talk about making sure to align what you really want with what you say you want, and be careful since if you get things wrong that could be dangerous.

Fortunately, as the numbers tell you, the meh essays were fairly few; however, as a book reviewer I had to read them. You can use a variant of Ebert's rule:

If you don't laugh in the first 15 minutes of a comedy, you won't laugh in the remaining 105 minutes.

I also hasten to point out, these are just my opinions, and as well known, *de gustibus non est disputandum.*

8.3 Should You Read This Book? The Elephant in the Room

(The next paragraph is almost word-for-word what I wrote in the review of *A Map that Reflects the Territory.*)

And now for the elephant in the room: Why buy a book if the essays are on the web for free? I have addressed this issue in the past:

1. I've reviewed 4 blog books. See the next four links for the reviews:

   `https://www.cs.umd.edu/~gasarch/BLOGPAPERS/lipton.pdf`

   `https://www.cs.umd.edu/~gasarch/BLOGPAPERS/liptonregan.pdf`

   `https://www.cs.umd.edu/~gasarch/BLOGPAPERS/tao.pdf`

   `https://www.cs.umd.edu/~gasarch/bookrev/FRED/Lesswrong.pdf`

2. I have written my own blog book: *Problems with a point: Explorations in Math and Computer Science by Gasarch and Kruskal*

   See here for its entry on amazon:

   `https://www.amazon.com/Problems-Point-Exploring-Computer-Science/dp/9813279974`

Here is an abbreviated quote from my book that applies to the book under review.

The Elephant in the Room

So why should you buy this book if it's available for free?

1. Trying to find which entries are worth reading would be hard. There are a lot of entries and it really is a mixed bag.

2. There is something about a book that makes you want to read it. Having words on a screen just doesn't do it. I used to think this was my inner-Luddite talking, but younger people agree, especially about math-on-the-screen.

9 Acknowledgments

I thank the entire Lesswrong community for existing and putting forth reason as a way of life (though be careful eating Manioc).

I thank Oliver Habryka for giving me this opportunity to review these books.

I thank Nathan Hayes and Ben Pace for proofreading and useful comments.

I thank Ruby Bloom and Justis-from-Lesswrong2.0 for helping me post this review to the LessWrong forum in text form (having a pointer to the review I learned was not as good).

I thank Fred Green, editor of SIGACT News for proofreading and not asking me to cut anything for length.

I thank lalaithon, gjm, and Richard Kennaway for comments they made when I posted this on lesswrong, which I incorporated into the review.

All of the people acknowledged help make this review LessWrong.