Best of LessWrong: June 2020

1. The ground of optimization
2. Growing Independence
3. We've built Connected Papers - a visual tool for researchers to find and explore academic papers
4. Simulacra Levels and their Interactions
5. Wireless is a trap
6. Self-sacrifice is a scarce resource
7. Everyday Lessons from High-Dimensional Optimization
8. A Personal (Interim) COVID-19 Postmortem
9. Five Ways To Prioritize Better
10. Our take on CHAI's research agenda in under 1500 words
11. A revolution in philosophy: the rise of conceptual engineering
12. Philosophy in the Darkest Timeline: Basics of the Evolution of Meaning
13. Optimized Propaganda with Bayesian Networks: Comment on "Articulating Lay Theories Through Graphical Models"
14. Reply to Paul Christiano on Inaccessible Information
15. Betting with Mandatory Post-Mortem
16. Don't Make Your Problems Hide
17. Half-Baked Products and Idea Kernels
18. My weekly review habit
19. Plausible cases for HRAD work, and locating the crux in the "realism about rationality" debate
20. Inaccessible information
21. News ⊂ Advertising
22. Superexponential Historic Growth, by David Roodman
23. Sparsity and interpretability?
24. What are some Civilizational Sanity Interventions?
25. Public Static: What is Abstraction?
26. The Skill of Noticing Emotions
27. May Gwern.net newsletter (w/GPT-3 commentary)
28. Most reliable news sources?
29. GPT-3 Fiction Samples
30. Turns Out Interruptions Are Bad, Who Knew?
31. What's Your Cognitive Algorithm?
32. A Practical Guide to Conflict Resolution: Comprehension
33. [AN #102]: Meta learning by GPT-3, and a list of full proposals for AI alignment
34. Results of $1,000 Oracle contest!
35. How alienated should you be?
36. Coronavirus as a test-run for X-risks
37. Quick Look #1 Diophantus of Alexandria
38. Where to Start Research?
39. Preparing for "The Talk" with AI projects
40. Relevant pre-AGI possibilities
41. What is meant by Simulcra Levels?
42. Using a memory palace to memorize a textbook.
43. Mediators of History
44. Prediction = Compression [Transcript]
45. Locality of goals
46. Life at Three Tails of the Bell Curve
47. Three characteristics: impermanence
48. Institutional Senescence
49. [AN #103]: ARCHES: an agenda for existential safety, and combining natural language with deep RL
50. If AI is based on GPT, how to ensure its safety?

Best of LessWrong: June 2020

1. The ground of optimization
2. Growing Independence
3. We've built Connected Papers - a visual tool for researchers to find and explore academic papers
4. Simulacra Levels and their Interactions
5. Wireless is a trap
6. Self-sacrifice is a scarce resource
7. Everyday Lessons from High-Dimensional Optimization
8. A Personal (Interim) COVID-19 Postmortem
9. Five Ways To Prioritize Better
10. Our take on CHAI's research agenda in under 1500 words
11. A revolution in philosophy: the rise of conceptual engineering
12. Philosophy in the Darkest Timeline: Basics of the Evolution of Meaning
13. Optimized Propaganda with Bayesian Networks: Comment on "Articulating Lay Theories Through Graphical Models"
14. Reply to Paul Christiano on Inaccessible Information
15. Betting with Mandatory Post-Mortem
16. Don't Make Your Problems Hide
17. Half-Baked Products and Idea Kernels
18. My weekly review habit
19. Plausible cases for HRAD work, and locating the crux in the "realism about rationality" debate
20. Inaccessible information
21. News ⊂ Advertising
22. Superexponential Historic Growth, by David Roodman
23. Sparsity and interpretability?
24. What are some Civilizational Sanity Interventions?
25. Public Static: What is Abstraction?
26. The Skill of Noticing Emotions
27. May Gwern.net newsletter (w/GPT-3 commentary)
28. Most reliable news sources?
29. GPT-3 Fiction Samples
30. Turns Out Interruptions Are Bad, Who Knew?
31. What's Your Cognitive Algorithm?
32. A Practical Guide to Conflict Resolution: Comprehension
33. [AN #102]: Meta learning by GPT-3, and a list of full proposals for AI alignment
34. Results of $1,000 Oracle contest!
35. How alienated should you be?
36. Coronavirus as a test-run for X-risks
37. Quick Look #1 Diophantus of Alexandria
38. Where to Start Research?
39. Preparing for "The Talk" with AI projects
40. Relevant pre-AGI possibilities
41. What is meant by Simulcra Levels?
42. Using a memory palace to memorize a textbook.
43. Mediators of History
44. Prediction = Compression [Transcript]
45. Locality of goals
46. Life at Three Tails of the Bell Curve

47. Three characteristics: impermanence
48. Institutional Senescence
49. [AN #103]: ARCHES: an agenda for existential safety, and combining natural language with deep RL
50. If AI is based on GPT, how to ensure its safety?

The ground of optimization

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This work was supported by OAK, a monastic community in the Berkeley hills. This document could not have been written without the daily love of living in this beautiful community. The work involved in writing this cannot be separated from the sitting, chanting, cooking, cleaning, crying, correcting, fundraising, listening, laughing, and teaching of the whole community.

What is optimization? What is the relationship between a computational optimization process — say, a computer program solving an optimization problem — and a physical optimization process — say, a team of humans building a house?

We propose the concept of an optimizing system as a physically closed system containing both that which is being optimized and that which is doing the optimizing, and defined by a tendency to evolve from a broad basin of attraction towards a small set of target configurations despite perturbations to the system. We compare our definition to that proposed by Yudkowsky, and place our work in the context of work by Demski and Garrabrant's *Embedded Agency*, and Drexler's *Comprehensive AI Services*. We show that our definition resolves difficult cases proposed by Daniel Filan. We work through numerous examples of biological, computational, and simple physical systems showing how our definition relates to each.

Introduction

In the field of computer science, an optimization algorithm is a computer program that outputs the solution, or an approximation thereof, to an optimization problem. An optimization problem consists of an objective function to be maximized or minimized, and a feasible region within which to search for a solution. For example we might take the objective function $(x^2 - 2)^2$ as a minimization problem and the whole real number line as the feasible region. The solution then would be $x = \sqrt{2}$ and a working optimization algorithm for this problem is one that outputs a close approximation to this value.

In the field of operations research and engineering more broadly, optimization involves improving some process or physical artifact so that it is fit for a certain purpose or fulfills some set of requirements. For example, we might choose to measure a nail factory by the rate at which it outputs nails, relative to the cost of production inputs. We can view this as a kind of objective function, with the factory as the object of optimization just as the variable x was the object of optimization in the previous example.

There is clearly a connection between optimizing the factory and optimizing for x, but what exactly is this connection? What is it that identifies an algorithm as an optimization algorithm? What is it that identifies a process as an optimization process?

The answer proposed in this essay is: an optimizing system is a physical process in which the configuration of some part of the universe moves predictably towards a small set of target configurations from any point in a broad basin of optimization, *despite perturbations during the optimization process*.

We do not imagine that there is some engine or agent or mind performing optimization, separately from that which is being optimized. We consider the whole system jointly — engine and object of optimization — and ask whether it exhibits a tendency to evolve towards a predictable target configuration. If so, then we call it an optimizing system. If the basin of attraction is deep and wide then we say that this is a robust optimizing system.

An optimizing system as defined in this essay is known in dynamical systems theory as a dynamical system with one or more attractors. In this essay we show how this framework can help to understand optimization as manifested in physically closed systems containing both engine and object of optimization.

In this way we find that optimizing systems are not something that are designed but are discovered. The configuration space of the world contains countless pockets shaped like small and large basins, such that if the world should crest the rim of one of these pockets then it will naturally evolve towards the bottom of the basin. We care about them because we can use our own agency to tip the world into such a basin and then let go, knowing that from here on things will evolve towards the target region.

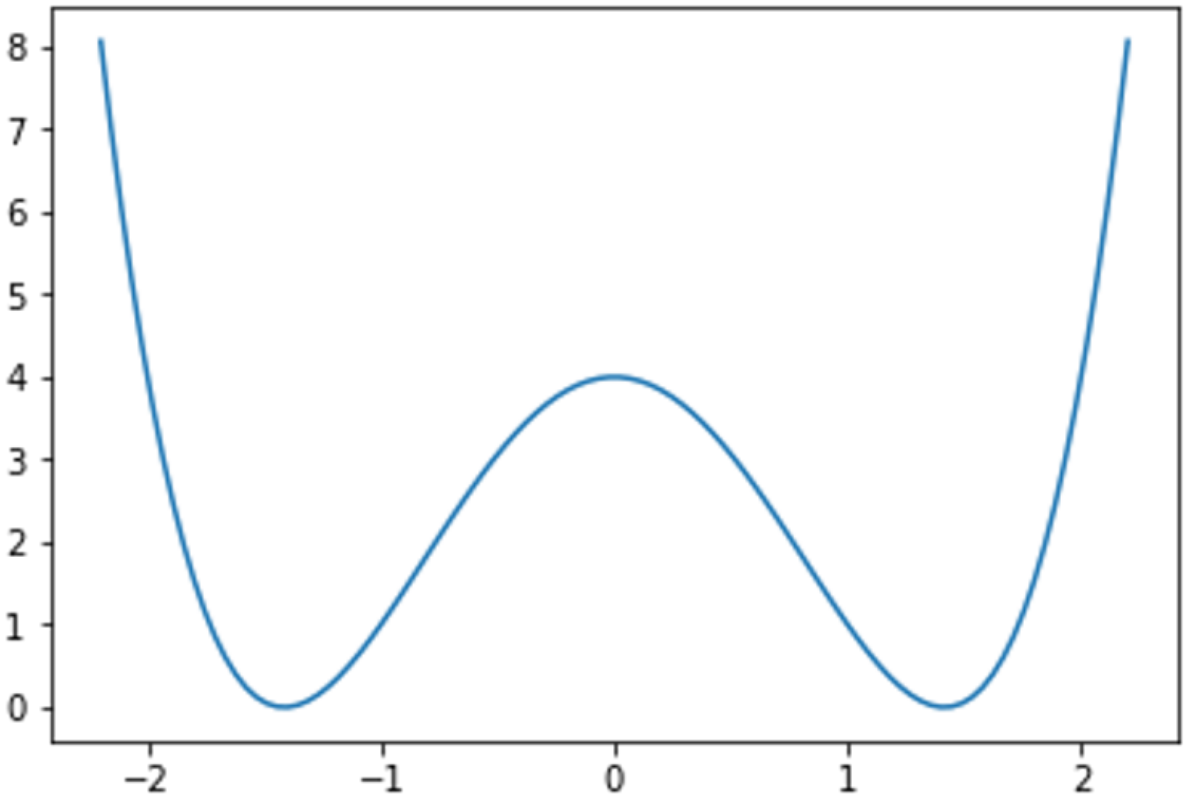
All optimization basins have a finite extent. A ball may roll to the center of a valley if initially placed anywhere within the valley, but if it is placed outside the valley then it will roll somewhere else entirely, or perhaps will not roll at all. Similarly, even a very robust optimizing system has an outer rim to its basin of attraction, such that if the configuration of the system is perturbed beyond that rim then the system no longer evolves towards the target that it once did. When an optimizing system deviates beyond its own rim, we say that it dies. An existential catastrophe is when the optimizing system of life on Earth moves beyond its own outer rim.

Example: computing the square root of two

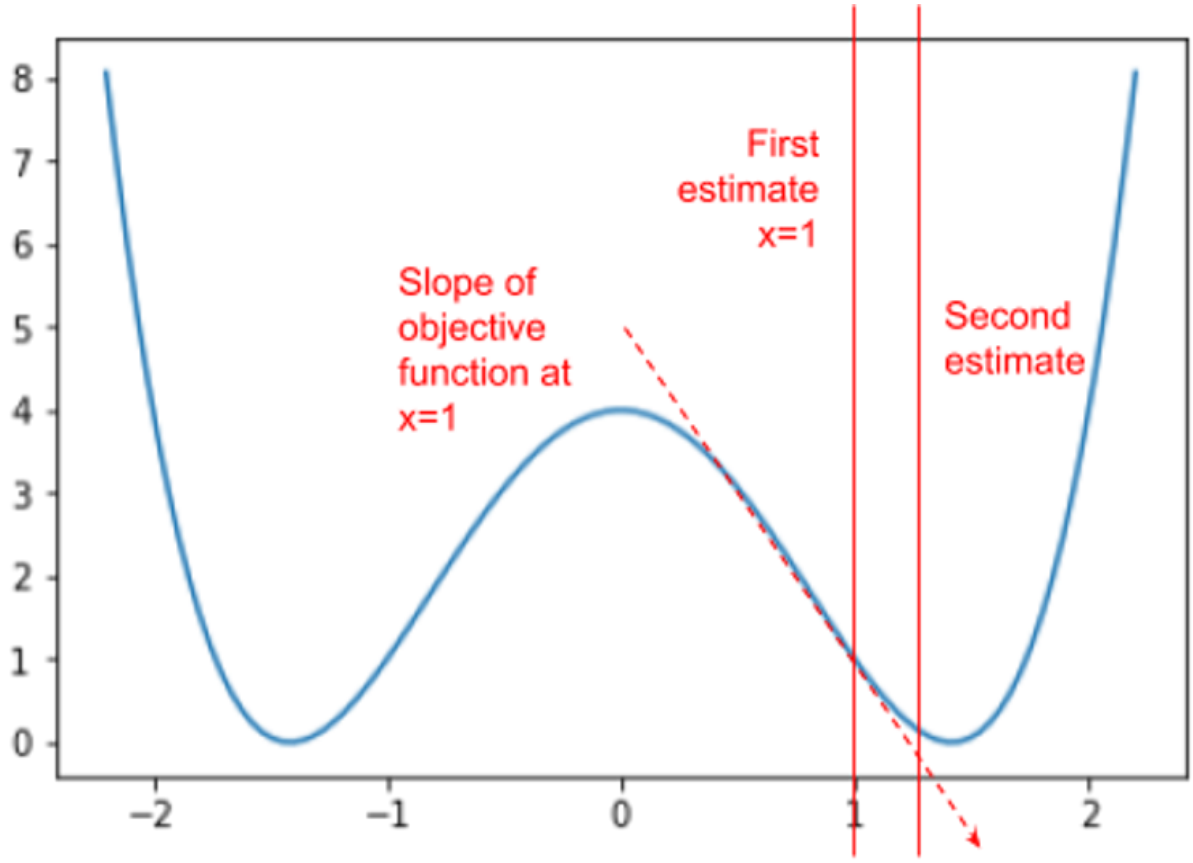
Say I ask my computer to compute the square root of two, for example by opening a python interpreter and typing:

```
>>> print(math.sqrt(2))
1.41421356237
```

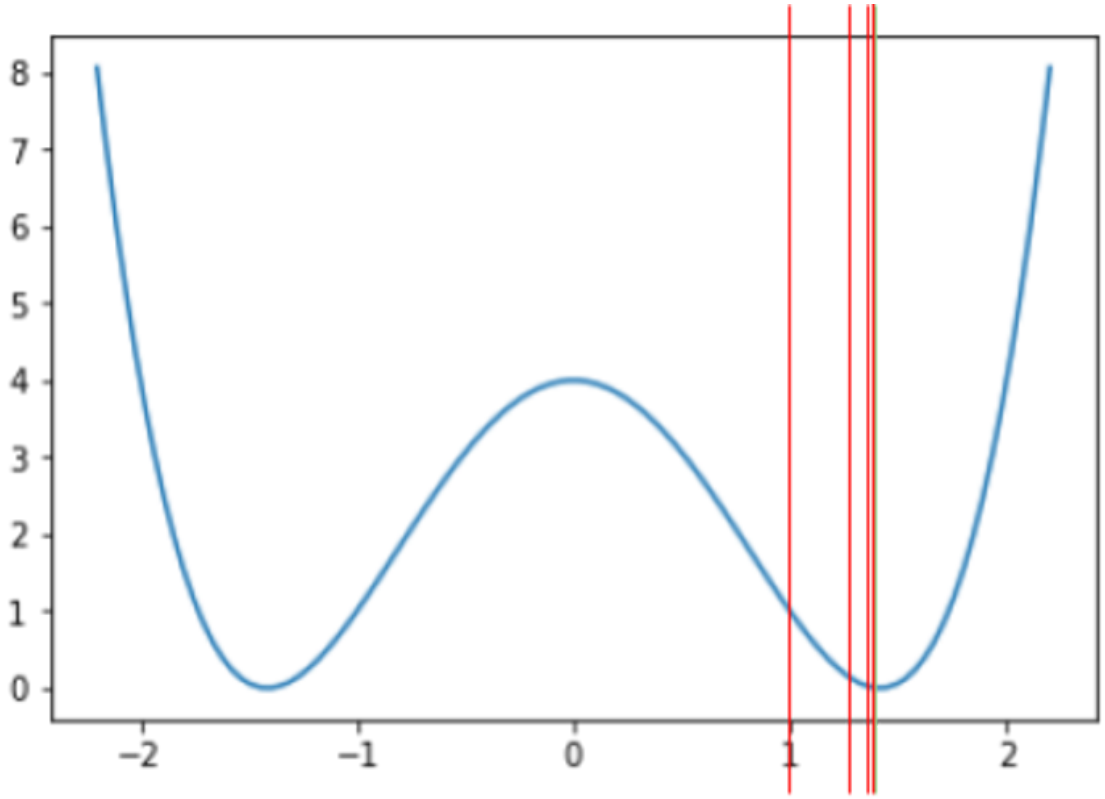
The value printed here is actually calculated by solving an optimization problem. It works roughly as follows. First we set up an objective function that has as its minimum value the square root of two. One function we could use is $y = (x^2 - 2)^2$



Next we pick an initial estimate for the square root of two, which can be any number whatsoever. Let’s take 1.0 as our initial guess. Then we take a gradient step in the direction indicated by computing the slope of the objective function at our initial estimate:



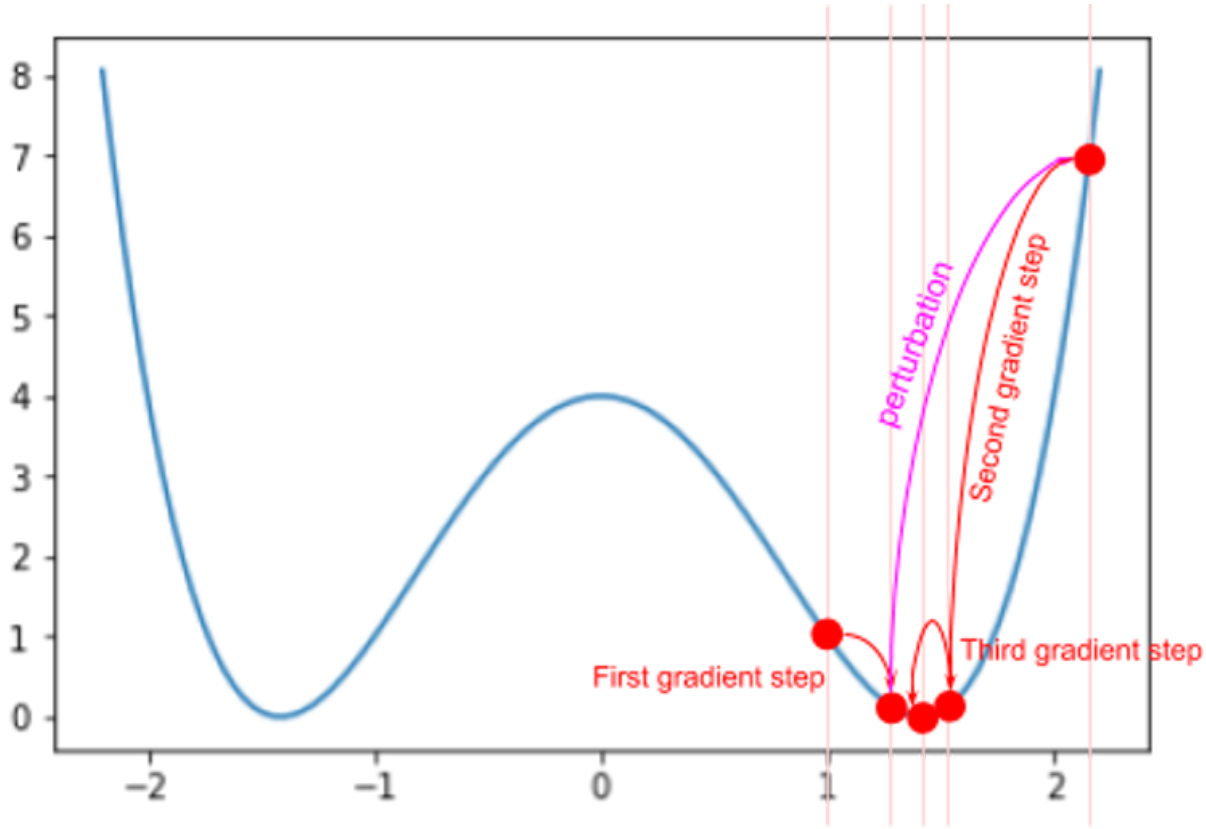
Then we repeat this process of computing the slope and updating our estimate over and over, and our optimization algorithm quickly converges to the square root of two:



This is gradient descent, and it can be implemented in a few lines of python code:

```
current_estimate = 1.0
step_size = 1e-3
while True:
        objective = (current_estimate**2 - 2) ** 2
        gradient = 4 * current_estimate * (current_estimate**2 - 2)
        if abs(gradient) < 1e-8:
                break
        current_estimate -= gradient * step_size
```

But this program has the following unusual property: we can modify the variable that holds the current estimate of the square root of two at any point while the program is running, *and the algorithm will still converge to the square root of two*. That is, while the code above is running, if I drop in with a debugger and overwrite the current estimate while the loop is still executing, what will happen is that the next gradient step will start correcting for this perturbation, pushing the estimate back towards the square root of two:



If we give the algorithm time to converge to within machine precision of the actual square root of two then the final output will be bit-for-bit identical to the result we would have gotten without the perturbation.

Consider this for a moment. For most kinds of computer code, overwriting a variable while the code is running will either have no effect because the variable isn't used, or it will have a catastrophic effect and the code will crash, or it will simply cause the code to output the wrong answer. If I use a debugger to drop in on a webserver servicing an http request and I overwrite some variable with an arbitrary value just as the code is performing a loop in which this variable is used in a central way, bad things are likely to happen! Most computer code is not robust to arbitrary in-flight data modifications.

But this code that computes the square root of two *is* robust to in-flight data modifications, or at least the "current estimate" variable is. It's not that our perturbation has no effect: if we change the value, the next iteration of the algorithm will compute the objective function and its slope at a completely different point, and each iteration after that will be different to how it would have been if we hadn't intervened. The perturbation may change the total number of iterations before convergence is reached. But ultimately the algorithm will *still output an estimate of the square root of two*, and, given time to fully converge, it will output the exact same answer it would have output without the perturbation. This is an unusual breed of computer program indeed!

What is happening here is that we have constructed a physical system consisting of a computer and a python program that computes the square root of two, such that:

- for a set of starting configurations (in this case the set of configurations in which the "current estimate" variable is set to each representable floating point

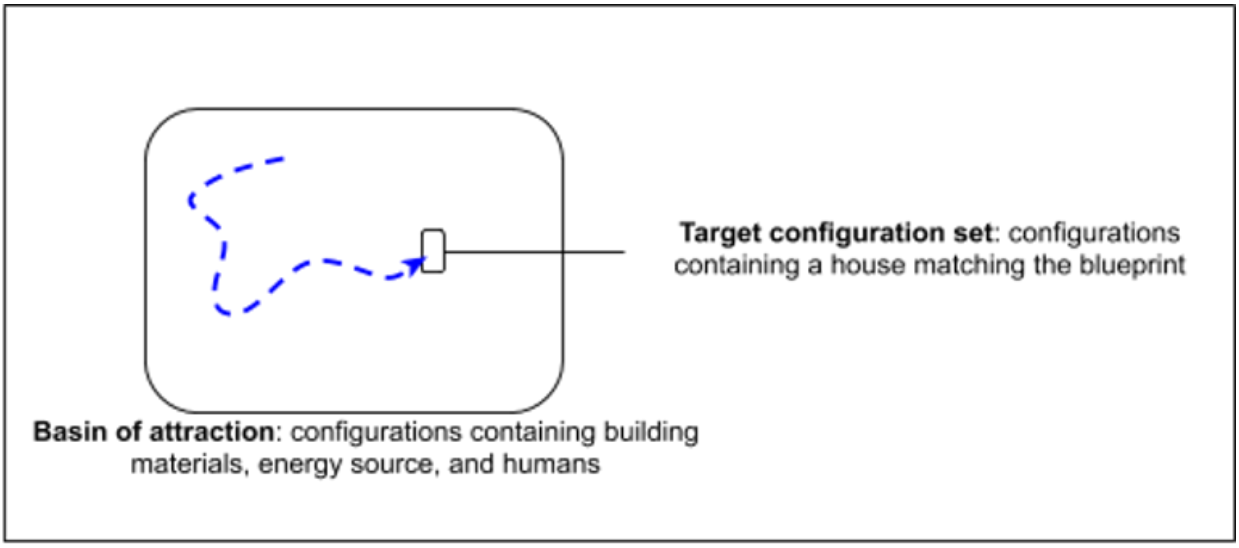
number),

- the system exhibits a tendency to evolve towards a small set of target configurations (in this case just the single configuration in which the "current estimate" variable is set to the square root of two),
- and this tendency is robust to in-flight perturbations to the system's configuration (in this case robustness is limited to just the dimensions corresponding to changes in the "current estimate" variable).

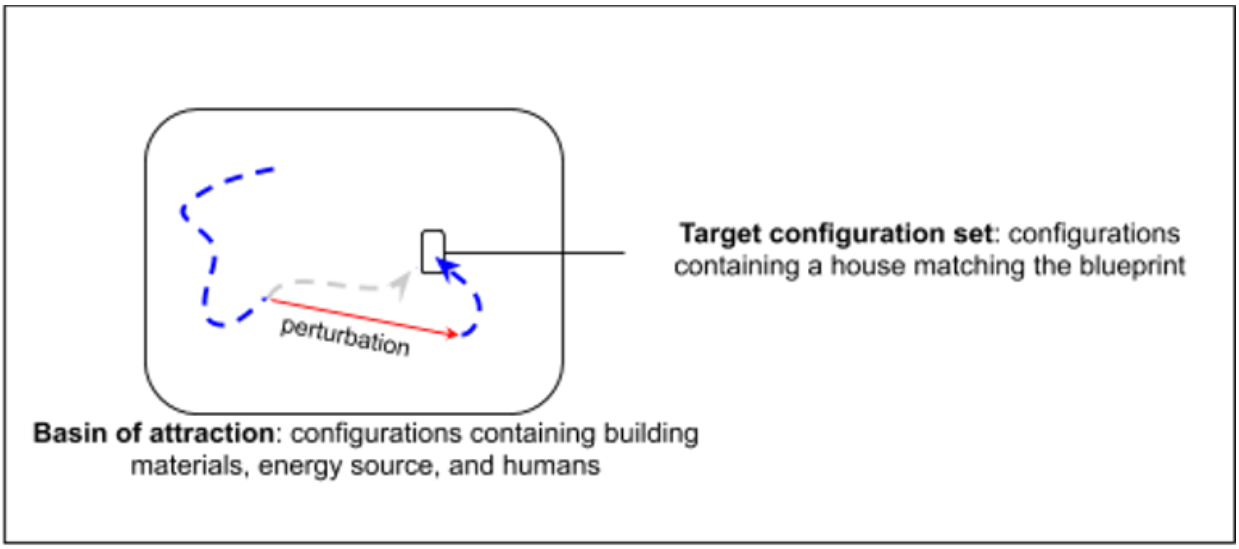
In this essay I argue that systems that converge to some target configuration, and will do so despite perturbations to the system, are the systems we should rightly call "optimizing systems".

Example: building a house

Consider a group of humans building a house. Let us consider the humans together with the building materials and construction site as a single physical system. Let us imagine that we assemble this system inside a completely closed chamber, including food and sleeping quarters for the humans, lighting, a power source, construction materials, construction blueprint, as well as the physical humans with appropriate instructions and incentives to build the house. If we just put these physical elements together we get a system that has a tendency to evolve under the natural laws of physics towards a configuration in which there is a house matching the blueprint.



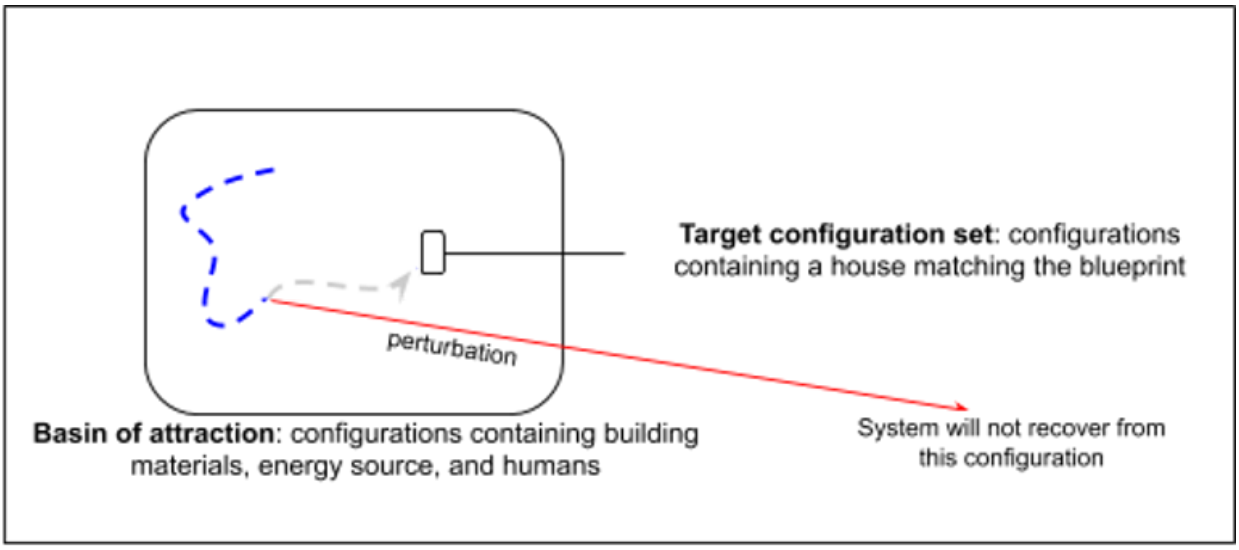
We could perturb the system while the house is being built — say by dropping in at night and removing some walls or moving some construction materials about — and this physical system will recover. The team of humans will come in the next day and find the construction materials that were moved, put in new walls to replace the ones that were removed, and so on.



Just like the square root of two example, here is a physical system with:

- A basin of attraction (all the possible arrangements of viable humans and building materials)
- A target configuration set that is small relative to the basin of attraction (those in which the building materials have been arranged into a house matching the design)
- A tendency to evolve towards the target configurations when starting from any point within the basin of attraction, *despite in-flight perturbations to the system*

Now this system is not infinitely robust. If we really scramble the arrangement of atoms within this system then we'll quickly wind up with a configuration that does not contain any humans, or in which the building materials are irrevocably destroyed, and then we will have a system without the tendency to evolve towards any small set of final configurations.



In the physical world we are not surprised to find systems that have this tendency to evolve towards a small set of target configurations. If I pick up my dog while he is sleeping and move him by a few inches, he still finds his way to his water bowl when he wakes up. If I pull a piece of bark off a tree, the tree continues to grow in the same upward direction. If I make a noise that surprises a friend working on some math homework, the math homework still gets done. Systems that contain living beings regularly exhibit this tendency to evolve towards target configurations, and tend to do so in a way that is robust to in-flight perturbations. As a result we are familiar with physical systems that have this property, and we are not surprised when they arise in our lives.

But physical systems in general do not have the tendency to evolve towards target configurations. If I move a billiard ball a few inches to the left while a bunch of billiard balls are energetically bouncing around a billiard table, the balls are likely to come to rest in a very different position than if I had not moved the ball. If I change the trajectory of a satellite a little bit, the satellite does not have any tendency to move back into its old orbit.

The computer systems that we have built are still, by and large, more primitive than the living systems that we inhabit, and most computer systems do not have the tendency to evolve robustly towards some set of target configurations, so optimization algorithms as discussed in the previous section, which do have this property, are somewhat unusual.

Defining optimization

An optimizing system is a system that has a tendency to evolve towards one of a set of configurations that we will call the *target configuration set*, when started from any configuration within a larger set of configurations, which we call the *basin of attraction*, and continues to exhibit this tendency with respect to the same target configuration set despite perturbations.

Some systems may have a single target configuration towards which they inevitably evolve. Examples are a ball in a steep valley with a single local minimum, and a computer computing the square root of two. Other systems may have a set of target configurations and perturbing the system may cause it to evolve towards a different member of this set. Examples are a ball in a valley with multiple local minima, or a tree growing upwards (perturbing the tree by, for example, cutting off some branches while it is growing will probably change its final shape, but will not change its tendency to grow towards *one* of the configurations in which it has reached its maximum size).

We can quantify optimizing systems in the following ways.

Robustness. Along how many dimensions can we perturb the system without altering its tendency to evolve towards the target configuration set? What magnitude perturbation can the system absorb along these dimensions? A self-driving car navigating through a city may be robust to perturbations that involve physically moving the car to a different position on the road in the city, but not to perturbations that involve changing the state of physical memory registers that contain critical bits of computer code in the car's internal computer.

Duality. To what extent can we identify subsets of the system corresponding to "that which is being optimized" and "that which is doing the optimization"? Between engine and object of optimization; between agent and world. Highly dualistic systems may be robust to perturbations of the object of optimization, but brittle with respect to perturbations of the engine of optimization. For example, a system containing a 2020s-era robot moving a vase around is a dualistic optimizing system: there is a clear subset of the system that is the engine of optimization (the robot), and object of optimization (the vase). Furthermore, the robot may be able to deal with a wide variety of perturbations to the environment and to the vase, but there are likely to be numerous small perturbations to the robot itself that will render it inert. In contrast, a tree is a non-dualistic optimizing system: the tree does grow towards a set of target configurations, but it makes no sense to ask which part of the tree is "doing" the optimization and which part is "being" optimized. This latter example is discussed further below.

Retargetability. Is it possible, using only a microscopic perturbation to the system, to change the system such that it is still an optimizing system but with a different target configuration set? A system containing a robot with the goal of moving a vase to a certain location can be modified by making just a small number of microscopic perturbations to key memory registers such that the robot holds the goal of moving the vase to a different location and the whole vase/robot system now exhibits a tendency to evolve towards a different target configuration. In contrast, a system containing a ball rolling towards the bottom of a valley cannot generally be modified by any *microscopic* perturbation such that the ball will roll to a different target location. A tree is an intermediate example: to cause the tree to evolve towards a different target configuration set — say, one in which its leaves were of a different shape — one would have to modify the genetic code simultaneously in all of the tree's cells.

Relationship to Yudkowsky's definition of optimization

In *Measuring Optimization Power*, Eliezer Yudkowsky defines optimization as a process in which some part of the world ends up in a configuration that is high in an agent's preference ordering, yet has low probability of arising spontaneously. Yudkowsky's

definition asks us to look at a patch of the world that has already undergone optimization by an agent or mind, and draw conclusions about the power or intelligence of that mind by asking how unlikely it would be for a configuration of equal or greater utility (to the agent) to arise spontaneously.

Our definition differs from this in the following ways:

- We look at whole systems that evolve naturally under physical laws. We do not assume that we can decompose these systems into some engine and object of optimization, or into mind and environment. We do not look at systems that are "being optimized" by some external entity but rather at "optimizing systems" that exhibit a natural tendency to evolve towards a target configuration set. These optimizing systems may contain subsystems that have the properties of agents, but as we will see there are many instances of optimizing systems that do not contain dualistic agentic subsystems.
- When discerning the boundary between optimization and non-optimization, we look principally at robustness — whether the system will continue to evolve towards its target configuration set in the face of perturbations — whereas Yudkowsky looks at the improbability of the final configuration.

Relationship to Drexler's *Comprehensive AI Services*

Eric Drexler has written about the need to consider AI systems that are not goal-directed agents. He points out that the most economically important AI systems today are not constructed within the agent paradigm, and that in fact agents represent just a tiny fraction of the design space of intelligent systems. For example, a system that identifies faces in images would be an intelligent system but not an agent according to Drexler's taxonomy. This perspective is highly relevant to our discussion here since we seek to go beyond the narrow agent model in which intelligent systems are conceived of as unitary entities that receive observations from the environment, send actions back into the environment, but are otherwise separate from the environment.

Our perspective is that there is a specific class of intelligent systems — which we call optimizing systems — that are worthy of special attention and study due to their potential to reshape the world. The set of optimizing systems is smaller than the set of all AI services, but larger than the set of goal-directed agentic systems.

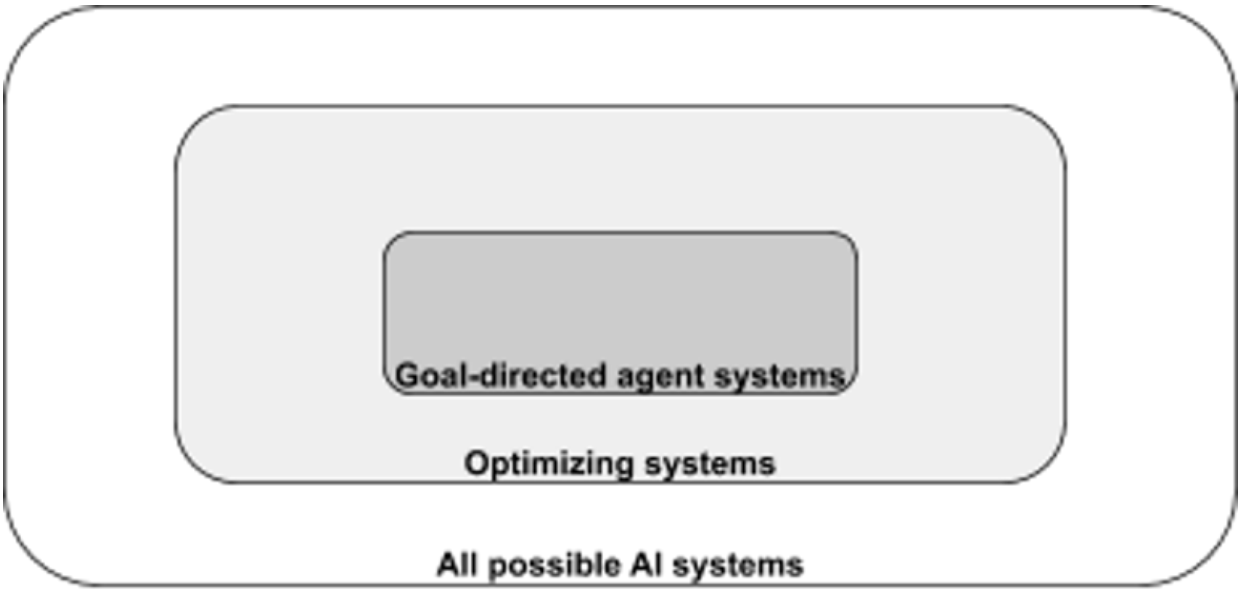


Figure: relationship between our optimizing system concept and Drexler's taxonomy of AI systems

Examples of systems that lie in each of these three tiers are as follows:

- A system that identifies faces in images by evaluating a feed-forward neural network is an AI system but not an optimizing system.
- A tree is an optimizing system but not a goal-directed agent system (see section below analyzing a tree as an optimizing system).
- A robot with the goal of moving a ball to a specific destination is a goal-directed agent system.

Relationship to Garrabrant and Demski's *Embedded Agency*

Scott Garrabrant and Abram Demski have written about the many ways that a dualistic view of agency in which one conceives of a hard separation between agent and environment fails to capture the reality of agents that are reducible to the same basic building-blocks as the environments in which they are embedded. They show that if one starts from a dualistic view of agency then it is difficult to design agents capable of reflecting on and making improvements to their own cognitive processes, since the dualistic view of agency rests on a unitary agent whose cognition does not affect the world except via explicit actions. They also show that reasoning about counterfactuals becomes nonsensical if starting from a dualistic view of agency, since the agent's cognitive processes are governed by the same physical laws as those that govern the environment, and the agent can come to notice this fact, leading to confusion when considering the consequences actions that are different from the actions that the agent will, in fact, output.

One could view the *Embedded Agency* work as enumerating the many logical pitfalls one falls into if one takes the "optimizer" concept as the starting point for designing

intelligent systems, rather than "optimizing system" as we propose here. The present work is strongly inspired by Garrabrant and Demski's work. Our hope is to point the way to a view of optimization and agency that captures reality sufficiently well to avoid the logical pitfalls identified in the *Embedded Agency* work.

Example: ball in a valley

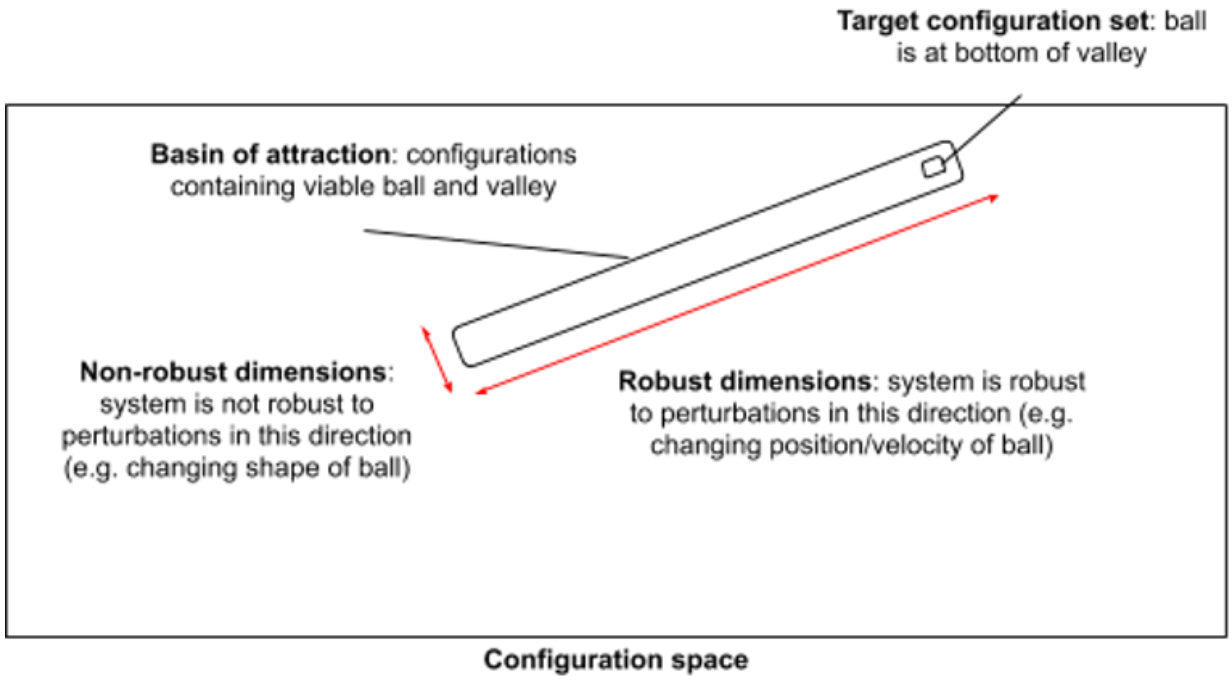
Consider a physical ball rolling around in a small valley. According to our definition of optimization, this is an optimizing system:

Configuration space. The system we are studying consists of the physical valley plus the ball

Basin of attraction. The ball could initially be placed anywhere in the valley (these are the configurations comprising the basin of attraction)

Target configuration set. The ball will roll until it ends up at the bottom of the valley (the set of local minima are the target configurations)

We can perturb the ball while it is "in flight", say by changing its position or velocity, and the ball will still ultimately end up at one of the target configurations. This system is robust to perturbations along dimensions corresponding to the spatial position and velocity of the ball, but there are many more dimensions along which this system is not robust. If we change the shape of the ball to a cube, for example, then the ball will not continue rolling to the bottom of the valley.



Example: ball in valley with robot

Consider now a ball in a valley as above, but this time with the addition of an intelligent robot holding the goal of ensuring that the ball reaches the bottom of the valley.

Configuration space. The system we are studying now consists of the physical valley, the ball, and the robot. We consider the evolution of and perturbations to this whole joint system.

Target configuration set. As before, the target configuration is the ball being at the bottom of the valley

Basin of attraction. As before, the basin of attraction consists of all the possible spatial locations that the ball could be placed in the valley.

We can now perturb the system along many more dimensions than in the case where there was no robot. For example, we could introduce a barrier that prevents the ball from rolling downhill past a certain point, and we can then expect a sufficiently intelligent robot to move the ball over the barrier. We can expect a sufficiently well-designed robot to be able to overcome a wide variety of hurdles that gravity would not overcome on its own. Therefore we say that this system is more robust than the system without the robot.

There is a sequence of systems spanning the gap between a ball rolling in a valley, which is robust to a narrow set of perturbations and therefore we say exhibits a weak degree of optimization, up to a robot with a goal of moving a ball around in a valley, which is robust to a much wider set of perturbations, and therefore we say exhibits a stronger degree of optimization. Therefore the difference between systems that do and do not undergo optimization is not a binary distinction but a continuous gradient of increasing robustness to perturbations.

By introducing the robot to the system we have also introduced new dimensions along which the system is fragile: the dimensions corresponding to modifications to the robot itself, and in particular the dimensions corresponding to modifications to the code running on the robot (i.e. physical perturbations to the configuration of the memory cells in which the code is stored). There are two types of perturbation we might consider:

- Perturbations that destroy the robot. There are numerous ways we could cut wires or scramble computer code that would leave the robot completely non-operational. Many of these would be physically microscopic, such as flipping a single bit in a memory cell containing some critical computer code. In fact there are now *more* ways to break the system via microscopic perturbations compared to when we were considering a ball in a valley without a robot, since there are few ways to cause a ball not to reach the bottom of a valley by making only a microscopic perturbation to the system, but there are many ways to break modern computer systems via a microscopic perturbation.

- Perturbations that change the target configurations. We could also make physically microscopic perturbations to this system that change the robot's goal. For example we might flip the sign on some critical computations in the robot's code such that the robot works to place the ball at the highest point rather than the lowest. This is still a physical perturbation to the valley/ball/robot system: it is one that affects the configuration of the memory cells containing the robot's computer code. These kinds of perturbations may point to a concept with some similarity to that of an agent. If we have a system that can be perturbed in a way that preserves the robustness of the basin of convergence but changes the target

configuration towards which the system tends to evolve, and if we can find perturbations that cause the target configurations to match our own goals, then we have a way to navigate between convergence basins.

Example: computer performing gradient descent

Consider now a computer running an iterative gradient descent algorithm in order to solve an optimization problem. For concreteness let us imagine that the objective function being optimized is globally convex, in which case the algorithm will certainly reach the global optimum given sufficient time. Let us further imagine that the computer stores its current best estimate of the location of the global optimum (which we will henceforth call the "optimizand") at some known memory location, and updates this after every iteration of gradient descent.

Since this is a purely computational process, it may be tempting to define the configuration space at the computational level — for example by taking the configuration space to be the domain of the objective function. However, it is of utmost importance when analyzing any optimizing system to ground our analysis in a physical system evolving according to the physical laws of nature, just as we have for all previous examples. The reason this is important is to ensure that we always study complete systems, not just some inert part of the system that is "being optimized" by something external to the system. Therefore we analyze this system as follows.

Configuration space. The system consists of a physical computer running some code that performs gradient descent. The configurations of the system are the physical configurations of the atoms comprising the computer.

Target-configuration set. The target configuration set consists of the set of physical configurations of the computer in which the memory cells that store the current optimized state contain the true location of the global optimum (or the closest floating point representation of it).

Basin of attraction. The basin of attraction consists of the set of physical configurations in which there is a viable computer and it is running the gradient descent algorithm.

Example: billiard balls

Let us now examine a system that is not an optimizing system according to our definition. Consider a billiard table with some billiard balls that are currently bouncing around in motion. Left alone, the balls will eventually come to rest in some configuration. Is this an optimizing system?

In order to qualify as an optimizing system, a system must (1) have a tendency to evolve towards a set of target configurations that are small relative to the basin of attraction, and (2) continue to evolve towards the same set of target configurations if perturbed.

If we reach in while the billiard balls are bouncing around and move one of the balls that is in motion, the system will now come to rest in a different configuration. Therefore this is not an optimizing system, because there is no set of target

configurations towards which the system evolves despite perturbations. A system does not need to be robust along all dimensions in order to be an optimizing system, but a billiard table exhibits no such robust dimensions at all, so it is not an optimizing system.

Example: satellite in orbit

Consider a second example of a system that is not an optimizing system: a satellite in orbit around Earth. Unlike the billiard balls, there is no chaotic tendency for small perturbations to lead to large deviations in the system's evolution, but neither is there any tendency for the system to come back to some target configuration when perturbed. If we perturb the satellite's velocity or position, then from that point on it is in a different orbit and has no tendency to return to its previous orbit. There is no set of target configurations towards which the system evolves despite perturbations, so this is not an optimizing system.

Example: a tree

Consider a patch of fertile ground with a tree growing in it. Is this an optimizing system?

Configuration space. For the sake of concreteness let us take a region of space that is sealed off from the outside world — say 100m x 100m x 100m. This region is filled at the bottom with fertile soil and at the top with an atmosphere conducive to the tree's growth. Let us say that the region contains a single tree.

We will analyze this system in terms of the arrangement of atoms inside this region of space. Out of all the possible configurations of these atoms, the vast majority consist of a uniform hazy gas. An astronomically tiny fraction of configurations contain a non-trivial mass of complex biological nutrients making up soil. An even tinier fraction of configurations contain a viable tree.

Target-configuration set. A tree has a tendency to grow taller over time, to sprout more branches and leaves, and so on. Furthermore, trees can only grow so tall due to the physics of transporting sugars up and down the trunk. So we can identify a set of target configurations in which the atoms in our region of space are arranged into a tree that has grown to its maximum size (has sprouted as many branches and leaves as it can support given the atmosphere, the soil that it is growing in, and the constraints of its own biology). There are many topologies in which the tree's branches could divide, many positions that leaves could sprout in, and so on, so there are many configurations within the target configuration set. But this set is still tiny compared to all the ways that the same atoms could be arranged without the constraint of forming a viable tree.

Basin of convergence. This system will evolve towards the target configuration set starting from any configuration in which there is a viable tree. This includes configurations in which there is just a seed in the ground, as well as configurations in which there is a tree of small, medium, or large size. Starting from any of these configurations, if we leave the system to evolve under the natural laws of physics then the tree will grow towards its maximum size, at which point the system will be in one of the target configurations.

Robustness to perturbations. This system is highly robust to perturbations. Consider perturbing the system in any of the following ways:

- Moving soil from one place to another
- Removing some leaves from the tree
- Cutting a branch off the tree

These perturbations might change which particular target configuration is eventually reached — the particular arrangement of branches and leaves in the tree once it reaches its maximum size — but they will not stop the tree from growing taller and evolving towards a target configuration. In fact we could cut the tree right at the base of the trunk and it would continue to evolve towards a target configuration by sprouting a new trunk and growing a whole new tree.

Duality. A tree is a non-dualistic optimizing system. There is no subsystem that is responsible for "doing" the optimization, separately from that which is "being" optimized. Yet the tree does exhibit a tendency to evolve towards a set of target configurations, and can overcome a wide variety of perturbations in order to do so. There are no man-made systems in existence today that are capable of gathering and utilizing resources so flexibly as a tree, from so broad a variety of environments, and there are certainly no man-made systems that can recover from being physically dismembered to such an extent that a tree can recover from being cut at the trunk.

At this point it may be tempting to say that the engine of optimization is natural selection. But recall that we are studying just a single tree growing from seed to maximum size. Can you identify a physical subset of our 100m x 100m x 100m region of space that is this engine of optimization, analogous to how we identified a physical subset of the robot-and-ball system as the engine of optimization (i.e. the physical robot)? Natural selection might be the process by which the initial system came into existence, but it is not the process that drives the growth of the tree towards a target configuration.

It may then be tempting to say that it is the tree’s DNA that is the engine of optimization. It is true that the tree’s DNA exhibits some characteristics of an engine of optimization: it remains unchanged throughout the life of the tree, and physically microscopic perturbations to it can disable the tree. But a tree replicates its DNA in each of its cells, and perturbing just one or a small number of these is not likely to affect the tree’s overall growth trajectory. More importantly, a single strand of DNA does not really have agency on its own: it requires the molecular machinery of the whole cell to synthesize proteins based on the genetic code in the DNA, and the physical machinery of the whole tree to collect and deploy energy, water, and nutrients. Just as it would be incorrect to identify the memory registers containing computer code within a robot as the "true" engine of optimization separate from the rest of the computing and physical machinery that brings this code to life, it is not quite accurate to identify DNA as an engine of optimization. A tree simply does not decompose into engine and object of optimization.

It may also be tempting to ask whether the tree can "really" be said to be undergoing optimization in the absence of any "intention" to reach one of the target configurations. But this expectation of a centralized mind with centralized intentions is really an artifact of us projecting our view of our self onto the world: we believe that we have a centralized mind with centralized intentions, so we focus our attention on optimizing systems with a similar structure. But this turns out to be misguided on two counts: first,

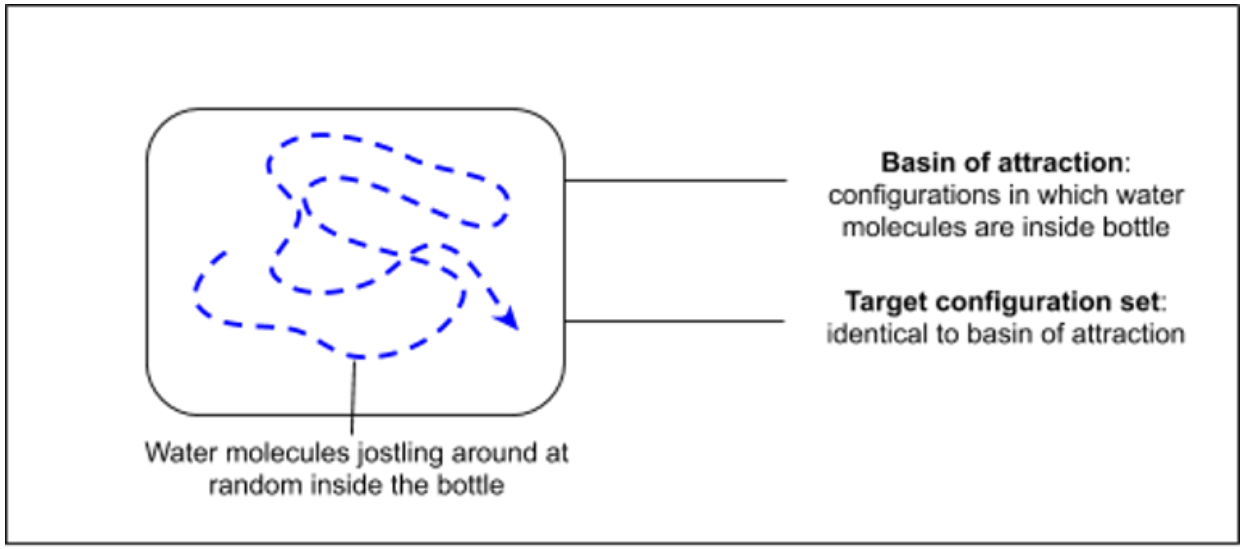
the vast majority of optimizing systems do not contain centralized minds, and second, our own minds are actually far less centralized than we think! For now we put this question of whether optimization requires intentions and instead just work within our definition of optimizing systems, which a tree definitely satisfies.

Example: bottle cap

Daniel Filan has pointed out that some definitions of optimization would nonsensically classify a bottle cap as an optimizer, since a bottle cap causes water molecules in a bottle to stay inside the bottle, and the set of configurations in which the molecules are inside a bottle is much smaller than the set of configurations in which the molecules are each allowed to take a position either inside or outside the bottle.

In our framework we have the following:

- The system consists of a bottle, a bottle cap, and water molecules. The configuration space consists of all the possible spatial arrangements of water molecules, either inside or outside the bottle.
- The basin of attraction is the set of configurations in which the water molecules are inside the bottle
- The target configuration set is the same as the basin of attraction



Configuration space: all possible spatial configurations of water molecules

This is not an optimizing system for two reasons.

First, the target configuration set is no smaller than the basin of attraction. To be an optimizing system there must be a tendency to evolve from any configuration within a basin of attraction towards a smaller target configuration set, but in this case the system merely remains within the set of configurations in which the water molecules are inside the bottle. This is no different from a rock sitting on a beach: due to basic chemistry there is a tendency to remain within the set of configurations in which the molecules comprising the rock are physically bound to one another, but it has no

tendency to evolve from a wide basin of attraction towards a small set of target configuration.

Second, the bottle cap system is not robust to perturbations since if we perturb the position of a single water molecule so that it is outside the bottle, there is no tendency for it to move back inside the bottle. This is really just the first point above restated, since if there were a tendency for water molecules moved outside the bottle to evolve back towards a configuration in which all the water molecules were inside the bottle, then we would have a basin of attraction larger than the target configuration set.

Example: the human liver

Filan also asks whether one's liver should be considered an optimizer. Suppose we observe a human working to make money. If this person were deprived of a liver, or if their liver stopped functioning, they would presumably be unable to make money. So are we then to view the liver as an optimizer working towards the goal of making money? Filan asks this question as a challenge to Yudkowsky's definition of optimization, since it seems absurd to view one's liver as an optimizer working towards the goal of making money, yet Yudkowsky's definition of optimization might classify it as such.

In our framework we have the following:

- The system consists of a human working to make money, together with the whole human economy and world.
- The basin of attraction consists of the configurations in which there is a healthy human (with a healthy liver) having the goal of making money
- The target configurations are those in which this person's bank balance is high. (Interestingly there is no upper bound here, so there is no fixed point but rather a continuous gradient.)

We can expect that this person is capable of overcoming a reasonably broad variety of obstacles in pursuit of making money, so we recognize that this *overall* system (the human together with the whole economy) is an optimizing system. But Filan would surely agree on this point and his question is more specific: he is asking whether the liver is an optimizer.

In general we cannot expect to decompose optimizing systems into an engine of optimization and object of optimization. We can see that the system has the characteristics of an optimizing system, and we may identify parts, including in this case the person's liver, that are necessary for these characteristics to exist, but we cannot in general identify any crisp subset of the system as that which is *doing* the optimization. And picking various subcomponents of the system (such as the person's liver) and asking "*is this the part that is doing the optimization?*" does not in general have an answer.

By analogy, suppose we looked at a planet orbiting a star and asked: "which part here is *doing* the orbiting?" Is it the planet or the star that is the "engine of orbiting"? Or suppose we looked at a car and noticed that the fuel pump is a complex piece of machinery without which the car's locomotion would cease. We might ask: is this fuel pump the true "engine of locomotion"? These questions don't have answers because they mistakenly presuppose that we can identify a subsystem that is uniquely

responsible for the orbiting of the planet or the locomotion of the car. Asking whether a human liver is an "optimizer" is similarly mistaken: we can see that the liver is a complex piece of machinery that is necessary in order for the overall system to exhibit the characteristics of an optimizing system (robust evolution towards a target configuration set), but beyond this it makes no more sense to ask whether the liver is a true "locus of optimization".

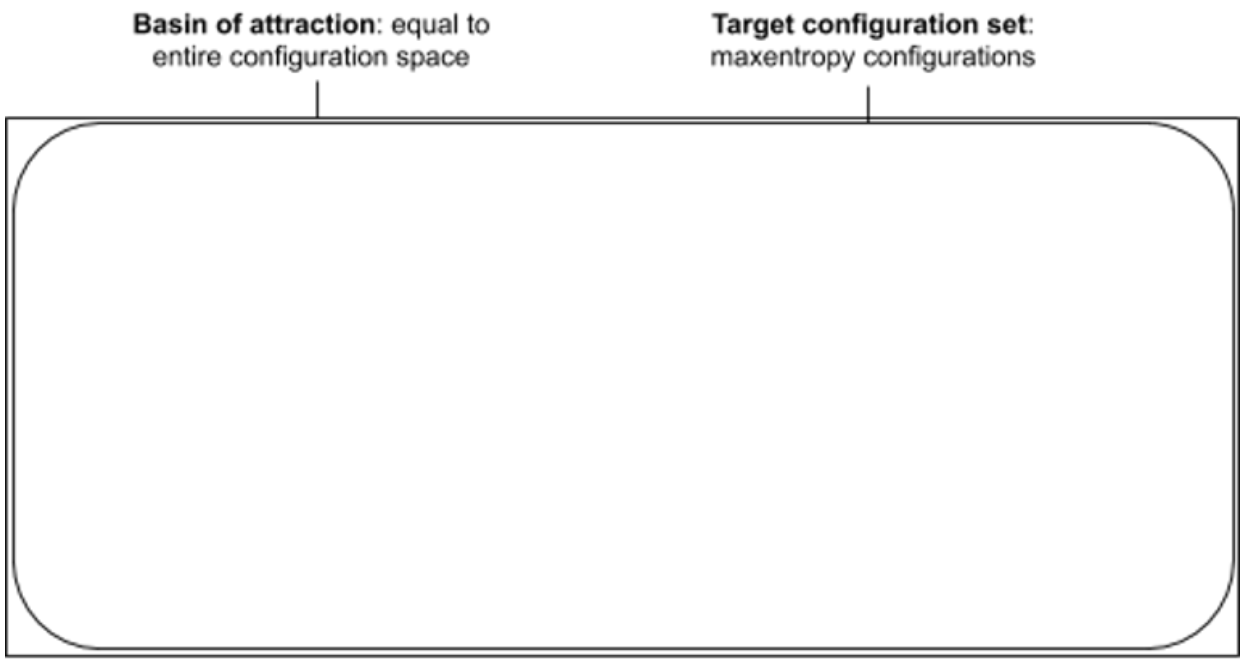
So rather than answering Filan's question in either the positive or the negative, the appropriate move is to dissolve the concept of an optimizer, and instead ask whether the overall system is an optimizing system.

Example: the universe as a whole

Consider the whole physical universe as a single closed system. Is this an optimizing system?

The second law of thermodynamics tells us that the universe is evolving towards a maximally disordered thermodynamic equilibrium in which it cycles through various maxentropy configuration. We might then imagine that the universe is an optimizing system in which the basin of attraction is all possible configurations of matter and energy, and the target configuration set consists of the maxentropy configurations.

However, this is not quite accurate. Out of all possible configurations of the universe, the vast majority of configurations are at or close to maximum entropy. That is, if we sample a configuration of the universe at random, we have only an astronomically tiny chance of finding anything other than a close-to-uniform gas of basic particles. If we define the basin of attraction as all possible configurations of matter in the universe and the target configuration set as the set of maxentropy configurations, then the target configuration set actually contains almost the entirety of the basin of attraction, with the only configurations that are in the basin of attraction but not the target configuration set being the highly unusual configurations of matter containing stars, galaxies, and so on.



For this reason the universe as a whole does not qualify as an optimizing system under our definition. (Or perhaps it would be more accurate to say that it qualifies as an extremely weak optimizing system.)

Power sources and entropy

The second law of thermodynamics tells us that any closed system will eventually tend towards a maximally disordered state in which matter and energy is spread approximately uniformly through space. So if we were to isolate one of the systems explore above inside a sealed chamber and leave it for a very long period then eventually whatever power source we put inside the sealed chamber would become depleted, and then eventually after that every complex material or compound in the system would degrade into its base products, and then finally we would be left with a chamber filled with a uniform gaseous mixture of whatever base elements we originally put in.

So in this sense there are no optimizing systems at all, since any of the systems above evolve towards their target configuration sets only for a finite period of time, after which they degrade and evolve towards a maxentropy configuration.

This is not a very serious challenge to our definition of optimization since it is common throughout physics and computer science to study various "steady-state" or "fixed point" systems even though the same objection could be made about any of them. We say that a thermometer can be used to build a heat regulator that will keep the temperature of a house within a desired range, and we do not usually need to add the caveat that eventually the house and regulator will degrade into a uniform gaseous mixture due to the heat death of the universe.

Nevertheless, two possible ways to refine our definition are:

1. We could stipulate that some power source is provided externally to each system we analyze, and then perform our analysis conditional on the existence of that power source.

2. We could specify a finite time horizon and say that "a system is an optimizing system if it tends towards a target configuration set up to time T".

Connection to dynamical systems theory

The concept of "optimizing system" in this essay is very close to that of a dynamical system with one or more attractors. We offer the following remarks on this connection.

- A general dynamical system is any system with a state that evolves over time as a function of the state itself. This encompasses a very broad range of systems indeed!

- In dynamical system theory, an attractor is the term used for what we have called the target configuration set. A fixed point attractor is, in our language, a target configuration set with just one element, such as when computing the square root of two. A limit cycle is, in our language, a system that eventually stably loops through a sequence of states all of which are in the target configuration set, such as a satellite in orbit.

- We have discussed systems that evolve towards target configurations along some dimensions but not others (e.g. ball in a valley). We have not yet discovered whether dynamical systems theory explicitly studies attractors that operate along a subset of the system’s dimensions.

- There is a concept of "well-posedness" in dynamical systems theory that justifies the identification of a mathematical model with a physical system. The conditions for a model to be well-posed are (1) that a solution exists (i.e. the model is not self-contradictory), (2) that there is a unique solution (i.e. the model contains enough information to pick out a single system trajectory), and (3) that the solution changes continuously with the initial conditions (the behavior of the system is not too chaotic). This third condition may present an interesting avenue for future investigation as it seems related to but not quite equivalent to our notion of robustness since robustness as we define it additionally requires that the system continue to evolve towards the same attractor state despite perturbations. Exploring this connection may present an interesting avenue for future investigation.

Conclusion

We have proposed a concept that we call "optimizing systems" to describe systems that have a tendency to evolve towards a narrow target configuration set when started from any point within a broader basin of attraction, and continue to do so despite perturbations.

We have analyzed optimizing systems along three dimensions:

- Robustness, which measures the number of dimensions along which the system is robust to perturbations, and the magnitude of perturbation along these dimensions that the system can withstand.

- Duality, which measures the extent to which an approximate "engine of optimization" subsystem can be identified.
- Retargetability, which measures the extent to which the system can be transformed via microscopic perturbations into an equally robust optimizing system but with a different target configuration set.

We have argued that the "optimizer" concept rests on an assumption that optimizing systems can be decomposed into engine and object of optimization (or agent and environment, or mind and world). We have described systems that do exhibit optimization yet cannot be decomposed this way, such as the tree example. We have also pointed out that, even among those systems that can be decomposed approximately into engine and object of optimization (for example, a robot moving a ball around), we will not in general be able to meaningfully answer the question of whether arbitrary subcomponents of the agent are an optimizer not (c.f. the human liver example).

Therefore, while the "optimizer" concept clearly still has much utility in designing intelligent systems, we should be cautious about taking it as a primitive in our understanding of the world. In particular we should not expect questions of the form "is X an optimizer?" to always have answers.

Growing Independence

Note: this is based on my experience with my two kids, currently four and six. It may not generalize as much as I think it does.

People start out dependent on their parents for food, changing, contact, motion, and even sleep timing. Typically they end up as adults, no longer dependent on their parents at all. Part of my approach to parenting has been that I want to let my kids be as independent as possible, as early as possible. Not only does it make their lives better, because they can meet their own needs how they want, but it makes my life easier, because they can handle more on their own. Sometimes this involves a bit more effort up front, but I think it's substantially less effort in total.

Examples:

- If Lily (6y) comes to me and says "Anna (4y) pushed me," my first response will probably be "have you talked to Anna?" I'll still help some, often by listening to them negotiate and clarifying rules ("you can't push people, even when they happen to be between you and your desired toy") but over time they've gotten much better at this. There's a whole post worth of thoughts that could go here on what's worked and what hasn't, but at this point they can get up an hour before we do and (nearly always) resolve their own conflicts without waking us.

- The kids will often ask for help while I'm cooking. If I'm in the middle of something, which I usually am, I'll say something like "I can help you as soon as I finish mixing this". During that time they're often able to solve their own problem. If they do still need help when I'm ready, they get my full attention. This is acting as a cost, paying with their time, which filters their requests so I only get the ones where it's worth it to them. And then while they're bored waiting for me they'll often try a bit harder at doing up the snaps on their shirt or whatever, and often that extra focused effort is what they need to do it on their own.

  Similarly, when Anna was learning to ride her trike and she got to a sidewalk bump that was hard to pedal over, she would call for help. I found that if I walked far enough behind her she would keep trying while she waited for me to catch up, and then often didn't need me by the time I was there.

- If I hear crying, I don't automatically do something about it. As I was writing this post I heard Anna get up. Then I heard some crying. Not "I've been badly hurt crying" but some sort of frustration. It didn't last very long, and I didn't move, just listening. A few minutes later Anna came down and said good morning. She had a lot she wanted to tell me about the clothes she had picked out. When she was done I asked what they crying had been, and whether she was ok, and she said that Lily hadn't been willing to come out and play with her even though her bedroom light was on. I clarified (again... this one keeps coming up) that Lily isn't required to play with her, and that even if someone's light is on that doesn't necessarily mean they want to come out of their room and get up. Then we cuddled up and read a book together.

- When the kids started being able to climb things, I would spot them. Often they wanted me to lift them or support them in their climbing, and I wouldn't. They would also want to be lifted down at the end, but the rule would be "if you can

climb up, you can climb down." I was willing to give them advice or guide their foot when they couldn't see where to place it, but they still needed to do the climbing. At this point I'll spot them if they ask me to, or maybe say things like "if you're going to climb that high you need to find an adult to spot you." With tree climbing I'm willing to be a stepstool if asked, but I won't lift them.

- Recently Lily dropped her fork, and asked me to pick it up. I said that this seemed like the sort of thing she could do? Anna volunteered to pick it up for her, and was very happy to be helpful. This wasn't what I was going for, but a different nice outcome.

- I was out with both kids, and Lily wanted to go home while Anna wanted to keep picking dandelions. We were on our block, around the corner from our house. Lily and I talked about how she could go home: she would walk home (no street crossings needed and she knew where to go) and ring the doorbell. If someone let her in she was set, otherwise she would walk back to where I was. After Lily set off I posted in the house chat that Lily would be ringing the doorbell soon, and once Lily was inside Julia replied to let me know.

- We noticed that Anna kept giving or trading things to Lily, and then regretting it. For example, she gave Lily an elephant stuffy she got for her birthday, and then talked for months about how she was sad and wished it were still hers. We talked to Lily about this, and told the kids that any gifts Anna made were provisional, and she had three days to change her mind. We also told them that if they wanted to make permanent trades they needed to bring the proposed trade to a grownup first, who could jog Anna's memory ("Anna, do you remember how you felt when you...") and make sure they really wanted to go through with it. I think the first rule never ended up getting used, and the second rule got used maybe once? We've since let both rules fade away.

- When the kids were little they would sometimes ask for a drink of water in the middle of the night. As soon as they were old enough that we trusted they wouldn't spill it, we gave them sippy cups of water to keep by their beds. The changed the frequent "I'm thirsty" for the less frequent "my water cup is empty." And, better, they started checking their cup when going to bed, usually telling us when it was running out. A few weeks ago Anna woke us up, for the first time in a while: her cup was empty. I told her I wouldn't fill her cup, but described how she could get a drink from the bathroom. She was mad that I wouldn't do it for her, but after I went back to bed I heard her walk out, get a drink, and go back to bed. She hasn't woken us since.

- About a year ago I brought Lily to an amusement park. Near the end of the day there was a roller coaster she wanted to go on, but it was too scary for me, so I told her I wouldn't go on it with her. She asked if she could go on it by herself, but you needed to be 48" to ride alone. She told me she was going to find someone else to ride with her, and I didn't object. She wandered a bit with me until she identified someone who I think she thought was sufficiently non-threatening (middle-aged woman hanging out with family) and Lily asked me if I would be willing to ask on her behalf. I declined, expecting Lily would be too shy, but Lily went up, explained the situation, and asked if they would go with her. They were a bit confused, confirmed the situation with me, asked me if I was ok with it, and I emphasised that it really was fine if they said no. They decided to do it, and as far as I could tell both had a really good time.

- When Lily was ~1.5, she was just learning to walk and was standing at the top of a short flight of stairs. I was below her, in a place where I could catch her if she fell, but as she continued looking around and seeming stable I started playing my mandolin which I was wearing on a strap. I wasn't expecting she would fall, but she did, and while I dropped the mandolin and went to catch her I didn't get her fully and she bonked her chin. She lost two teeth, and I'm sure it hurt a lot. This is probably the event I most regret in parenting so far, and pushed me in the more cautious direction.

- As soon as our kids could walk we started teaching them how to stay out of the street. This was some work, but when we fully trusted that they would stop at the corner they gained the freedom to run ahead on their own. When they were little I couldn't let them get too far ahead, though, or other adults who didn't know that these particular kids knew to stay out of the street would get worried and try to protect them.

- Our house has big heavy doors, which means the kids can't get out by themselves. I made a kid door out to the back yard, and put kid-height railings on the steps. Now if they want to go outside on their own they can.

- One time we couldn't find Lily. We looked all over the house, and she was just not there. When I saw the kid door was unlocked, that told me she'd gone out, but then she wasn't in the back yard. Apparently she'd left? This was very unlike her, and we were worried, though I knew she wouldn't have gone far since she wouldn't cross any streets. I started to run around the block, and just as I'd gone around the first corner I saw Lily happily running from the other direction. She'd decided she'd like to run around the block on her own as an adventure. I talked with her about how it wasn't ok to go off on her own like that yet, and next time she should check in with a grownup first. When we put in the kid door I should have been clearer with them about how they needed to stay in our yard.

- Lily asked me if she could cut her own hair, and I explained that kids who cut their own hair generally end up with hair they're unhappy with. She asked if I would cut it; I declined. She asked our housemate Ruthie if she would cut it, Ruthie checked with me ("it's Lily's hair, so it's fine with me") and Ruthie gave her a nice cut that Lily was very happy with. Julia was out at the time, and when she came home she was upset that I had let Lily cut her hair on a whim. Asking her now she wrote, "I'm less in favor of giving the kids free rein here because I think it's important to make sure any long-lasting choices are made in an informed way, and I don't think anyone made sure Lily understood that for the next several months she wouldn't be able to do some of the hairstyles she sometimes requested. I also likely would have required a waiting period of a week or two to see if she still wanted it. As someone who hated spending third grade growing out my bangs, the possible downside of haircut decisions is more salient to me than it is to Jeff." Afterwards, we talked for a while trying to find other places where we might have similar disagreements about what to let the kids do (tattos? piercings? cutting up their clothes?) This is a good illustration of how it's important to be on the same page as your partner about what you're ok with letting the kids do.

- Anna and I were out with her trike, and she asked me to carry it home for her. We weren't very far from the house, and I declined. She said she was just going to leave it there. I told her that if she left the trike it would be available for

anyone to take. And that I would probably take it, but it would then be my trike. She decided to ride her trike home.

- When we started wearing covid masks, the kids didn't want them. I explained that the recommendations had changed, we were trying to help build a norm of mask-wearing, masks keep people from spreading their germs, and even our family could have the coronavirus without realizing yet. This was enough for Lily, who's pretty pro-social, but Anna didn't like having something on her face. I told her that if she wasn't willing to wear a mask she'd have to stay on our property, which meant inside or in the back yard. She initially (firmly!) said she was fine with that, but when she realized this meant she wouldn't be able to ride her trike around the block she changed her mind and asked me to help her put her mask on.

  A few weeks later, when wearing masks was routine, Anna was still asking us to put her's on each time. I thought she could probably start doing it for herself, when she next asked if I'd put it on her I said "can you do it?" She said no. I said that I thought she was big enough to do it herself, and that I was still willing to do it I needed her to try doing it herself first. With a bit of coaching (one ear, then the other) she got it on without any other help from me. She was so proud! Since then she's managed it herself every time.

  This morning I was out with the kids and I noticed that Anna's mask was around her neck. "Anna, your mask?" "Papa, I'm doing the honeysuckle, and my mask gets in the way." "Ok, as long as you put it right back on when you're done."

- I rarely tell the kids "no". Instead, if it's something that I don't think is a good idea or won't work, I explain why. "The last time you climbed a tree not wearing pants you scraped your legs a lot, and were pretty sad about it." "If you want to use a sharp needle you'll need to find an adult who's willing to supervise so you don't stab yourself." "If you sign up to cook dinner you need to make sure you prepare enough food for everyone, with food everyone can eat." "If you want make that much of a mess you'll need to find a grownup who'll commit to cleaning up if you don't."

- There are still some hard constraints. They have to go through their bedtime routine and go to bed. They have to sit at the table for their meals (though we don't make them eat, just spend at least 10min in front of their food). We try to make these really predictable, and we'll use counting and timeouts if they're not following them. Any consequence we're imposing should happen as quickly as possible, because that lets you use much weaker consequences for the same amount of behavior change.

Some common threads:

- I'm willing to invest large amounts of time in teaching and advice, but won't do things for them unless I'm pretty sure they can't do it themselves.

- I'm happy to talk in detail about why we do things the way we do, and am open to being convinced in cases where they think the rules should be different.

- I want them to be practicing making decisions and living with the consequences, but not beyond what's currently safe for them or beyond what they can productively learn from. When I think they're making a bad decision I'll try to

bring up information I think they're overlooking, but I'll only very rarely take the choice away from them.

- I let them solve their own problems, and let them practice figuring out when to bring in help.

I have three main motivations here. The first is teaching: eventually they'll need to make good decisions on their own, and the sooner they start the more practice they'll be able to get. The second is a kind of long-term laziness: once they can do things for themselves it's less work for me. And the third is respect: they're people and as much as possible they should get to choose how their lives go.

We've built Connected Papers - a visual tool for researchers to find and explore academic papers

Hi LessWrong. I'm a long time lurker and finally have something that I'm really proud to share with you.

After a long beta, we are releasing Connected Papers to the public!

Connected papers is a unique, visual tool to help researchers and applied scientists find and explore papers relevant to their field of work.

First - let's look at a couple of examples graphs for work that is representative of this community:

Nick Bostrom:

https://www.connectedpapers.com/main/7bba95b3d145564025e26b49ca67f13f884f8560/Superintelligence-Paths-Dangers-Strategies/graph

Eliezer Yudkowsky, Nate Soares:

https://www.connectedpapers.com/main/61c368138b0211323e8773174ac3132122da07ef/Functional-Decision-Theory-A-New-Theory-of-Instrumental-Rationality/graph

Did you find new and interesting papers to read? Would this be helpful as an introduction to the literature of a new field of study?

The problem

Almost every research project in academia or industry involves phases of literature review. Many times we find an interesting paper, and we'd like to:

- Find different methods and approaches to the same subject
- Track down the state of the art research in the field
- Identify seminal works and background reading
- Explore and immerse ourselves in the topic and become aware of the trends and dynamics in the literature

Previously, the best ways to do this were to browse reference lists, or hope to find good keywords in textual search engines and databases.

Enter Connected Papers

It started as a side project between friends. We've felt the pains of academic literature review and exploration for years and kept thinking about how to solve it.

For the past year we've been meeting on weekends and prototyping a tool that would allow a very different type of search process for academic papers. When we saw how much it improved our own research and development workflows — and got increasingly more requests from friends and colleagues to use it — we committed to release it to the public.

You know… for science.

So how does it work?

Connected Papers is not a citation tree. Those have been done before.
In our graph, papers are arranged according to their similarity. That means that even papers that do not directly cite each other can be strongly connected and positioned close to each other in the graph.

To get a bit technical, our similarity is based primarily on the concepts of *co-citation* and *bibliographic coupling* (aka co-reference). According to this measure, two papers that have highly overlapping citations and references are presumed to have a higher chance of treating a related subject matter.

Reading the graph

Our graph is designed to make the important and relevant papers pop out immediately

With our layout algorithm, similar papers cluster together in space and are connected by stronger lines (edges). Popular papers (that are frequently cited) are represented by bigger circles (nodes) and more recent papers are represented by a darker color.

So for example, finding an important new paper in your field is as easy as identifying the dark large node at the center of a big cluster.

List view

In some cases it is convenient to work with just a list of connected papers. For these occasions, we've built the List view which you can access by clicking "Expand" at the top of the left panel. Here you can view additional paper details as well as sort and filter them according to various properties.

Prior and derivative works

The Prior works feature lists the top common ancestral papers for the connected papers in the graph. It usually includes **seminal works** in the field that heavily influenced the next generation.

Meanwhile, the Derivative works feature is the opposite: it shows a list of common descendants of the papers in the graph. It usually includes relevant **state of the art** papers or **systematic reviews and meta-analyses in** the field.

We have found these features to be especially useful when we have a paper from one era of research and we would like to be directed to the preceding and succeeding generations of research on the same topic

Help us spread the word

Connected Papers will only grow by word of mouth. **Please share Connected Papers in your scientific community!**

We are very eager to see how the broader academic community adopts and responds to this tool. We welcome all forms of feedback and would love to brainstorm together about how it can further evolve and improve.

Simulacra Levels and their Interactions

Previously: Covid-19: My Current Model, On Negative Feedback and Simulacra

This post aims to unpack and explain simulacra levels of action using the threat of covid-19 as its central example. My intention is for future posts to then apply this model to many covid-related dynamics.

In Elizabeth's Negative Feedback and Simulacra, she examined several example situations on which information was being processed on multiple simulacra levels at once. On Negative Feedback and Simulacra was my take on those examples.

To re-familiarize ourselves with the simulacra levels, here's the introduction Elizabeth offered to them in her post:

> My friend Ben Hoffman talks about simulacra a lot, with this rough definition:
>
> First, words were used to maintain shared accounting. We described reality intersubjectively in order to build shared maps, the better to navigate our environment. I say that the food source is over there, so that our band can move towards or away from it when situationally appropriate, or so people can make other inferences based on this knowledge.
>
> 2. The breakdown of naive intersubjectivity – people start taking the shared map as an object to be manipulated, rather than part of their own subjectivity. For instance, I might say there's a lion over somewhere where I know there's food, in order to hoard access to that resource for idiosyncratic advantage. Thus, the map drifts from reality, and we start dissociating from the maps we make.
>
> 3. When maps drift far enough from reality, in some cases people aren't even parsing it as though it had a literal specific objective meaning that grounds out in some verifiable external test outside of social reality. Instead, the map becomes a sort of command language for coordinating actions and feelings. "There's food over there" is perhaps construed as a bid to move in that direction, and evaluated as though it were that call to action. Any argument for or against the implied call to action is conflated with an argument for or against the proposition literally asserted. This is how arguments become soldiers. Any attempt to simply investigate the literal truth of the proposition is considered at best naive and at worst politically irresponsible.
>
> But since this usage is parasitic on the old map structure that was meant to describe something outside the system of describers, language is still structured in terms of reification and objectivity, so it substantively resembles something with descriptive power, or "aboutness." For instance, while you cannot acquire a physician's privileges and social role simply by providing clear evidence of your ability to heal others, those privileges are still justified in terms of pseudo-consequentialist arguments about expertise in healing.
>
> 4. Finally, the pseudostructure itself becomes perceptible as an object that can be manipulated, the pseudocorrespondence breaks down, and all assertions are

> nothing but moves in an ever-shifting game where you're trying to think a bit ahead of the others (for positional advantage), but not too far ahead.

If that doesn't make sense, try this anonymous comment on the post

> Level 1: "There's a lion across the river." = There's a lion across the river.
>
> Level 2: "There's a lion across the river." = I don't want to go (or have other people go) across the river.
>
> Level 3: "There's a lion across the river." = I'm with the popular kids who are too cool to go across the river.
>
> Level 4: "There's a lion across the river." = A firm stance against trans-river expansionism focus grouped well with undecided voters in my constituency.

Almost everyone would rather not be eaten by a lion. I certainly would rather not be eaten by a lion.

Whether or not I am eaten by a lion still does not drive much of my decision making. It is highly implausible that I will be eaten by a lion.

I hope that if in the future it becomes plausible that I may get eaten by a lion, how to not get eaten by a lion would then drive much of my decision making.

If the presence of lions in various places would not put anyone in any danger, that makes it much less expensive for me to be wrong about where they are. The less people are concerned about the consequences of having or inflicting incorrect object-level models of the world, the less concerned they will be with Level 1 and with the Level 1 accuracy of their statements.

The prioritization of various simulacra levels becomes a habit. If you are used to interpreting "There's a lion across the river" almost entirely as "I'm with the popular kids who are too cool to go across the river," because that's what it almost always means in your village, it may be very difficult for someone to say "No, really, I'm not associating with the cool kids right now. There's literally an actual lion across the actual river and if you cross the river you will die."

There is no good way to sacrifice the cool points in order to communicate the presence of a lion. Even if it works at first, soon there will be a tendency for the new wording to become the canonical form of "I'm with the popular kids who are too cool to go across the river."

If everyone's instinct is to interpret "There's a lion across the river" as *both* "There is an actual lion across the actual river" and *also* "I'm with the kids who are too cool to cross the river" then there is a chance.

There is still a barrier. Whoever wants to share knowledge of the lion will become less cool by doing so. Ideally, for high enough stakes, this stops being a problem in multiple ways. If lives are at stake, especially one's own or one's loved ones, being cool looks less important than avoiding the lion. Ideally, being the person who saved us from the lion is also considered kind of cool, allowing one to both starve lions and look cool. That only works if everyone realizes the lion was there. But the payoff could be very large. So there's a chance.

Whereas, if things are too forsaken, one *loses the ability to communicate about the lion at all.* There is no combination of sounds one can make that makes people think there is an actual lion across an actual river that will actually eat them if they cross the river.

I'm not trying to be subtle here. You can guess where this is going.

There's a Virus Across The Ocean

Level 1

"There's a pandemic headed our way from China" means "there's a pandemic headed our way from China."

"There's no pandemic headed our way from China" means "there's **no** pandemic headed our way from China."

Level 2

"There's a pandemic headed our way from China" means "I want you to act as if you think there might be a pandemic on our way from China" while hoping to still be interpreted by the listener as meaning "There's a pandemic headed our way from China."

The speaker hopes to be interpreted as still meaning "There's a pandemic headed our way from China."

This may involve any of the following:

"I want you to click on this headline about a pandemic headed our way from China and/or subscribe to and share my newsletter, website or channel."

"I want you to have more negative affect towards China and/or other foreigners."

"I want you to be afraid."

"I want to shut down economic activity."

"I want to sell you toilet paper, masks and hand sanitizer at premium prices."

"I want you to thank me later for any or all of the above."

"There is no pandemic headed our way from China" means "I want you to act as if you think there is **not** a pandemic headed our way from China."

The speaker hopes to be interpreted as still meaning "There is not a pandemic headed our way from China."

This may include any of the following:

"I want you to click on this headline about there not being a pandemic headed our way from China and/or subscribe to and share my newsletter, website or channel."

"I want you to have more positive affect towards China and/or other foreigners."

“I do not want you to be afraid.”

“I want to stop you from hoarding toilet paper, masks and hand sanitizer.”

“I want to avoid shutting down economic activity.”

“I want you to thank me later for any or all of the above.”

Level 1 vs. Level 2

The key difference is in the intended effect of the statement made, and to what extent the statement is correlated with the true state of the physical world.

Level 2 statements *do not have to be untrue.* Nor need it be something you do not believe. It is often said that “the truth is the best lie.”

Nor need the statement be selfish. Many Level 2 statements really are intended *for the subject’s ‘own good.’*

What makes a statement Level 2 rather than Level 1 is that *you don’t care whether or not it is true*. Instead, you care about what actions it causes people to take, and whether or not you like those actions.

If there is a *tiger* across the river, but the group is insufficiently afraid of tigers, you might claim there is a lion across the river instead.

If there is a tiger across the river, but you have a shotgun and are itching to get a tiger’s head as a trophy, you might claim there is not a lion across the river, so that we will cross the river. Whether or not you are aware of a lion across the river doesn’t matter. A lion would prevent you from hunting tigers. You want to hunt a tiger.

The Four Direct Communicators

Sticking to the first two levels for now, we can divide into quadrants and see five types of communication strategies.

Let V = “There is a virus across the ocean that is likely to cause a pandemic here.”

Let ~V = “There is not a virus across the ocean that is likely to cause a pandemic here.”

Let C = The consequences of being told V.

Let ~C = The consequences of being told ~V.

Let N = The consequences of being told nothing. For simplicity assume everyone either believes that C>N, or that ~C>N.

The Oracle only looks at Level 1. The Oracle says V if and only if The Oracle believes V with sufficient confidence, says ~V if they believe ~V with sufficient confidence, and says nothing or “I don’t know” otherwise. They care not whether C>~C or ~C>C.

The Trickster only looks at Level 2. The Trickster says V if they think C>~C. They say ~V if they think ~C>C. Otherwise they say nothing. They care not whether V is true.

The Nihilist cares about neither Level 1 nor Level 2. They say whatever they feel like saying, then eat at Arby's.

The Sage looks at *both* Level 1 and Level 2 and avoids actions that violate either principle. They say V if *both* they believe V *and* they believe C>~C. They say ~V if *both* they believe ~V *and* they believe ~C>C. If they believe that V but ~C>C, or they believe that ~V but C>~C, they say nothing.

The Pragmatist looks at *both* Level 1 and Level 2 and assigns some value to each, then takes action that balances both concerns. They recognize that there is a cost to saying that which is not, but that cost is not infinite. Thus, if The Sage would talk, they talk. If The Sage would not talk, they may talk anyway. Up to a point, they'll speak the truth and take the direct consequences even if those consequences are bad. Past a certain point, they'll be willing to lie.

There's obviously a continuum all around, especially between Sage and Pragmatists. *Almost* everyone has *some* breaking point where they would lie for a sufficiently powerful cause. *Most* people place value on telling the truth beyond known specific and direct consequences, and thus it takes some threshold of bad other consequences to get them to be quiet.

Once you know which of these types a person is, you can trust them in some sense.

It's easy to interpret an Oracle or a Sage. When a Sage is silent, you can trust that either they don't know the answer, or they believe telling you the answer would be bad. Often, absent strong glomarization, this lets you figure out the answer.

But it's also easy, *once you know they're a trickster*, to trust a Trickster in their own way. You simply interpret their statements as manipulations rather than observations.

Level 3

"There's a pandemic headed our way from China" means "I wish to associate with the group that claims there is a pandemic headed our way from China."

This may include any of the following:

"I want to affirm my membership in my in-group, and that I dislike the out-group."

"I want to be part of the group of people with power, and/or who are winning."

"I want to be seen serving those with power and spreading their messages."

"I want to be seen as smart, on the ball, ahead of the curve, scientific and other neat stuff like that."

"I want to be part of the group that cares about people."

"I want to be part of the group that believes/defies experts."

"I want to be part of the group of so-called 'responsible experts' on this."

"I want to be part of the group that isn't afraid to tell you hard truths."

"I want to be part of the group that doesn't have bad traits like racism."

“The high status move is to endorse this position at this time.”

“The publication I write for wants to hear this.”

And so on.

“There’s no pandemic headed our way from China” means “I wish to associate with the group that claims there is **not** a pandemic headed our way from China.”

It may include any or all of exactly the same things, depending on your local situation, and where you are in the timeline.

The polarity of many of these motivations has changed. In some cases, it has changed multiple times.

Level 4

“There’s a pandemic headed our way from China” means “It is advantageous for me to say there is a pandemic headed our way from China.”

“This set of verbal incantations focuses attention on things where focused attention helps me, and away from places where focused attention hurts me.”

“This set of verbal incantations will make people think I am associated and allied with those who it is advantageous for them to think I am associated and allied with.”

“This set of verbal incantations associates me with good words and emotions, or my enemies with bad words and emotions.”

“It would be advantageous for me if the group I am associated with is viewed as advocating the claim that there is a pandemic headed our way from China.”

“This claim fits the heuristics of claims that will make me look responsible.”

“This claim fits the heuristics of claims that will make me look strong and/or powerful and/or in the know and/or a winner.”

“This claim fits the heuristics of claims that give one leverage and/or power.”

“This claim fits the heuristics of claims that will make me someone others think that others will view as a valuable ally, especially others also operating at Level 4.”

“This claim fits the heuristics of claims that will stop me from being scapegoated.”

“This claim fits the heuristics of creating optionality, and/or putting rivals and opponents into situations where they will look bad.”

“There’s no pandemic headed our way from China” means the same thing, except in this situation the additional incantation ‘no’ seems appropriate.

Note that the Level 4 actor has in an important sense lost the ability to think or plan.

It might or might not impact this calculation whether or not your statement is true (Level 1), or whether it will be believed (Level 2), or what coalitions this statement signals your membership in or support of (Level 3). The primary way in which Level 1+2 considerations impact this decision is indirectly, through their impact on Level 3.

That’s all part of a calculation, and matters only to the extent that it effects the consequences of saying the thing.

I find Level 4 is the hardest to grok, by far. It does not come naturally to me.

A potentially easier way to understand Level 4 is to think about how it fits into the contrasts between the first three levels, as will now be discussed.

Level 1 vs. Level 2 vs. Level 3

Consider the first three levels of action and consequence. Level 1 cares about the object level. Level 2 cares about the consequences of changing perception of the object level. Level 3 cares about which coalition your statement associates you with, and which coalitions would approve or disapprove of your statement.

We can think of each of the possible three contrasts. Is Oceania at war with East Asia, Eurasia or both?

We can then add Level 4 into the mix, wherever it belongs based on that division.

Level 1 vs. Levels 2+3+4: Truth versus Untruth

This division feels the most natural to me and people similar to me. Here, Level 4 actions definitely fall into the 2+3 camp, with this full division being level 1 vs. levels 2+3+4.

Level 1 statements correspond to object-level truth. If everyone makes level 1 statements, everyone’s map improves. Decisions get better. Action that has good object-level consequences can be taken. People can be trusted in a pure sense.

Higher level statements corrupt all that. Trust is destroyed. People’s beliefs no longer converge towards the truth. Actions taken are what those with power, those who manipulate and form alliances more effectively, think they want to happen. Since they don’t have the true picture of the situation either, they often don’t like those consequences, and regret their choices to the extent that they care about those consequences.

What difference, this perspective asks, does it make whether you said “Don’t prepare, there’s nothing to worry about” because you wanted to save masks for health care workers, or be able to buy up all the hand sanitizer for resale, or because you were worried about prejudice against Asians, or you wanted to keep the economy open, or you wanted to look responsible and calming rather than irresponsible and alarmist, or that’s what all the authoritative media and government sources were saying and you wanted to be seen as someone who holds the party line?

(Or, at Level 4, if you believe that it will increase the power of your group to be seen as advocating for this position, because it will improve your image or it is popular?)

From this perspective, the only important thing is, *you didn’t care if what you said was true.* You said it because it was useful to you to say it. That’s what matters.

Thus, we can repeat the 2×2 from above. The only difference is now ‘the consequences’ include coalition politics and other more abstract things.

It is easy to see why primarily Level 1 activities are helpful in dealing with Covid-19. It is also easy to see why one might view all primarily Level 2, 3 and 4 activities as assumed to be unhelpful.

Level 1+3 vs. Levels 2+4: Authentic vs. Inauthentic

This division can definitely be weird when first pointed out or considered, but it makes a decent amount of sense.

If I say that V because I want to show that I am a member of the group that believes V, then that is a signal of group membership rather than evidence for V. But one can see this as an *authentic* signal of group membership. I *really do* wish to associate with the V-advocates and not with the anti-Vs.

The risk when sending this signal honestly is that one can confuse my statement of group membership with a claim of V.

I can be making an 'honest' statement about which coalition I am supporting, and you can get the wrong impression that V is true.

Or, I could be making an honest statement that V, and you could get the impression that I wish to belong to the V-advocating coalition. This also distorts my map of reality in a potentially dangerous way.

If you can tell when someone is engaging in Level 3 actions versus Level 1 actions, then one can preserve the sanctity and trustworthiness of the Level 1 actions.

The Oracle who only cares about Level 1 is easy to interpret.

So is **The Drone**, who only cares about Level 3 and will be discussed below. The drone's claim tells you their belief of what their side is currently advocating. No more, no less.

Alternatively, one can be an advocate for a side that presents the case in the best possible light, while only making true statements. This is a variation on the 1+2 strategy of The Sage. One can call this communication strategy **The Lawyer**. The Lawyer will say only things that have positive impact on level 1 *and* positive impact on level 3. Alternatively, they might say only things that have positive impact on level 1 *and* positive combined impact on levels 2 and 3. Or they might need to fulfill all three requirements.

By contrast, statements on Level 2 or Level 4 can be entirely false.

Level 2 does this to manipulate your Level 1 map, and therefore your actions.

Level 4 does this as part of a system that manipulates your Level 3 map, and therefore your actions, believing that this is what drives human action.

If such considerations can dominate, or frequently do, then everything becomes a game.

In this perspective, primarily Level 3 actions are a positive driving force. Such considerations can motivate humans to align with accurate maps and helpful behaviors, under at least some conditions where pure object-level considerations

would not work. This is one way we coordinate around washing our hands, locking down or wearing a mask.

Whereas Level 2 and Level 4 actions distort that, and are what lead to inaccurate maps and unhelpful actions.

Levels 1+2 vs. Levels 3+4: Facts vs. Politics

This is natural in the sense that it's higher levels versus lower levels. One can view the first two levels as *caring about the object level at all.*

You might lie. But you're lying because truth matters, beliefs determine physical actions and actions have consequences. Your desire for actions that follow from bad maps of the underlying territory is unfortunate. But at least there are maps of the territory involved, however flawed, and people are trying to cause actions that have consequences they themselves want to occur, even if those consequences are not good for others.

One can view Level 2 statements and actions as a sort of corruption of Level 1, but one still grounded in reality. They're fighting dirty, but they're still fighting. One can still speak one's mind, there is still a marketplace of ideas and over time truth retains a competitive advantage.

Whereas Level 3 is an entirely different thing that has nothing but contempt for the idea that facts matter, or actions have consequences distinct from how they are viewed by others.

Hence the claim that "Facts Don't Matter."

Facts Don't Matter signifies that it does not matter if it is common knowledge that someone is lying.

Thus, Facts Don't Matter is the dominance of level 3+4 considerations over level 1+2 considerations. Why should I care if the words I say correspond accurately to the physical world's past, present or future? What matters is their impact on my membership in my coalition, and the success of myself and that coalition in playing political games.

An embodiment of this distinction is the resonance of the statement "I Demand A Plausible Lie." This is a request to cease purely Level 3+4 behavior and at least adapt some Level 2 considerations. It insists that one be allowed to maintain a map of reality at all outside of politics, and that political considerations be bound at least a little by reality.

To the purely political actor, the implausible lie is better. If the lie is implausible, then those repeating it have sent a costly signal of loyalty, and cut ties with lower levels. You don't have to worry they repeated the statement because it happens to match the physical world, or that they will refuse to repeat the next one if it fails to match.

Note also that if you only are playing politics, you might be able to act directly in a way that has a direct effect. What you cannot do is make or carry out physical plans involving multiple steps. In the best of times actually planning is very hard. This makes it impossible.

Levels 1+2+4 vs. Level 3: The Drone vs. the Agent

When grouping levels 1+2 against level 3, it feels natural to me to put level 4 with level 3. When talking about this with Ben Hoffman, it became clear there was also a view, where it is level 3 rather than level 4 that feels alien and hard to grok, that naturally groups level four instead with the first two levels.

In this view, the Drone, who cares only about level 3 considerations, is the odd one out.

Anyone acting on any other level is an *agent.* They are *acting on* systems. Level 1 acts upon the physical world. Level 2 acts upon other people's models of the physical world. Level 4 acts upon people's models of other people and their dynamics.

Whereas the Drone lacks agency and free will entirely. The Drone does what they see others in their group are doing, says what others in their group are saying. More than anyone else, the Drone is a dead player.

Levels 1+2+3 vs. Level 4: The People vs. The Lizards

One could also group the first three levels together and contrast them with the fourth. This thinking is that there is ordinary decent interaction, humans being human, as represented by the first three levels. Then there are the schemers who prey upon us, twist everything in their sick games and play us against each other. We vote for the lizards, as Douglas Adams reminds us, because if we don't, the wrong lizard might win.

The lizards do not care about Covid-19. It would be a category error to say that the lizards care about things at all. That implies they believe in the existence of things, or prefer one state of those things to another state, and act upon that in some way. That's not their jam.

Instead of having goals and trying to achieve them, the lizards have systems of power accumulation. They follow habits of behavior that move away from potentially blameworthy actions towards ones that will be seen as good, to sculpt perception of them as powerful and their opponents as weak, and so on. On multiple levels they have a complete inability to plan, even more so than in the last section. They are the politician who prepares two speeches, one pro and one anti, and gives whichever sounds better.

This perspective says the problem is mainly the lizards – or, alternatively, that you want to make sure you are one of the lizards. Get rid of the lizards, and you won't get a paradise, but you'll get systems that move towards truth and justice, and that can plan and do useful things.

A Gentle Glossary of Strategies

From here, L-1 is level 1, L-2 is level 2, L-3 is level 3, L-4 is level 4.

Let's summarize the players. This is what they would say. What they would do is similar.

These players are roles that individuals take in situations. Few people will embody one of them at all times in all circumstances. Sometimes you see a lion across the river.

Nothing: The Nihilist says some things, then eats at Arby's.

L-1: The Oracle speaks the truth, even if their voice trembles.

L-2: The Trickster says that which causes beliefs that cause the actions they want.

L-1 and L-2: The Sage says only true things that don't have bad consequences.

L-3: The Drone sings songs and carries signs, mostly saying hurray for our side.

L-1 and L-3: The Lawyer says the true things that comprise the best argument for their position.

L-3 and L-4: The Politician ignores the object level and only considers politics.

L-4: The Lizard trusts their instincts and does that which creates or captures power.

L-All: The Pragmatist balances impact at all levels they are aware of slash care about when deciding what to say.

Several of these roles have important divisions into two or more related but distinct approaches. A key question is whether considerations act as veto points, or if they are weighed against each other. Further discussion is beyond scope here but I hope it happens in the future.

(Think this is missing 1-3 additional roles. Discussion question, what is The Idealist?)

Paths Forward

All actions and statements operate on all four levels at once, to the extent that they have implications on those levels.

The *intent* of a statement is often entirely on one level. That's not how humans or Bayesians interpret an action. If you want to improve the physical world without having any higher-level side effects, that's going to require extra effort. Avoiding meaningful implications on levels 2, 3 and 4 is *hard work.* The same bleeding effect occurs when aiming for higher levels, as well. For concrete discussion of this, see my previous post on simulacra levels.

As I say in the previous section, many of these roles have multiple variations, and have a lot of complexity inside them. Posts that explore them in detail would be worthwhile.

There is also the issue of the overall simulacra level of a group, organization or civilization. What is the default interpretation of information or action? What is the assumed motivation? What does that say about the group's dynamics and its ability to do things? I'm still trying to work through these things. There's clearly a somewhat distinct way of thinking about 3rd and 4th level simulacra that is built around these questions rather than thinking about individual actions. I *suspect* that the two are fully compatible and describe aspects of the same thing, but I'm still working that out and will talk about it more when I better understand it.

It's also possible there are two or more different models that are using the same four-level language structure, that share their concepts of levels one and two but disagree about how level four works, and to some extent about level three. The more we talk about it, and the more concrete we can make our examples, the better we can sort all this out. The important thing is to get models that are useful.

The original intent of this post was to go on to analysis of other issues surrounding Covid-19. I was hoping to make clear what I meant by the more disputed statements in my Covid-19 model summary from two weeks ago, and also how and why I believe those dynamics occurred, and what dynamics one can expect going forward. But this post is long enough, so I've pushed that into future posts.

Wireless is a trap

I used to be an anti-wire crusader. I hated the clutter of cables, and my tendency to unconsciously chew on them if they got anywhere near my face. But running into bug after tricky wireless bug—mostly while trying to make my video calls work better—I've apostasized. The more I've learned about wifi, Bluetooth and related protocols, the more I'm convinced that they're often worse, on net, than wires.

For instance: most people, when their video call stutters, blame their Internet service provider. That's understandable, since most ISPs are overpriced oligopolists with barely-usable software and horrible customer service. However, every time I can remember helping someone track down the source of their connection problems, the culprit has turned out to be their wifi. And often, the easiest fix has been to run a cable.

Wifi (and bluetooth, etc.) sucker you in by making it seem like they "just work." But if you investigate, you'll often find that the wireless link is operating in a degraded state that performs much worse than a wired equivalent. Since this degradation is silent, it's often not obvious that the problem is the wireless—instead, you'll probably conclude that it's your device/software/self.

Over and over again, I've seen people fix some wireless-related problem and go "wow, I had no idea how much better this could be!"

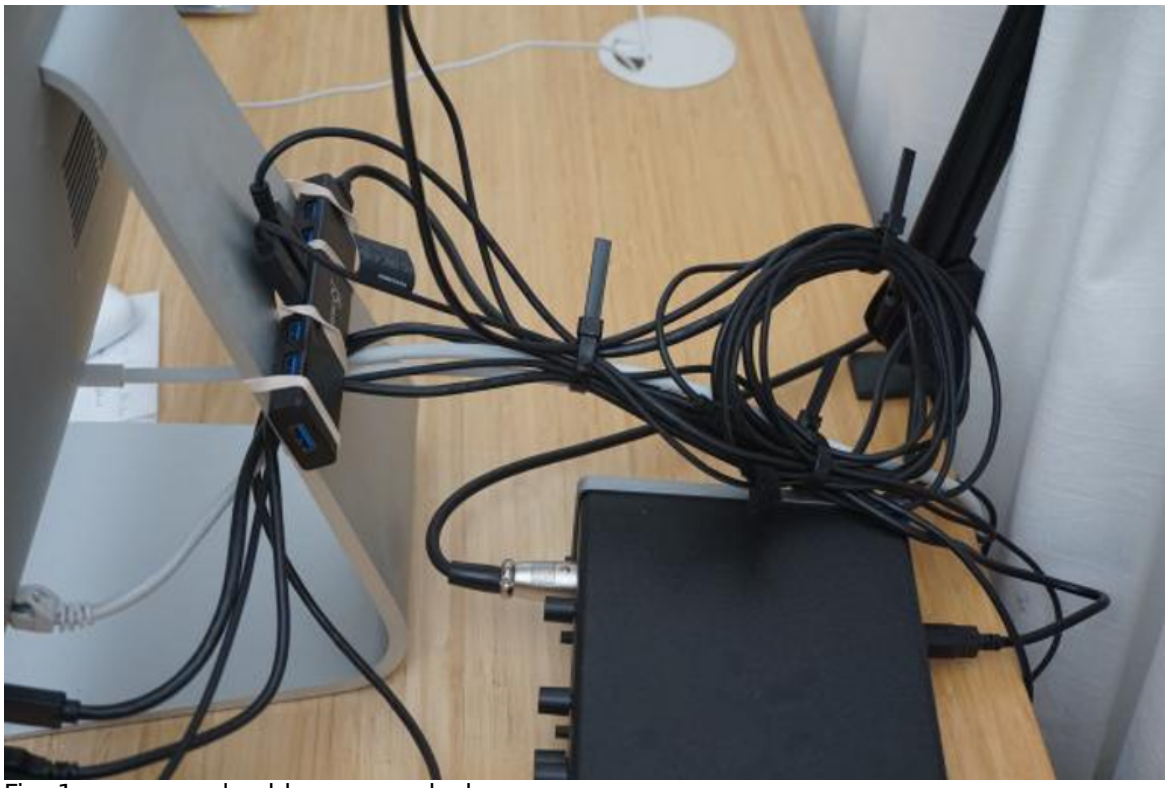


Fig. 1a: managed cables on my desk.

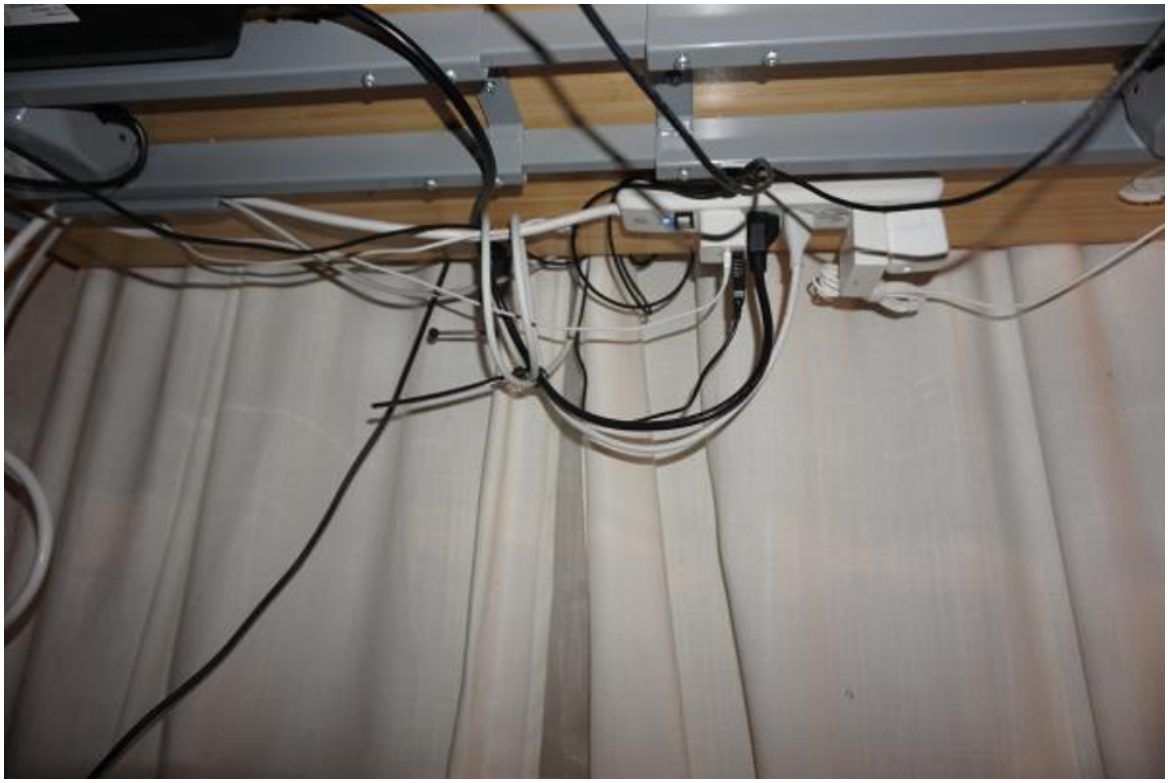


Fig. 1b: less-managed cables underneath.

Recently, I finally ragequit and replaced all my desk's wireless devices with wired ones. While I had to invest a bit in figuring out cable management (and break my habit of chewing on headphone cables), I was able to achieve nearly the same level of tidiness, with *much* better reliability, quality and speed. I no longer have to worry about my equipment failing to pair, running out of battery, or spontaneously giving me garbled robot voice during a livestreamed talk. It's dramatically reduced my level of device-related agony.

To illustrate the degree of agony I'm talking about, below I'll cover some of the subtle, hard-to-notice but severe problems I've run into with wireless protocols. If you're convinced, try out some wires—you, too, can figure out whether you've been a victim of the wireless trap.

Wifi

Interference. If multiple wireless networks are operating on the same "channel" (radio frequency band), their transmissions can interfere with each other. When that happens your device needs to re-send the same information, which makes your wifi slow.

You might think this could be solved by having routers automatically figure out the least interference-prone channel to use, but many of them seem to be quite bad at this. (Also, the old 2.4GHz wifi protocol was only allowed to use three non-overlapping channels.)

That means in dense areas (e.g. apartment buildings), routers will often pick a bad channel and end up interfering with each other. There's no way for your router or device to notify you if it's experiencing interference, so you'll only learn about it if you know how fast your router "should" be and notice that it's slower.

Dead zones. If you're too far from your router, your computer may not be able to reliably receive the signal that the router is sending, or vice-versa. How far is "too far" can also be affected, sometimes in weird ways, by whatever walls or ceilings are in the way. Unless you

know a lot about how radio waves interact with building materials, it's hard to predict where your dead zones are.

The worst part is that many dead zones aren't fully dead: your computer and the router will try to retransmit each data packet multiple times before giving up, and often it will eventually go through. If that's what mostly happens, instead of a dead zone you'll end up with a "slow zone" where your internet works, but is subtly crappy.

Of course, unless you're keeping a close eye on your network performance statistics and how they relate to your spatial location, you'd never notice a slow zone. If you noticed anything, it would be that sometimes your internet is randomly worse than other times.

Polling. Any program on your computer can ask your wireless card to enumerate the nearby networks. This causes it to go into "polling mode," where it spends less time transmitting data and more time listening for routers advertising their network info (it can't transmit and receive at the same time). Thus, it will cause a sudden burst of network delays that can e.g. cause your video call to stutter or freeze for a few seconds.

Most programmers don't realize that wireless polling interferes with network performance, so they ask the OS to poll with wild abandon. I've been burned by this many times.

The most egregious instance was when I noticed that my video calls sometimes stuttered with an oddly regular frequency. Here's the tortuous process by which I tracked down the culprit:

- I pinged my router every second for about 10 minutes, then plotted the output in Excel and confirmed that the slow pings were exactly 30 seconds apart. That made me guess it was probably a software problem.
- I asked for help debugging it on Facebook and someone recommended enabling macOS wireless debug logging.
- I enabled the debug logging and noticed that several apps, when I had them open, would requesting network scans, at times correlated to the increased ping latencies.
- After narrowing it down to those few apps, I asked how to stop them from doing that on AskDifferent.
- Someone on AskDifferent tracked the problem down to Qt, a software framework for making user interfaces, used by apps with millions of users.

Qt included a component which would poll for networks every 30 seconds whenever a "network access manager" was instantiated, causing pretty much *any Qt app* using the network to degrade your wifi for ~5 out of every 30 seconds.

There were already multiple bug reports for this issue, one of which was declared "closed" by an engineer because they allowed users to use an environment variable to disable the polling.

Of course, this is an unbelievably useless "solution" because most users won't realize that their wifi is degraded; those who do won't realize that it's Qt's fault; and those who realize will still have a hard time Googling for the appropriate fix (let alone implementing it, unless they can code).

This behavior is so user-unfriendly, and the "fix" so laughable, that it seems likely that the Qt developers somehow failed to realize the severity of the problem—I'd guess it ruined video calls for on the order of a million people, since, for instance, it affected qBittorrent which has been downloaded 75m times. Most of those million people were probably not technical enough to figure out how to "set the `QT_BEARER_POLL_TIMEOUT` environment variable to `-1`."

(Fortunately, it does look like in 2017—three years after the original bug report—they finally realized they should just *stop polling* and fixed the bug the right way.)

Qt was the worst offender, but it's far from the only one. Even macOS had a bug for a while where the same thing would happen when you opened Spotlight (which I do frequently during video calls, if someone asked me to look at a particular file, or if I want to ~~zone out and read the internet~~ multitask). I had to fix it by disabling individual Spotlight result types until I found out which one was causing the problem. So it seems even Apple's own developers don't realize that wifi polling is a hazard.

Bluetooth audio

High latency. Most Bluetooth headsets introduce around 150-300ms of latency (the time between my computer receiving the audio from the Internet, and the sound coming out of the headphones). That means that if I'm chatting with a friend in New York, the audio data will take about 50ms to get from them to my computer, and, say, 200ms—*4x as long*—to get from my computer to my ears. Since high latency ruins the natural flow of conversations, I'd like to eliminate as much of it as I can.

It's possible to find lower-latency Bluetooth headsets if they support the right "codec," like "AptX Low Latency." Of course, in addition to supporting the codec in theory, they have to agree with your computer to use it, which can sometimes fail. (The option to inspect which codec is being used is, of course, buried in various hidden menus and settings depending on your OS.)

Low quality. Related to the codec issue, many bluetooth devices will play high-quality audio when the microphone is turned off, but degrade to much lower-quality audio when it's turned on. You can test this for yourself if you have a bluetooth headset: play music on it, then open your microphone settings to the page where it shows the mic input volume. You'll probably hear the audio cut out for a second, then return at lower quality. (This happens even with devices you might expect to be high-end, like my Airpods Pro + 2018 Macbook Air.)

Bluetooth general

Dongles. Even though all computers now have built-in Bluetooth, many Bluetooth accessories today still ship with proprietary dongles. I assume this is because the manufacturer was worried about inconsistencies or incompatibilities between their own Bluetooth implementation and your computer's built-in Bluetooth hardware/drivers.

And that mistrust seems correct—for instance, on my Mac's built-in bluetooth, my Logitech MX Master displayed noticeable jank (stopping, then jumping, instead of moving smoothly). I've seen this happen on three different Macs, so it seems likely to be a software problem. When I switched to Logitech's specific dongle, though, it stopped.

Similarly, when connected to Mac bluetooth, my Jabra Evolve 75 headset would frequently have the mic or sound drop. It (mostly) worked fine on its own dongle. I'm not sure whether to blame Jabra/Logitech or Apple (or the Bluetooth standards body) for these problems.

Either way, this explosion of dongles is silly and inconvenient. At one point I had to buy a USB hub just for my four dongles (keyboard / mouse / headphones / microphone). The original intention of Bluetooth was to unify different wireless devices in a single wireless radio and protocol, much like Harald Bluetooth unified Denmark, but it seems to have mostly failed at this.

Reliability. Even with proprietary dongles, Bluetooth devices are much less reliable than wired. For instance, I wrote above that my Logitech MX Master worked fine once I switched to

the proprietary dongle, but that's not quite true: it worked fine for a while, then one day it started janking again for no discernible reason. (That was the day I finally apostasized and threw out my desk's Bluetooth gear.)

I encounter other Bluetooth bugs that require me to un-pair and re-pair a few times a week. For instance, my AirPods sometimes "desynchronize" so that one is playing back audio a few milliseconds ahead of the other, causing a strange and really unpleasant echo effect.

Interference. One possible reason for poor reliability is that Bluetooth and 2.4GHz wifi interfere with each other. Much like wifi interference, your devices will never warn you if they're experiencing interference; you'll notice only because they suddenly become kind of crappy.

Charging. Not the fault of Bluetooth per se, but a downside of using too many wireless devices is that it's really annoying to remember to keep them all charged. Mine tended to die at the least opportune times, e.g. during video calls.

Conclusion

Most of these problems shared a few things in common:

- Things didn't break completely, they just degraded. That's probably the right call, but it meant that I didn't immediately notice there was a problem.
- Compounding this, I had no idea how well the device "should" work, so I took a long time to notice that it was in a degraded state.
- Even once I was aware of the problem, it was hard or impossible to understand the root cause and fix it because I didn't know the right diagnostics (or they didn't exist).

I want my tools to be *predictable*—to have consistent performance and fail in ways that I understand. Wireless protocols are inherently more complex (because many devices share the same airspace) and have more different ways to fail, so they're much less predictable than wires. For me, the convenience often isn't worth that cost.

I still use wireless gear when it's clearly worth it—for instance, I use wifi for my laptop since it moves around a lot, and I use a wireless charger for my phone since I don't really care how fast it charges. But for serious work, I'll spend the time to fiddle around with cable routing and wire everything.

This makes me wonder what the world would look like if we took 10% of the effort we currently spend on removing wires from everything, and put it into ingenious cable routing solutions instead. I'd bet that a lot of wireless-dependent activities like video calls would be way more pleasant.

Self-sacrifice is a scarce resource

> “I just solved the trolley problem.... See, the trolley problem forces you to choose between two versions of letting other people die, but the actual solution is very simple. You sacrifice yourself.”
>
> - The Good Place

High school: Naive morality

When I was a teenager, I had a very simple, naive view of morality. I thought that the *right* thing to do was to make others happy. I also had a naive view of how this was accomplished - I spent my time digging through trash cans to sort out the recyclables, picking up litter on the streets, reading the Communist Manifesto, going to protests, you know, high school kind of stuff. I also poured my heart into my dance group which was almost entirely comprised of disadvantaged students - mainly poor, developmentally disabled, or severely depressed, though we had all sorts. They were good people, for the most part, and I liked many of them simply as friends, but I probably also had some sort of intelligentsia savior complex going on with the amount of effort I put into that group.

The moment of reckoning for my naive morality came when I started dating a depressed, traumatized, and unbelievably pedantic boy with a superiority complex even bigger than his voice was loud. I didn’t like him. I think there was a time when I thought I loved him, but I always knew I didn’t like him. He was deeply unkind, and it was like there was nothing real inside of him. But my naive morality told me that dating him was the *right* thing to do, because he liked me, and because maybe if I gave enough of myself I could fix him, and then he would be kind to others like he was to me. Needless to say this did not work. I am much worse off for the choices I made at that time, with one effect being that I have trouble distinguishing between giving too much of myself and just giving basic human decency.

And even if it were true that pouring all of my love and goodness out for a broken person could make them whole again, what good would it be? There are millions of sad people in the world, and with that method I would only be able to save a few at most (or in reality, one, because of how badly pouring kindness into a black hole burns you out). If you really want to make people’s lives better, that is, if you really care about human flourishing, you can’t give your whole self to save one person. You only have one self to give.

Effective altruism, my early days

When I first moved to the Bay, right after college, I lived with five other people in what could perhaps practically but certainly not legally be called a four-bedroom apartment. Four of the others were my age, and three of us (including me) were vegan. The previous tenants had left behind a large box of oatmeal and a gallon of cinnamon, so that was most of what I ate, though I sometimes bought a jar of peanut butter to spice things up or mooched food off of our one adult housemate. I was pretty young and pretty new to EA and I didn’t think it was morally permissible to spend money, and many of my housemates seemed to think likewise. Crazy-burnout-guy work was

basically the only thing we did - variously for CEA, CHAI, GiveWell, LessWrong, and an EA startup. My roommate would be gone when I woke up and not back from work yet when I fell asleep, and there was work happening at basically all hours. One time my roommate and I asked Habryka if he wanted to read Luke's report on consciousness with us on Friday night and he told us he would be busy; when we asked with what he said he'd be working.

One day I met some Australian guys who had been there in the *really* early days of EA, who told us about eating out of the garbage (really!) and sleeping seven to a hallway or something ridiculous like that, so that they could donate fully 100% of their earnings to global poverty. And then I felt bad about myself, because even though I was vegan, living in a tenement, half-starving myself, and working for an EA org, I *could* have been doing more.

It was a long and complex process to get from there to where I am now, but suffice it to say I now realize that being miserable and half-starving is not an ideal way to set oneself up for any kind of productive work, world-saving or otherwise.

You can't make a policy out of self-sacrifice

I want to circle back to the quote at the beginning of this post. (Don't worry, there won't be any spoilers for The Good Place). It's supposed to be a touching moment, and in some ways it is, but it's also frustrating. Whether or not self-sacrifice was correct in that situation misses the point; the problem is that *self-sacrifice cannot be the answer to the trolley problem.*

Let's say, for simplicity's sake, that me jumping in front of the trolley will stop it. So I do that, and boom, six lives saved. But if the trolley problem is a metaphor for any real-world problem, there are millions of trolleys hurtling down millions of tracks, and whether you jump in front of one of those trolleys yourself or not, millions of people are still going to die. You still need to come up with a policy-level answer for the problem, and the fact remains that the policy that will result in the fewest deaths is switching tracks to kill one person instead of five. You can't jump in front of a million trolleys.

There may be times when self-sacrifice is the best of several bad options. Like, if you're in a crashing airplane with Eliezer Yudkowsky and Scott Alexander (or substitute your morally important figures of choice) and there are only two parachutes, then sure, there's probably a good argument to be made for letting them have the parachutes. But the point I want to make is, **you can't make a policy out of self-sacrifice**. Because there's only one of you, and there's only so much of you that can be given, and it's not nearly commensurate with the amount of ill in the world.

Clarification

I am not attempting to argue that, in doing your best to do the right thing, you will never have to make decisions that are painful for you. I know many a person working on AI safety who, if the world were different, would have loved nothing more than to be a physicist. I'm glad for my work in the Bay, but I also regret not living nearer to my parents as they grow older. We all make sacrifices at the altar of opportunity cost, but that's true for everyone, whether they're trying to do the right thing or not.

The key thing is that those AI safety researchers are not making themselves miserable with their choices, and neither am I. We enjoy our work and our lives, even if there are other things we might have enjoyed that we've had to give up for various reasons. Choosing the path of least regret doesn't mean you'll have no regrets on the path you go down.

The difference, as I see it, is that the "self-sacrifices" I talked about earlier in the post made my life *strictly* worse. I would have been strictly better off if I hadn't poured kindness into someone I hated, or if I hadn't lived in a dark converted cafe with a nightmare shower and tried to subsist off of stale oatmeal with no salt.

You'll most likely have to make sacrifices if you're aiming at anything worthwhile, but be careful not to follow policies that deplete the core of yourself. You won't be very good at achieving your goals if you're burnt out, traumatized, or dead. Self-sacrifice is generally thought of as virtuous, in the colloquial sense of the word, but moralities that advocate it are unlikely to lead you where you want to go.

Self-sacrifice is a scarce resource.

Everyday Lessons from High-Dimensional Optimization

Suppose you're designing a bridge. There's a massive number of variables you can tweak: overall shape, relative positions and connectivity of components, even the dimensions and material of every beam and rivet. Even for a small footbridge, we're talking about at least thousands of variables. For a large project, millions if not billions. Every one of those is a dimension over which we could, in principle, optimize.

Suppose you have a website, and you want to increase sign-ups. There's a massive number of variables you can tweak: ad copy/photos/videos, spend distribution across ad channels, home page copy/photos/videos, button sizes and positions, page colors and styling… and every one of those is itself high-dimensional. Every word choice, every color, every position of every button, header, divider, sidebar, box, link… every one of those is a variable, adding up to thousands of dimensions over which to optimize.

Suppose you're a startup founder planning your day - just a normal workday. There's a massive number of variables you could tweak: dozens of people you could talk to, dozens of things you could talk to any of them about, and all the possible *combinations* of people and topics. There's emails, code, designs and plans you could write. Every choice is a dimension over which you could optimize.

Point being: real-world optimization problems are usually pretty high dimensional.

Unfortunately, many techniques and intuitions which work well for low-dimensional optimization do not scale up well to higher dimensions. This post talks about some of these problems, and what they look like in real life.

Try It and See

Let's start with a baseline: try something at random, see how well it works. If it doesn't work, or doesn't work better than your current best choice, then throw it out and try something else at random.

In low-dimensional problems, this isn't a bad approach. Want to decide which brand of soap to use? Try it and see. There just aren't that many possible choices, so you'll pretty quickly try all of them and settle on the best.

On the other hand, we probably don't want to design a bridge by creating a design completely at random, checking whether it works, then throwing it out and trying another design at random if it doesn't.

In general, the number of possible states/designs/configurations in a space increases exponentially with the number of dimensions. A problem in two or three dimensions, with k choices for each variable, will only have k^2 or k^3 possibilities. A problem in a hundred thousand dimensions will have k^100000 - well in excess of the number of electrons in the universe, even if there's only two choices per variable. The number of possible bridge designs is exponentially huge; selecting designs at random will not ever find the best design, or anything close to it.

Let’s look at a less obvious example.

From Studies on Slack:

> Imagine a distant planet full of eyeless animals. Evolving eyes is hard: they need to evolve Eye Part 1, then Eye Part 2, then Eye Part 3, in that order. Each of these requires a separate series of rare mutations.
>
> Here on Earth, scientists believe each of these mutations must have had its own benefits – in the land of the blind, the man with only Eye Part 1 is king. But on this hypothetical alien planet, there is no such luck. You need all three Eye Parts or they’re useless. Worse, each Eye Part is metabolically costly [...]
>
> So these animals will only evolve eyes in conditions of relatively weak evolutionary pressure.

See the mistake?

When evolutionary pressure is low, we explore the space of organism-designs more-or-less at random. Now, if we imagine mutations just stepping left or right on a one-dimensional line, with the three eye parts as milestones along that line, then the picture above makes sense: as long as evolutionary pressure is low, we’ll explore out along the line, and eventually stumble on the eye.

But the real world doesn't look like that. Evolution operates in a space with (at least) hundreds of thousands of dimensions - every codon in every gene can change, genes/chunks of genes can copy or delete, etc. The "No Eye" state doesn't have *one* outgoing arrow, it has hundreds of thousands of outgoing arrows, and "Eye Part 1" has hundreds of thousands of outgoing arrows, and so forth.

As we move further away from the starting state, the number of possible states increases exponentially. By the time we're as-far-away as Whole Eye (which involves a whole lot of mutations), the number of possible states will outnumber the atoms in the universe. If evolution is uniformly-randomly exploring that space, it will not ever stumble on the Whole Eye - no matter how weak evolutionary pressure is.

Point is: the hard part of getting from "No Eye" to "Whole Eye" is not the fitness cost in the middle, it's figuring out which direction to go in a space with hundreds of thousands of directions to choose from at every single step.

Conversely, the weak evolutionary benefits of partial-eyes matter, not because they grant a fitness gain in themselves, but because they bias evolution in the direction toward Whole Eyes. They tell us which way to go in a space with hundreds of thousands of directions.

This same reasoning applies to high-dimensional problem solving in general. Need to design a bridge? A webpage? A profitable company? A better mousetrap? The hard part isn't obtaining enough slack to build the thing; the hard part is figuring out which direction to go in a space with hundreds of thousands of possible directions to choose from at every single step.

The E-Coli Optimization Algorithm

Here’s a simple but somewhat-less-brute-force optimization algorithm:

- Pick a random direction
- Move that way
- Check whether the thing you're optimizing is improving
- If so, keep going
- If not, go some other direction

This algorithm is exactly how e-coli follow chemical gradients to find food. It's also how lots of real-world problem-solving proceeds: try something, if it helps then do more, if it doesn't help then try something else. It's babble and prune with weak heuristics for babble-generation.

This works great in low dimensions. Trying to find the source of a tasty smell or a loud noise? Walk in a random direction, if the smell/noise gets stronger then keep going, otherwise try another direction. There's only two or three dimensions along which to explore, so you'll often choose directions which point you close to the source. It works for e-coli, and it works for many other things too.

On the other hand, imagine designing a bridge by starting from a design, changing something at random, checking whether it helps, then keeping the change if it did help or throwing it out otherwise. You *might* eventually end up with a workable design, but even at best this would be a very slow way to design a bridge.

If you're a programmer, you've probably seen how well it works to write/debug a program by making random changes, keeping them if they help, and throwing them away if not.

In general, e-coli optimization can work as long as there's a path of local improvements to wherever we want to go. In principle, it's a lot like gradient descent, except slower: gradient descent always chooses the "locally-best" direction, whereas e-coli just chooses a random direction.

How much slower is e-coli optimization compared to gradient descent? What's the cost of experimenting with random directions, rather than going in the "best" direction? Well, imagine an inclined plane in n dimensions. There's exactly one "downhill" direction (the gradient). The n-1 directions perpendicular to the gradient don't go downhill at all; they're all just flat. If we take a one-unit step along each of these directions, one after another, then we'll take n steps, but only 1 step will be downhill. In other words, only ~O(1/n) of our travel-effort is useful; the rest is wasted.

In a two-dimensional space, that means ~50% of effort is wasted. In three dimensions, 70%. In a thousand-dimensional space, ~99.9% of effort is wasted.

Obviously this model is fairly simplistic, but the qualitative conclusion makes at least some sense. When an e-coli swims around looking for food in a three-dimensional puddle of water, it will randomly end up pointed in approximately the right direction reasonably often - there are only three independent directions to pick from, and one of them is right. On the other hand, if we're designing a bridge and there's one particular strut which fails under load, then we'd randomly try changing hundreds or thousands of other struts before we tried changing the one which is failing.

This points toward a potentially more efficient approach: if we're designing a bridge and we can easily identify which strut fails first, then we should change that strut. Note, however, that this requires reasoning about the structure of the system - i.e. the struts - not just treating it as a black box. That's the big upside of e-coli optimization: we can apply it to black boxes.

Beyond Black Boxes

For low-dimensional systems, try-it-and-see or e-coli optimization work reasonably well, at least in a big-O sense. But in high-dimensional problems, we need to start reasoning about the internal structure of the system in order to solve problems efficiently. We break up the high-dimensional system into a bunch of low-level components, and reason about their interactions with each other. Mathematically, this involves things like:

- Causality: if I change a line in a program, how do the effects of that change propagate to the rest?
- Constraints: the load on a particular bolt must be below X; what constraints does this imply for the load on adjacent components?
- Backpropagation: to which parameter is the system's overall performance most sensitive?

This sort of reasoning requires an expensive investment in understanding the internal gears of the system, but once we have that understanding, we can use it to achieve better big-O problem-solving performance. It allows us to identify which low-dimensional components matter most, and focus optimization effort on those. We can change the strut which is failing, or the line of code with a bug in it, rather than trying things at random. The more dimensions the system has, the larger the efficiency gain from a gearsy approach.

A Personal (Interim) COVID-19 Postmortem

I think it's important to clearly and publicly admit when we were wrong. It's even better to diagnose why, and take steps to prevent doing so again. COVID-19 is far from over, but given my early stance on a number of questions regarding COVID-19, this is my attempt at a public personal review to see where I was wrong.

I have been pushing for better forecasting and preparation for pandemics for years, but I wasn't forecasting on the various specific questions about Pandemics on most platforms until at least mid-March, and I failed in several ways.

Mea Culpa

I was late to update about a number of things, and simply wrong in some cases even on the basis of known information. The failures include initially being slow to recognize the extent of the threat, starting out dismissive about masks, being more concerned about hospital-based transmission than ended up being justified, being overconfident in the response of the US government, and in early March, over-confidently getting a key fact wrong about transmission being at least largely via aerosol droplet versus physical contact. I have a number of excuses, of course. Most other experts agreed with my views, my grandfather passed away in January, followed by his wife in early March, I was under a lot of stress, I was very busy with my personal life, I was trying to do a number of other high-priority projects, I was not paying attention to the details, and so on. But predictive accuracy doesn't care about WHY you were wrong, especially since there are always such excuses. And the impact of my poor judgement was also likely misleading to others in the community.

At the same time, I feel the perhaps egotistical need to note where I was correct early, and what I got right - followed by a clearer description of my failures. I started saying there would be PPE shortages due to COVID-19 by January, and was writing about the supply chain issues well before COVID. I submitted this paper November last year with Dave Denkenberger, which was largely finished last summer, and it was accepted in February, which then took 3 months to get published. The delay was in part due to other demands on my time, but in retrospect, if it had been available 3 months earlier, it would have been far, far more impactful.

I also understood the failure mode we ended up seeing, and in my 2018 paper, discussing overconfidence in claims that pandemics would be rare, I argued that among the most critical risks was failure to respond to emerging pandemics which could in theory be controlled quickly enough. On the other hand, my failure to realize that this is exactly what was happening is perhaps compounded by the fact that I understood the dynamics, and should have been able to identify what was going on.

Lastly, I maintain I was correct in warning about the poorly thought out and in some cases outright dangerous "preparation" in some quarters of the rationality community proposed in March, such as advocating use of bleach and ozone in closed areas for disinfection. Some people in the community were stockpiling N-95 masks and food and buying up second hand ventilators, and as I said at the time, were at best being

selfish and defecting. On the other hand, as I mention below, I was insufficiently clear about the need for better preparation, and waited far too long to speak.

Some of My Mistakes, and Related Comments

Slow to recognize the extent of the threat.

I said we should be very concerned in January, albeit not very publicly. I took until early March to start suggesting that it was clear that the US would expect to see large numbers of deaths. I was skeptical of valuable efforts early on, and didn't start really publicly sounding the alarm and reacting until even later. I was later than most of this community in recognizing the risks.

Skeptical about Border Closures

In a conversation that started Jan 27th, I was asked about shutting down borders to prevent spread. I was dismissive, in large part based on the expert consensus. I'm unsure whether this was a mistake on the object level, since I think that at that early point, the facts were unclear enough, and trade wars really are bad. I also expected response to be better, based on previous cases.

I do not think that border shutdowns were feasible, and historically they have not been. Quarantines at borders were and are logistically impossible. And full border closures for COVID-19 were also not very effective most places until very late in the spread, (Mongolia and Vietnam are the exceptions that disprove the rule.) Even late in the pandemic spread, lots of transmission occurred from places where there had been few or no cases at the time people entered. However, when discussing it, I excused my early claims that it was too economically damaging and would have been ineffective by substituting a different argument about political feasibility - one which I think is correct, but was not my original consideration. This was bad epistemic practice, and I should have been clearer that in retrospect, if they could have been put in place, travel bans would have been a much better idea. I still think my later excuse, that they were politically impossible, holds up - but I had not fully thought through the question until well after my early response.

Dismissive about masks.

The research on use of masks was unclear and I don't want to claim it was retrospectively obvious, but as a matter of decision making given uncertain risks, people should have started wearing ***homemade masks*** in public much earlier. We will still need to see how much impact promoting mask wearing in public has had, but at the very least it functioned as a clear and important public signal that COVID was serious, which promotes physical distance and other critical factors.

On the other hand, I said at the time, and still maintain that I was correct in suggesting that buying up P95 and surgical masks in February and March was defecting, since it was already clear that those supplies were needed desperately in

hospitals. And Fauci has now said as much (as a level-1+2 sage, in my view.). In retrospect, I think it would have been better, consequentially, to push for cloth masks earlier, but current modeling and our understanding of spread make it clear that mask wearing by itself is only marginally effective. I was instead focused on promoting handwashing, which I think is still undersold in importance, and thought that continued focus on masks would be a net negative. I was wrong, and others here were correct.

Not clear enough about the importance of preparation.

I've long said, following all of the experts, that people should have 2 weeks supply of food and basic supplies. Especially people in California, where earthquakes are far more common than severe pandemics. Further preparation should have been unneeded early on - but in fact, most people don't do this, and the people who were advocating making sure that you were prepared for a worse outcome were correct.

On the other hand, there is an argument I've seen here, and by others in the rationality community elsewhere, that encouraging people to buy critical supplies and hoard early in a crisis sends a price signal to get companies to produce. The argument is that this type of hoarding masks and other PPE will convince manufacturers to make more. I thought, and still think, that this is at least partly misunderstanding the way that price signals and supply chain delays propagate. Anyone who's familiar with MIT System Dynamics' Beer Game and the bullwhip effect would tell you that companies that ramped up production in response to demand quickly (rather than projections and an understanding of longer term demand) were being stupid, not prudent, and companies that tried this in exactly this area were burned in the past for doing so. If that isn't clear enough, notice that it took a couple months for the toilet paper and flour "shortages" to be worked out, despite the fact that there was sufficient supply, and there were not actual production supply shortages. Yes, markets are largely efficient, but they aren't magical ways to eliminate production and distribution delays, much less to insulate companies from actual market dynamics - and China and other southeast Asian countries had already stepped up mask production massively by mid-January. Most of the current supply comes from those factories, so the supposed benefits of price signals from buying masks in February seem not to have been actually effective in speeding anything up.

Oversold Hospital-based transmission.

Part of my concern about hoarding of masks and other equipment was that I thought we would once again see a pattern of large transmission events being centered around hospitals. Thankfully, this didn't happen - hospitals have gotten far better at isolation of patients, and they shut down non-essential services early. We did still see many, many cases and deaths in hospital staff, and this was very clearly in large part due to a lack of supply of PPE. Still, it wasn't the critical locus of spread I expected it to be.

Overconfidence in the response of (certain agencies in) the US government.

This was a huge mistake on my part. I have been concerned about the current administration for years, have repeatedly warned that it is destroying government agencies. Despite that, I was (in retrospect very unreasonably) still confident that the CDC was going to handle the situation well. They had handbooks on influenza pandemic preparedness, I had personally discussed pandemic preparedness plans with senior people at CDC just a few years ago, and I was overconfident in the ability to respond. Based on that, in turn, I was confident that the level of concern being voiced by the CDC was a reflection of their planning and ongoing preparation. The CDC has planned for preparation for this exact case for years, and I assumed they would carry out those plans. I was wrong.

It seems, though it is still somewhat unclear, that center directors were told by the director and the head of HHS that they needed not to speak out about the risks, specific recommendations were vetoed, and (easily the worst screw up,) they let the FDA ban private tests, seemingly at the direction of the administration, to hide the extent of the spread. I'm still confused by the level of non-reaction among non-political SES staff and GS-14s. We have seen many people in various agencies come forward with complaints during this administration, but CDC seems to have just dropped the ball on their response. We will likely see in the coming years how much this was due to central directives not to react, versus alack of central directives to react, therefore failing due to passivity. I still want to assume the former, but that's in large part self-justification of my prior views.

I was wrong in trying to defend the CDC's overall response in March. It definitely isn't as clear as I thought at the time that they were, and would be, net positive. I do think that the emergence of Fauci as almost a national hero has been very helpful in getting people to listen to expert recommendations, even if this did come very late. This is a point on the side of getting most people to listen more and attack less. On the other hand, Lesswrong was overall better prepared because of their skepticism, so *at the very least* I was talking to the wrong crowd to defend them, and more likely should have been quicker to judge their actions as dangerous myself.

The FDA also surprised me with how badly they did, albeit the surprise was less severe because I had lower expectations. I thought they were getting less dangerous to US public health given the previous pushes to reduce regulation by the current administration. Scott Gottlieb was there for two years, and was probably the only Trump nominee I was actually super-happy about. Unfortunately, he left (a fact I wasn't paying attention to,) and it turns out that the incompetence of a sequence of new directors and rapid changes left the FDA even less prepared that they would have been. I would have expected a doctrinaire Republican appointee to seize the opportunity of a crisis to reduce regulation, and instead it seems they did nothing but block critical testing work for months.

I've long considered myself skeptical of government agencies abilities, and lean fairly heavily libertarian in many ways - albeit less than most others at lesswrong. I was still surprised by the level of ongoing, perhaps even malicious incompetence of the current administration. I'm still unclear if this is a Hanlon-dodge, or if they really have broken the US government so badly, so quickly. Other governments managed this far less poorly, so I'm unclear how generalizable the lesson is that governments are bad at everything. But I am glad I left the US.

Being a jerk commenting on a post attacking the CDC

Given that I'm posting a retrospective, there is a different type of mistake I made that I also need to address. In a lesswrong thread several months ago, there were a number of claims made about the CDC's response. I responded that I thought the post was an infohazard, would very plausibly lead to many more people dying, and as such, the posters should have asked for feedback from someone who could vet concerns about this, and that it should be taken down by site administrators. This was stupid, and I have apologized there, along with laying out what I hope is a fair analysis of what I know I did wrong, and what I still think I was correct about.

Speculation about Causes

There are lots of things I did wrong.

First, I think I was too close to the situation. I had spent a ton of time looking at the US's system specifically, and writing about the closely related -topic of influenza pandemics in my dissertation, then doing work for Open Philanthropy on GCBRs. All of this was during the Obama administration. I left the US a bit after Trump was elected, partly for that reason, and worked on related topics that had less to do with US policy. I'd like to say that's why I didn't update, but to be honest, I think I was just being stupid in accepting my cached thoughts about the risk and best responses, instead of re-evaluating.

I also had too-strong priors and "expert" ideas to be properly fox-like in my predictions, and not quick enough to update about how things were actually going based on the data. Because I was slow to move from the base-rate, I underestimated the severity of COVID-19 for too long. I'm unsure how to fix that, since most of the time it's the right move, and paying attention to every new event is very expensive in terms of mental energy. (Suggestions welcome!)

I also gave too much weight to others' forecasts. Good Judgement's predictions were WAY optimistic about this early on, and I was not forecasting the question, but I was assuming that their aggregate guess was better than that of individuals, especially people who aren't forecasters. This is usually correct, but here it was a mistake. (I now think that superforecasting is materially worse than I hoped it would be at noticing rare events early.) I also followed the herd too much from expert circles, and my twitter feed from infectious disease epidemiology circles was behind even my slow self in recognizing that this was a incipient disaster back in March.

Conclusion

COVID-19 went badly in some places, and went disastrously in others. This was largely predictable, and I failed to notice early enough. (The US is in deep, deep trouble, and this will continue for quite a while longer, with myriad longer term effects on the global economy, and on global stability of other types.) I'm chastened about the poorly calibrated overconfidence of my expert opinion.

I'm also partly unsure what the best next steps are for better-calibration. One key thing I did, several years ago, was explicitly try to rely more on other people's views in the rationality community to guide my decisions, and provide a clear source of feedback. I didn't do this as much as I should have in this case. (On the other hand, it was a large part of why I recognized the mistake as quickly as I did, albeit later than I could have - so it was at least a partial success.)

I'm hoping that this exercise is another way in which thinking through the situation gives me a valuable chance to reflect, and that I can get further feedback. I also hope that it's useful for others to perhaps learn from, but I'm unsure how transferable the lessons of my failures are.

Five Ways To Prioritize Better

This piece is cross-posted on my blog here.

I'm going to let you in on a secret of productivity.

Those people you admire, the ones who make you wonder how on earth they accomplish so much? Those people *might* work more hours than you or be more talented or more passionate. Or they might *not*.

But they *probably* work on better things, in better ways.

Now, before you protest that that's the same thing as being talented or smart or hardworking, let me unpack that claim. Working on better things means they carefully choose what's worth caring about, and what they won't give a fuck about. Working in better ways means they carefully choose to do the most important actions to accomplish those goals.

In short, working on better things, in better ways is … prioritization. Prioritizing well is the common thread behind successful people.

Because prioritizing well is so freaking important. Cumulative good choices can multiply your impact *tens or hundreds* of times. You can't work tens or hundreds as many hours in a day. You *can* choose what to do and how to do it most effectively.

But prioritization isn't just following through on the actions you already know you should do -- though that's important too. Knowing what you should do is actually hard. There's a long gap between wanting to make the world brighter and knowing how to do it. I want to acknowledge that.

So, bear with me a bit. This post is longer than usual because I'm trying to give you a bunch of examples to really get a taste of what prioritizing feels like. I think every one of the examples will help you understand better, but, if you have a hard time focusing for 14 pages, maybe read one tool now and come back another time for others?

I've organized the examples around five specific concepts that you can use to help apply the mindset more effectively: Theory of Change, Lean Tests, Bottlenecks, Ballpark Estimates, and the 80/20 Rule. I'll start each section with some concrete stories before summarizing the concept and how it helps you prioritize.

Note: The following vignettes, apart from the ones about me and the interview with Tara, are fictional. However, they are inspired by multiple real conversations I've had. If you find yourself thinking "this could never happen", rest assured that, though these examples are fictional, they are based on real conversations.

I. Theory of Change

1. Career planning

Phil is telling his friend Erica about his plans to switch into AI safety.

Phil: “So, I was thinking I’d stay at my job for another six months. I have a bonus coming then, so it’s a better time to make a career switch. ”

Erica: “Huh, I’m surprised that your current job is the best plan for your goals. I thought you said the most important goal was building your machine learning skills?”

Phil: “That’s right, but I’m picking up some machine learning on the job, so it’s also skill building!”

Erica: “But you’re only spending a bit of your time on ML, and it’s not like all of that learning will transfer to the jobs you’re applying for anyway. Wouldn’t you learn a lot more by independently studying some directly relevant ML?”

Phil: “Hmm…you’ve got a point. I could learn a lot more if I took a month off to just study. But that feels riskier, and I might have a hard time staying motivated.”

Erica: “But also consider, you told me that you thought that working on AI safety was several times as impactful as your current role, at least in expectation. That means that spending six more months before switching is like wasting a whole year or more of impact.”

Phil: “Man, you’re right. I could start applying now so I’m ready to switch right after I get my bonus. But, I don’t actually know if I’m ready! I don’t know if I have the skills I would need to get hired.”

Erica: “Then it sounds like you need to find out.”

Phil: “I know a couple of people I could ask, and I can look at the job posts to see what skills they’re looking for. That should help. And if I’m not ready, I’ll probably get a better idea exactly what I should study. Then I can make a timeline so I’m ready to switch as soon as I can.”

Erica: “How do you feel about that plan?”

Phil: “Pretty good. I think leaving my job feels really risky given how uncertain I am, and it was making me avoid thinking about switching. Investigating more without feeling like I need to commit to leaving helps.”

When Phil worked backwards from his end goal, he realized his original plan was actually a pretty bad plan (at least for him), because it wasn’t going directly for the goal. Despite a lot of uncertainty, Phil’s best guess is that applying sooner is much better. Waiting would probably mean he’d lose a few months that could have been spent on much more valuable work. But at worst, he might never make a plan that will actually have an impact unless he stops and thinks harder -- it’s really easy to default to the status quo, even when that’s a bad decision.

2. Publishing an op-ed

Elle wants to publish an op-ed. But she doesn’t really know how op-eds get published. So, she asks her friend Peter - who has published op-eds before - about the process.

Elle: “So, I was thinking I’d write an article, send it to a bunch of newspapers, and cross my fingers. What do you think?”

Peter: “Well, that’s probably not going to work out. You don’t just write an op-ed and send it to places; you need to really tailor the piece to the magazine you want to publish it in.”

Elle: “Huh, how would I do that?”

Peter: “First, you should check out other pieces on the platform you’re interested in. If a venue has already published several pieces on similar topics, then it’s likely that an editor there likes that topic and is more likely to accept a new perspective on the issue. Second…”

Elle lacked an accurate model of how her actions would lead to her goal because she didn’t understand the world well enough. Creating an accurate theory of change required learning what actions would generally suffice to accomplish her goal. If she hadn’t learned how the process actually worked, she might have wasted a lot of effort without ever getting the piece published. But she had no way of knowing that until she asked.

3. Studying for the GRE

Like many students before me, I didn’t really care if I remembered the content after the test. I just wanted to spend as little time as possible to get a GRE score I was happy with. But how to do that?

According to test prep websites, I should have worked through a GRE prep book (preferably theirs). After all, those are explicitly designed to prepare you for the GRE. They include everything I needed to know; concept review, vocabulary quizzes, and practice tests.

But I knew they would actually waste time.

See, I knew that spreading my time evenly over the material was an inefficient way to learn. I’ll learn a lot more studying stuff I don’t know yet, rather than reviewing stuff I already knew. Neither the study guide nor the vocab would have focused me on just my weakest areas, so I would have spent hours rereading familiar material.

Instead, I optimized my study for rapidly improving my weaknesses: take practice quant section tests (my weakest section), study the questions I missed, summarize the concepts behind the questions, repeat. Those three steps saved me dozens of hours.

Theory of Change

In each of these examples, the person wanted to accomplish a particular goal. They worked backwards from that goal to find the steps that would make them likely to succeed at the goal.

Elle figured out how the publishing world worked so that she knew what actions were likely to succeed. Phil found that a different path would allow him to have an impact sooner and reduce his chances of failing. I worked backwards to cut out unnecessary work and save time.

Each person needed to know their goal, figure out what steps would reliably lead to that goal, and then focus only on those steps. That's the theory of change -- this model of the causal chain of actions that lead to successfully accomplishing the end goal. You might need to investigate how the world works or go learn something if you don't have enough information, in order to accurately pick what will have an impact.

So, if you want to apply this tool, ask yourself -- what steps will *actually* make you likely to achieve your goals?

(If you want more, here's a great post on theory of change.)

Working out your goals and your theory of change is a prerequisite step to other prioritization techniques. For example, it's going to be hard to use the next concept (lean tests) without having a least a starting point for what you want to optimize.

II. Lean Tests

1. Starting a business

The summer after my freshman year of college, I was hired to start a company making personal biographies.

I immediately started finding contractors to create the books, getting prototypes made, and building a website. This seemed reasonable to me then – after all, that was my job description. And I did a good job. At the end of the summer, I had a full production plan ready for when the first customer purchased.

Except, no one ever purchased a single book.

If I had started by talking with potential customers, I could have known in a month that the idea was doomed from the start. The target audience had no interest in the elaborate $10,000 product my employer envisioned. But I didn't know that, because I created a product before I confirmed people were ready to buy it. I could have saved an entire summer of work if I had just tried talking to potential customers first.

2. Conducting research

Max is asking his friend Ellen for advice about his research project.

Max: "Aw man, I'm super bummed -- I spent *six months* researching this policy area. Then when I sent my write-up draft for feedback, someone sent me an unpublished document where they'd already done part of the research! Plus, now I need to do *more* research to answer their questions. I feel like I wasted so much time already, and I'm not even done. What could I have done?"

Ellen: "That sucks. Hmm, there's a post called Research as a Stochastic Decision Process that might help avoid similar situations in the future. Want to hear about it?"

Max: "Yeah!"

Ellen: “The really simple version is to first do the parts of the task that are most likely to fail or change what other steps you will do, rather than doing the easiest parts first. This way, you reduce uncertainty about which tasks are necessary as quickly as possible. For example, if your task has three steps and one is most likely to fail, you save time in expectation by doing that one first, because you might not need to do the other steps at all.”

Max: “Yeah... I did the easy tasks first here. Like, doing all the research was a lot of work, but it was easy to just keep reading. Asking for feedback didn’t take much time, but it was *hard*. I got really anxious whenever I thought about sending the emails, and just kept doing more research, and then more research. But having that feedback earlier would have totally changed what things I choose to research. It could have saved me hundreds of hours.”

3. Learning a new subject

Alex wanted to do some independent study to see if he would be a good fit for working on AI safety. Based on 80,000 Hours recommendations, he found a few promising math courses that he could work through on the weekends over a few months.

Before he got started, he asked a few acquaintances who worked on AI safety whether his plan made sense. They mostly thought it did, but agreed that several of the math topics he had planned to study weren’t immediately valuable. He could safely skip those for now.

Spending an hour sending those emails cut his months of study in half.

Lean Tests

In each of these examples, the person wanted to efficiently accomplish their goal. In the first two examples, the person failed to test quickly, and wasted time or failed entirely. In the third, by quickly testing their idea, the person exposed flaws early so that they could rapidly correct them or move on - increasing the proportion of their effort that actually made them succeed.

I spent an entire summer on a project that failed entirely because I didn’t test it early enough to change it into something that could succeed. Max could have saved hundreds of hours by asking for feedback upfront. Alex *did* save hundreds of hours by checking his plan before he implemented it.

Each person could have broken their tasks into chunks to iterate on, getting feedback each step of the way, rather than risk wasting time by investing a lot of effort without feedback.

Lean methodology is about continuously doing small tests to check that you’re on the right track. This allows you to iteratively making lots of corrections that move you towards your goal, even when you’re not sure what is required (or when you are sure but are wrong.)

By finding the flaws early, you can change course early, minimize wasted time, and reduce the risk of ultimate failure. Better to know a project will fail before you put in

months of effort that could have been spent on a project more likely to succeed. You could get feedback from more experienced people, your target audience, or the thing itself.

So, what is the first quick test you could create to get feedback and iterate?

III. Bottlenecks

1. Anxiety

Anna is early in her journalism career. She's talking to her partner Dan, a fellow writer, about her draft of a piece on factory farming.

Dan: "Hey, did you pitch your idea to your supervisor today like you planned?"

Anna: "No… I just got too nervous and couldn't make the words come out. I think I want to keep working on it before I talk to her."

Dan: "Anna, your piece is great! You've been working on it for a *year* already. It's not going to get better without feedback. All of your hard work won't matter until someone sees it."

Anna: "Yeah, you're right. But I just feel so scared when I try. My chest gets tight and it feels hard to breathe."

Dan: "Anna, maybe it's time to consider talking to someone about this. What do you think?"

Anna: "I've actually been thinking about it for a while now, and I think you're right."

2. Procrastination

Mary had meant to get started on her final presentation for her internship two weeks ago. But she felt a wave of dread whenever she started to think about it, so her thoughts slid away to something less awful each time. Now the presentation is in a week, and the feeling has risen to panic mode. Yet she still can't make herself even look at her research.

This isn't a new feeling for Mary. She's six months overdue on a write-up from her former research position. But whenever her former adviser sends an email asking for the report, a pit opens in Mary's stomach.

Mary started coaching to tackle her persistent procrastination. Over the next year, Mary and I worked together to build up her ability to make better plans, set up habits to reduce distractions, learn to ask for help early and often, and find commitment mechanisms that work for her.

After a year of working on this big bottleneck, she feels confident in doing her work by her deadlines. That work was an investment in herself that will pay off over the rest of her career.

3. Fatigue

Sometimes one issue will dominate everything else in regard to productivity, often a physical or mental health issue. In my case, it was fatigue. When you need to take three naps a day, you get less work done regardless of what other productivity tools you use.

So, it was worthwhile investing a bunch of time and effort to improve this. I tried a parade of experiments and tracked all the factors that I thought might influence my energy levels – including sleep times, sleep duration, hydration, exercise, medications, melatonin, doing a sleep study, temperature, naps, and nutrition. I tracked my energy levels and a changing subset of variables for 3-6 months, then compared the odds ratios for each variable.

Here, a bunch of small things cumulatively broke the bottleneck, primarily sleep duration (which was fixed by a consistent sleep schedule), exercise, hydration, and finally an antidepressant.

While I still have issues with fatigue, it's no longer the key thing holding me back.

Bottlenecks

In each of these examples, the person needed to get past one bottleneck that was holding them back from succeeding. Each bottleneck eclipsed other tasks, even valuable tasks, until the problem was addressed.

Anna needed to work on her social anxiety before her hard work would ever see the light of day. Mary found she could accomplish several times as much once she got her procrastination under control. I had the energy to start a blog after I reduced my fatigue.

Originally, bottlenecks referred to the "rate-limiting factors" that slow down the entire production line. In our examples, the bottlenecks are the tasks, beliefs, or problems that slow down or stop you from accomplishing other tasks. Each person made progress by identifying what was holding up other steps. Once they took care of the high-leverage factor, everything else went much faster.

So, what are the one or two things you could change about yourself or your environment to accomplish twice as much?

IV. Ballpark Estimates

1. Choosing projects

Will is talking with his PhD supervisor, Kate, about feeling overwhelmed by too many projects.

Will: "I think I just need to choose one or two to focus on, and put the rest on hold for now. But I *want* to work on all seven."

Kate: “Well, to start with, which ones are most important to work on?”

Will: “That’s the problem; they all seem important! Papers 1 and 2 have a good chance of being published in a good journal, which is important if I want to continue in academia. Paper 3 has a good collaborator, and I don’t want to let them down. And papers 4 and 5 are exciting. I think those ideas could really be impactful. Arrg...I just feel overwhelmed when I think about it, like I need to do them all.”

Kate: “Okay, let’s try a thought experiment. How much would you pay to have each of these projects magically completed? If it’s hard to think about paying with your own money, how much do you think Open Phil would pay to have the project completed?”

Will: “Argh, thats hard.” *15 minutes of brainstorming later* “Okay, paper 4 could be really big if it goes well, and 2 and 3 are *maybe* most important for my career. So I think I’d pay like $1000 each for papers 1 and 5, $1500 for papers 2 and 3, and $5000 for paper 4. But these are super crude, I’m really just guessing here.”

Kate: “Crude numbers are fine. You’re really just trying to get a better sense of how you intuitively value each of these. Those numbers help clarify the ranking and rough magnitude of difference between the projects. Sounds like paper 4 is the best, and then 2 and 3 are a bit better than 1 and 5, all else equal. Now, which projects do you expect to take the least time?”

Will: “Paper 3 for sure. My collaborator is doing a bunch of the work, so it’s probably half as big as the other papers. Between the others...really hard to say. Um, so maybe I’d sort them 3, 1, 4, 5, 2 from least to most time, but I’m really guessing here.”

Kate: “Based on value and time required, it sounds like you should spend most of your time on paper 4, plus some on paper 3.”

Will: “But all of these are estimations! I’m not confident, and I could be really wrong.”

Kate: “You’re not going to be confident. Things are uncertain and will be uncertain no matter how long you think about it. So you *have* to make your best choice despite your uncertainty. Estimates are a way to try to make that choice as well as you can. And you should absolutely spend more than 5 minutes thinking about them. So, take a few days to think about your estimates in more depth. Maybe ask a couple of advisors. But, when you’re done, go with the highest expected value and stop worrying about it. You can change your plan if you get new information. For now, you’re doing the best you can, and that’s good enough.”

Will: “I...think...I can do that. Thanks, Kate.”

2. Learning a new skill

Lyra is talking with her coworker Mike about her plans for independent study to improve her coding skills.

Lyra: “So, I’m debating between spending a bigger chunk of time to really understand computer architecture or doing several small learning projects around things that came up in my job. I’m having a hard time making progress on either idea because I keep flipping back and forth about which seems most important.”

Mike: “Can you try calculating the time required for the learning, and the time saved afterward, to calculate which is better?”

Lyra: “So I tried doing that earlier. I estimated the architecture learning would take me fifty to a hundred hours to do, and save maybe twenty or thirty hours a year. On the other hand, one of the smaller projects would only take five or ten hours, and it would save me a few minutes a day, which adds up to ten or fifteen hours a year.”

Mike: “That sounds like the smaller project is clearly a better deal - it would pay for itself within one year, while the bigger project would need more like three years to break even.”

Lyra: “I know, but the bigger project feels important anyway. I think... the bigger project isn't just about saving time. I care about it because it also opens up the option to do new things that I can’t do right now. But I’m not sure if that benefit is big enough to make it worth doing.”

Mike: “Could you run an experiment for five hours or so to see if it seems like you’re able to do new things?”

Lyra: “Yeah, that sounds good. If it looks like I’m not, then I can go ahead with the smaller project since that’s better for saving time.”

Even though Lyra didn’t follow her numbers exactly, they were helpful for clarifying the decision.

3. Job hunting

Mark has just graduated from university, and he wants to complete an ML masters to be an engineer at an AI safety org or earn to give if that doesn’t work out.

However, the programs cost between $15,000 and $24,000, and Mark isn’t comfortable going into debt. So his plan is to get a job now, apply for the masters’ programs in the fall, and save up money until he starts the following year.

Mark’s previous summer job has offered him a full-time role for $15 an hour. However, his undergrad thesis supervisor encourages him to look into software engineering jobs, which the supervisor thinks he’s qualified for. Unfortunately, he doesn’t have enough time to take the summer job while also applying.

Mark isn’t sure about it, but his supervisor convinces him to look into some job postings for local positions. So Mark puts together the following estimates.

	Salary per month working	Time before he can start work	Money saved by the end of the year
Summer Job	$2,400	0 months	$28,800
Software Engineer	~$4,000-$6,500	2-6 months	$24,000-$65,000

Based on those numbers, he decides it’s worth delaying starting at his previous job to spend a couple of months applying to engineering roles. If it works out, he’ll earn a lot more. If the job hunt doesn’t seem promising after a few months, he can go back to the other role.

Ballpark Estimates

In each of these examples, putting numbers on the uncertainty helped the person prioritize. Estimating didn't cleanly decide their priorities, but it reduced uncertainty. It revealed blind spots, such as missing important considerations or thinking two things were comparable when one was actually way more important. By quantifying, ranking, or crunching numbers, they were able to make a better guess at what should be prioritized.

Quantifying his expected value and time required helped Will increase his expected value per hour of work 2x compared to working on one of the papers at random. Quantifying return on investment for her time learning helped Lyra identify what factors she was overlooking, so now she can evaluate whether the project is worthwhile. Mark decided to apply to more ambitious jobs based on his numbers, and got a role making >50% more after three months of applying.

Classically, Fermi estimates are back-of-the-envelope calculations intended to quickly approximate the correct answer, usually when the real answer is difficult to get. Here, people estimated values and costs of different options, so they could approximate the return on their investment and compare opportunity costs.

Since we're frequently prioritizing amid uncertainty, even moderately reducing that uncertainty improves decisions. Often you have some data easily available even when you feel uncertain. Sometimes this is just making your intuition concrete. Sometimes it is actually gathering data and crunching numbers. You should take care not to be overconfident in your estimates, but even totally made up numbers can sometimes be useful, such as when you want to make a decision between competing intuitions.

So, what returns do you get on the time and effort invested? How does this compare with your other options?

V. 80/20 Rule

1. Working hours experiment

Bill was frustrated by a consistent dip in energy each afternoon. He felt less motivated during this time, and sluggish and slow even when he forced himself to work.

Working out in the afternoon helped him feel more energetic afterward, but taking a forty-five-minute break made him feel like he needed to work late.

He knew that subjective experience doesn't always match actual output, so he tried quickly recording how many words per hour he wrote for a week. Bill also noted how he subjectively felt during that time. At the end, the data suggested that his output dropped by nearly half during that period, and only gradually picked back up over the later afternoon.

He kept recording the data while he took some workout breaks. Although the data was noisy, he found that he got about as much done on days when he worked out and days when he didn't.

Given that, he decided to work out each afternoon without feeling obligated to work late.

2. Optimizing school

Sarah is a college freshman asking Alice, a senior who works in the same lab, for tips on how to succeed in college.

Sarah: “So, what are the most important things you do to get good grades?”

Alice: “Umm, I plan each day the night before so I know exactly what I need to do, and then I set aside a couple of hours when I turn off my phone and study without any distractions. That’s big. I usually do the most important task first so that I don’t risk running out of time. Oh, and I have a question in mind while I’m doing research, so that I don’t lose too much time going down rabbit holes.”

Sarah: “Is there something else that helps you reliably manage your workload?”

Alice: “So, I start my planning by looking at which projects are worth a lot of a grade or that I care a lot about learning, and I choose which projects deserve the most time and which to just do the bare minimum. For example, going from an okay paper to a great one takes a lot more work, so I’ll only do that if I really care about the paper. Otherwise, I’ll wait to start the paper until the day before it’s due, and then race like crazy. It forces me to get the paper done without spending too much time on it.”

Sarah: “Thanks, Alice. That sounds helpful. I’ve been feeling really overwhelmed by taking three really hard classes, and I really want to get As in all of them. I’m a bit of a perfectionist, I know.”

Alice: “It’s great that you’re asking for advice, Sarah, that will probably help you figure college out way faster than I did. But I want to ask, why do you care about getting good grades?”

Sarah: “Wha-what?”

Alice: “Why are As the thing you’re aiming for right now?”

Sarah: “That doesn’t even make sense. Grades are just what we’re supposed to do here. Wait, I’ll try to work out the reason… Because in school, grades are how we know we’re doing well or where we need to improve. And how future employers or grad schools know that we’re good.”

Alice: “That’s fair, but it’s only a small part of what people will care about in the future. You’re only a freshman, but you’re already working in this lab and your research seems really promising. You obviously love doing it. But you’re only doing five hours a week here because you say you don’t have enough time to do this and study. If you instead spent a lot more time on research and did really cool things there, I bet both grad schools and future employers would care about that more than a 4 point GPA.”

Sarah: “Hmm.”

Alice: “I mentioned earlier how important it is to decide which parts of a project deserve more time, and which to just put in the minimum. Well, it’s even more

important to carefully choose what projects or classes are worth a lot of time, and when it would be better to do the minimum you can in some classes, so that you can invest heavily in others."

Sarah: "I need to think about this. What you're saying kind of makes sense, but I'm worried that if I'm doing the bare minimum, my grades will drop too much."

Alice: "Good things to consider. I'm applying to med school, so my grades matter. But I chose to take easy classes for my gen eds and electives, so that I can put in a lot of time here at the lab without damaging my GPA. That's the right decision for me. I'll bet it's worthwhile for you to spend some time thinking about what you want to be perfectionist about."

3. High-value rest time

In my interview with Tara Mac Auley, she advised trying a bunch of leisure activities to decide which are most valuable for you.

> "If you take time to rest and you come back, and it doesn't feel better then probably the ways that you've chosen to rest aren't in fact the most restorative things you could be doing. And so I would suggest trying a lot of different things: a lot of different types of social activities, or physical activities, or intellectually engaging activities.
> I did this a lot when I was in my early 20s. I picked a random event from meetup.com every day for about two months, and I just had to go to whichever thing came up. And then I would write down beforehand whether I thought I would enjoy it and feel drained or refreshed from that activity. And then I would compare afterward what I actually felt and, I don't know, that was really informative and good for me."

She used this type of process to identify the activities that best leave her rejuvenated and rested.

> "Being near water and swimming, but not in a swimming pool, it has to be natural water. Being in nature. Reading a book, especially reading a book in a park or by a lake or something like that. Spending quality time with close friends or family just having a conversation for an hour with no particular goals, I find really rejuvenating. And eating a really nice meal; one where I can kind of savor all of the different tastes and textures.…I go out dancing a lot on my own, to go and see music artists that I enjoy, and I just dance like a crazy person until I'm really tired and then I go home, and that's amazing."

80/20 Rule

In each of these examples, the person wanted to prioritize the most valuable subset of possible actions. By identifying the higher-value actions, they could get more done for their effort.

Bill did an experiment to find out which hours of work provided the most value, so he could make better decisions about when to work. Sarah prioritized the highest value work to get good grades, and started thinking about how much more valuable that effort would be if she prioritized the highest value goals to begin with. Tara

experimented to find out which activities were the highest value fun for her, which she can now exploit 80/20 style.

Based on the idea that 80% of an output comes from 20% of the input, the 80/20 rule suggests that the value per unit of effort varies a lot across different actions. Because outputs vary so much, explore more can unearth dramatically better options. So, similar to Tara, you need to try many actions first in order to effectively identify the top-performing subset. Once you've identified which actions are most valuable, you can narrow your focus to just that subset. Then your output will increase significantly for the same effort.

So, what gives you the most value for the least effort? What can you cut with minimum loss so that you have extra resources to put toward what's most valuable?

Conclusion

All of these stories are of how people tried to identify the most valuable actions they could take to accomplish their goals. They reduced uncertainty, said no to other actions, and made choices based on their best guesses.

You might be tempted to say these examples don't feel important. That choosing which skill to learn or overcoming anxiety can't change the world. And maybe you're right, if you only look at that one step.

If you put all of these together, you have a mindset that searches for the most valuable goals, builds models to effectively accomplish them, iteratively tests assumptions against the world, logically weighs the opportunity cost, and judiciously spends time and effort to get the most impact possible. That mindset touches all your decisions.

And that's prioritization.

Because prioritization isn't something you do once a month. It's not a magical ability that lets you do everything - quite the opposite, in fact. It's the gut-deep sense that your time and effort are limited and you need to choose what to do, because you can't do everything.

But when you do that? When you put all of your reason and tools to the task of choosing the most valuable goals?

Then we have a chance. Choose important goals, and you could save lives from dying of malaria or build a future where pandemics don't wipe out hundreds of thousands. Accomplish those goals, and *the world becomes better*. If you need to take care of your own mental health or build skills first, then do it. You're still nudging the world in the right direction.

And if you don't? ...Then we're still right where we are now. We've lost out on some of the goodness and wonder the world could have had. There's the sense of being so close and just missing what could have been.

That's why I want to convey the mindset of what it feels like to prioritize.

So, here are five questions to take with you. Use them to make the world better.

1. What steps will *actually* make you likely to achieve your goals?
2. What is the first quick test you could create to get feedback and iterate?
3. What are the one or two things you could change about yourself or your environment to accomplish twice as much?
4. What returns do you get on the time and effort invested? How does this compare with your other options?
5. What gives you the most value for the least effort? What can you cut with minimum loss so that you have extra resources to put toward what's most valuable?

Enjoyed the piece? Subscribe to EA Coaching's newsletter to get more posts delivered to you.

Our take on CHAI's research agenda in under 1500 words

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This work was supported by OAK, a monastic community in the Berkeley hills. It could not have been written without the daily love of living in this beautiful community.

Last week I attended the annual workshop of Stuart Russell's research lab at UC Berkeley — the Center for Human-Compatible AI (CHAI). There were talks by Russell himself, as well as several graduates of the lab who now have research positions of their own at other universities. I got the clearest picture that I've yet encountered of CHAI's overall technical research agenda. This is my take on it.

Assistance games

Traditionally, AI researchers have formulated problems assuming that there will be a fixed objective provided by a human, and that the job of the AI system is to find a solution that satisfies the human's objective. In the language of sequence diagrams this looks as follows:

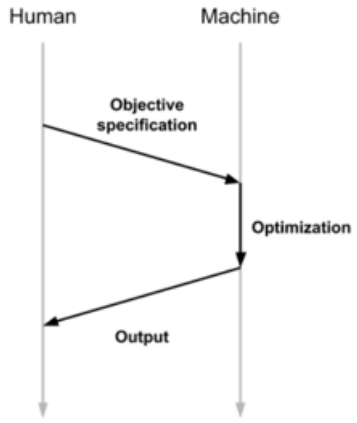


Figure 1: The "standard model" of AI research

For example, in a search problem the objective specification might be a graph over which the system is to search, a cost for each edge, and a goal state that terminates the search. The AI researcher then needs to to develop optimization algorithms that efficiently find a minimum-cost path to a goal state. Or in a supervised learning problem the objective specification might consist of a dataset of labelled examples and the AI researcher needs to develop optimization algorithms that efficiently find function approximations that extrapolate these labelled examples to future unlabelled examples.

CHAI's basic insight is to ask: why limit ourselves to a one-time objective specification event? We know that it is difficult to capture everything we care about in a formal metric (c.f. Goodhart's law). We know that humans aren't very good at foreseeing the

strange and sometimes deranged ways that powerful optimization can give you what you asked for but not what you wanted. Why should information about the human's objective be transmitted to the machine via a one-time data dump, after which it remains fixed for all time?

There are many alternative interaction patterns by which information about the human's objective could be transmitted to the machine. The human could observe the machine and provide it with feedback as it works. The machine could ask the human questions about its objective. The machine could observe the human and deduce its objective from its behavior. And so on.

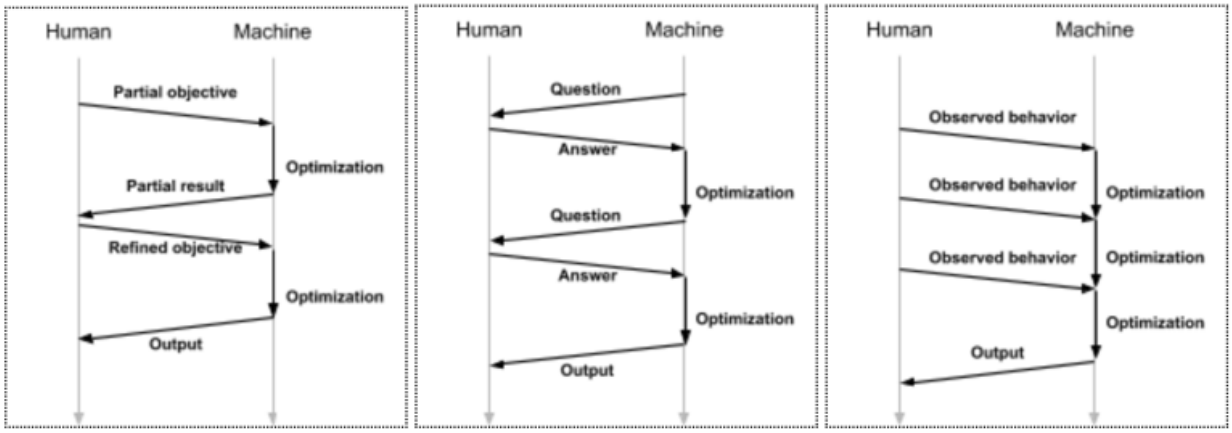


Figure 2: Examples of interaction patterns in assistance games

CHAI calls this an assistance game: the human wants something from the machine, and it is the machine's job to both (1) figure out what that is, and (2) fulfil it. The role of the AI researcher under this new model then is to explore the space of possible interaction patterns and find one that is conducive to the machine building an informed picture of what the human wants as quickly as possible.

The old model in which a complete objective is specified up front is actually just one special case of an assistance game: one in which the interaction pattern is that the machine receives all the information it will ever receive about the human's objective in a one-time up-front data dump. The unique thing about the old model -- and the reason it is both attractive and dangerous -- is that the machine never needs to entertain any uncertainty about what the human wants. It is given an objective up front and its job is just to fulfil it.

Using more nuanced interaction patterns require the machine to maintain uncertainty about what the human's objective is, which in turn requires optimization algorithms formulated so as to take into account this uncertainty. This suggests an exciting reformulation of each of the basic AI problem statements, and CHAI seems to be enthusiastically taking up this challenge, including with Russell's new edition of the standard AI textbook AI: A Modern Approach.

One of CHAI's early successes was the development of cooperative inverse reinforcement learning (CIRL). But CIRL is often mistaken as representing the entirety of CHAI's research agenda, whereas in fact CIRL is one particular approach to solving one particular kind of assistance game. Specifically it addresses an assistance game in which the machine observes demonstrations by a human, who is in turn incentivized to provide demonstrations that are of value to the machine. The original CIRL paper makes various further modelling assumptions and approximations in order to arrive at

a concrete algorithm. CIRL is an important contribution to the field but it is important to understand that the program of reformulating AI as an assistance game is broader than this one specific proposal.

I found myself wondering whether reinforcement learning might also count as a new-style assistance game. In reinforcement learning the machine begins by exploring, and the human provides a positive or negative reward each time it does something consistent with or opposed to the human's objective. In early reinforcement learning literature it was envisaged that there would be a literal human providing live feedback as learning progressed, but in modern reinforcement learning the reward signal is generally automated using a program that the human provides before learning commences. In this way modern RL looks more like the old model of Figure 1 since the reward algorithm is generally specified up-front and not modified during training. Also, reinforcement learning agents do not really maintain uncertainty about what the human's objective is: they just act to maximize their reward, and for that reason will take control of the reward signal if given the opportunity.

Going beyond agents

This new model of CHAI's relaxes one of the core assumptions of classical AI research -- that a fully-specified objective will be given at the beginning of time -- but there is still a strong assumption that both the human and the machine are well-modelled as having objectives.

Suppose we wish to build a machine that provides assistance to a rainforest. Can we view a rainforest -- the complete ecosystem containing all the various plants and animals that live there -- as having objectives? The plants and animals living in a rainforest spend a great deal of energy competing against one another, so it is difficult to view the rainforest as behaving according to any unified set of objectives. Yet it is possible to take actions that do great damage to a rainforest (clearing a large area of all trees, or introducing a synthetic pathogen) and conversely it is possible to take actions that protect the rainforest (preventing the clearing of trees or preventing the introduction of synthetic pathogen). Should our AI systems be able to observe a rainforest and deduce what it would mean to be of assistance to it?

Humans, too, are not perfectly modelled as agents. The agent model provides a compact and useful description of human behavior at a certain resolution, but as we look deeper into our own nature we find that we are not such unitary agents at all. We are in fact made of the same basic building blocks as the world around us and our view of ourselves as agents is only an approximate description. Should our AI systems model humans as agents, or can we do better? What is the next most detailed model up from the agent model? Can we build AI systems that play assistance games on the basis of this model?

Finally, machines are not ideal agents either. Any AI algorithm deployed on real hardware in the real world is made of the same basic building blocks as the world itself, and is only approximately modelled as an agent with a set of objectives. As we dial up the power of the AI systems we build and those AI systems are able to build increasingly detailed models of the world around them, this agent approximation is likely to break down. This is because an AI system that has a sufficiently detailed model of its environment will eventually discover and begin to reason about its own computing infrastructure, at which point it will need some way to deal with the paradoxes of counterfactual reasoning and logical uncertainty that arise when one can

accurately predict one’s own future behavior. If we have constructed our AI systems on the basis that they are well-modelled as agents, with no line of retreat from this assumption, then when our AI systems build detailed models of the world that conflict with this assumption, they are likely to misbehave.

A revolution in philosophy: the rise of conceptual engineering

Almost a decade ago, Luke Muehlhauser ran a series "Rationality and Philosophy" on LessWrong 1.0. It gives a good introductory account, but recently, still dissatisfied with the treatment of the two groups' relationship, I've started a larger "Meta-Sequence" project, so to speak, treating the subject in depth.

As part of that larger project, I want to introduce a frame that, to my knowledge, hasn't yet been discussed to any meaningful extent on this board: *conceptual engineering*, and its role as a solution to the problems of "counterexample philosophy" and "conceptual analysis"—the mistaken if implicit belief that concepts have "necessary and sufficient" conditions—in other words, Platonic *essences.* As Yudkowsky has argued extensively in "Human's Guide to Words," this is *not* how concepts work. But he's far from alone in advancing this argument, which has in recent decades become a rallying cry for a meaningful corner of philosophy.

I'll begin with a history of concepts and conceptual analysis, which I hope will present a productively new frame, for many here, through which to view the history of philosophy. (Why it was, indeed, a "diseased discipline"—and how it's healing itself.) Then I'll walk through a recent talk by Dave Chalmers (paper if you prefer reading) on conceptual engineering, using it as a pretense for exploring a cluster of pertinent ideas. Let me suggest an alternative title for Dave's talk in advance: "How to reintroduce all the bad habits we were trying to purge in the first place." As you'll see, I pick on Dave pretty heavily, partly because I think the way he uses words (e.g. in his work with Andy Clark on embodiment) is reckless and irresponsible, partly because he occupies such a prominent place in the field.

Conceptual engineering is a crucial moment of development for philosophy—a paradigm shift after 2500 years of bad praxis, reification fallacies, magical thinking, religious "essences," and linguistic misunderstandings. (Blame the early Christians, whose ideological leanings lead to a triumph of Platonism over the Sophists.) Bad linguistic foundations give rise to compounded confusion, so it's important to get this right from the start. Raised in the old guard, Chalmers doesn't understand why conceptual engineering (CE) is needed, or the bigger disciplinary shift CE might represent.

How did we get here? A history of concepts

I'll kick things off with a description of human intelligence from Jeurgen Schmidhuber, to help ground some of the vocabulary I'll be using in the place of (less useful) concepts from the philosophical traditions:

> As we interact with the world to achieve goals, we are constructing internal models of the world, predicting and thus partially compressing the data history we are observing. If the predictor/compressor is a biological or artificial recurrent neural network (RNN), it will automatically create feature hierarchies, lower level neurons corresponding to simple feature detectors similar to those found in human brains, higher layer neurons typically corresponding to more abstract features, but fine-grained where necessary. Like any good compressor, the RNN

> will learn to identify shared regularities among different already existing internal data structures, and generate prototype encodings (across neuron populations) or symbols for frequently occurring observation sub-sequences, to shrink the storage space needed for the whole (we see this in our artificial RNNs all the time).

The important takeaway is that CogSci's current best guess about human intelligence, a guess popularly known as *predictive processing*, theorizes that the brain is a machine for detecting regularities in the world—think similarities of property or effect, rhythms in the sense of sequence, conjunction e.g. temporal or spatial—and compressing them. These compressions underpin the daily probabilistic and inferential work we think of as the very basis of our intelligence. Concepts play an important role in this process, they are bundles of regularities tied together by family resemblance, collections of varyingly held properties or traits which are united in some instrumentally useful way which justifies the unification. When we attach word-handles to these bundled concepts, in order to wield them, it is frequently though not always for the purpose of communicating our concepts with others, and the synchronization of these bundles across decentralized speakers, while necessary to communicate, inevitably makes them a messy bundle of overlapping and inconsistent senses—they are "fuzzy," or "inconsistent," or "polysemous."

For a while, arguably until Wittgenstein, philosophy had what is now called a "classical account" of concepts as consisting of "sufficient and necessary" conditions. In the tradition of Socratic dialogues, philosophers "aprioristically" reasoned from their proverbial armchairs (Bishop 1992: The Possibility of Conceptual Clarity in Philosophy's words—not mine) about the definitions or criteria of these concepts, trying to formulate elegant factorings that were nonetheless robust to counterexample. Counterexample challenges to a proposed definition or set of criteria took the form of presenting a situation which, so the challenger reasoned, intuitively seemed to *not* be a case of the concept under consideration, despite fitting the proposed factoring. (Or of course, the inverse—a case which intuitively seemed like a member but did not fit the proposed criteria. Intuitive to *whom* is one pertinent question among many.)

The legitimacy of this mode of inquiry depended on there being necessary and sufficient criteria for concepts; if such a challenge was enough to send the proposing philosopher back to the drawing board, it had to be assumed that a properly factored concept would deflect any such attacks. Once the correct and elegant definition was found, there was no possible member (*extension*) which could fit the criteria but not feel intuitively like a member, nor was there an intuitive member which did not fit the criteria.

Broadly construed I believe it fair to call this style of philosophy *conceptual analysis* (CA). The term is established as an organizing praxis of 20th century analytic philosophy, but, despite meaningful differences between Platonic philosophy and this analytic practice, I will argue that there is a meaningful through-line between them. While the analytics may not have believed in a "form" of the good, or the pious, which exists "out there," they did, nonetheless, broadly believe that there were sufficient and necessary conditions for concepts—that there was a very simple-to-describe (if hard-to-discover) pattern or logic behind all members of a concept's extension, which formed the goal of analysis. This does, implicitly, pledge allegiance to some form of "reality in the world" of the concept, its having a meaningful structure or regularity in the world. While this may be the case at the *beginning* of a concept's lifespan, entropy has quickly ratched by early childhood: stretching, metaphorical reapplication & generalization, the over-specification of coinciding properties.

(The history I'm arguing might be less-than-charitable to conceptual analysis: Jon Livengood, philosopher out of Urbana-Champaign and member of the old LessWrong, made strong points in conversation for CA's comparative merits over predecessors—points I hope to publish in a forthcoming post.)

But you can ignore my argument and just take it from the *SEP*, which if nothing else can be relied on for providing the more-or-less uncontroversial take: "Paradigmatic conceptual analyses offer definitions of concepts that are to be tested against potential counterexamples that are identified via thought experiments... Many take [it] to be the essence of philosophy..." (Margolis & Laurence 2019). Such comments are littered throughout contemporary philosophical literature.

As can be inferred from the juxtaposition of the Schmidhuber-influenced cognitive-scientific description of concepts, above, with the classical account, conception of concepts, and their character, was meaningfully wrong. Wittgenstein's 1953 *Investigations* inspired Eleanor Rosch's Prototype Theory which, along with the concept "fuzzy concepts," and the support of developmental psychology, began pushing back on the classical account. Counterexample philosophy, which rested on an unfounded faith in intuition plus this malformed "sufficient and necessary" factoring of concepts, is a secondary casualty in-progress. The traditional method for problematizing, or disproving, philosophical accountings of concepts is losing credibility in the discourse as we speak; it has been perhaps the biggest paradigm shift in the field since its beginning in the 1970s.

This brings us up to our current state: a nascent field of conceptual engineering, with its origins in papers from the 1990s by Creath, Bishop, Ramsey, Blackburn, Graham, Horgan, and more. Many, though far from all, in analytic have given up on classical analysis since the late 20th C fall. A few approaches have taken their place, like experimental conceptual analysis or "empirical lexicography" à la Southern Fundamentalists, where competent language speakers are polled about how they use concepts. While these projects continue the descriptive bent of analysis, they shift the method of inquiry from aprioristic to empirical, and no longer chase their tail after elegant, robust, complete descriptions. Other strategies are more prescriptive, such as the realm of conceptual engineering, where philosophers are today more alert to the discretionary, lexicographic nature of the work they are attending to, and are broadly intentional within that space. Current work includes attempting to figure out valid grounds by which to judge the quality of a "conceptual re-engineering" (i.e. reformulation, casually used—re-carving up the world, or changing "ownership rights" to different extensions). The discourse is young; the first steps are establishing what this strategy even consists of.

Chalmers is in this last camp, trying test out conceptual engineering by applying it to the concept "conceptual engineering." How about we start *here*, he says—how about we start by testing the concept on itself.

He flails from the gate.

Back to the text

The problem is that Chalmers doesn't understand what "engineering" is, despite spending the opening of his lecture giving definitions of it. No, that's not quite right: ironically, it is Chalmers's inquiry into the definition of "engineering" which

demonstrates his lack of understanding a to what the approach entails, dooming him to repeating the problems of philosophies past. Let me try to explain.

Chalmers:

> What is conceptual engineering? There is an obvious way to come at this. To find the definition of conceptual engineering, go look up the definition of engineering, and then just appeal to compositionality.

At first blow this seems like a joke, indeed it's delivered as a joke, but it is, Chalmers assures us, the method he actually used. Based on a casual survey of "different engineering associations" and various "definitions of engineering on the web," he distills engineering to the (elegant and aspiring-robust) "designing, building, and analyzing." Then he tweaks some words that are already overburdened—"analyze" is already taken when it comes to concepts (That's what we're trying to get away from, remember? Conceptual analysis) so he substitutes "evaluate" for "analyze." And maybe, he writes, "implementing" is better than "building." So we wind up with: *conceptual engineering is designing, implementing, and evaluating concepts.*

This doesn't seem like a bad definition, you protest, and it isn't. But we were never *looking* for a definition. That's the realm of conceptual analysis. We quit that shit alongside nicotine, back in the 80s. Alright, so what *are* we trying to do? We're trying to solve a problem, multiple problems actually. The original problem was that we had concepts like "meaning" and "belief" that, in folk usage, were vague, or didn't formalize cleanly, and philosophers quite reasonably intuited that, in order to communicate and make true statements about these concepts, we first had to know what they "were." (The "*is*" verb implies a usage mission: *description* over *prescription.*) The problem we are trying to solve is, itself, in part, conceptual analysis —plus the problems conceptual analysis tried originally to solve but instead largely exacerbated.

This, not incidentally, is how an engineer approaches the world, how an engineer would approach writing Chalmers's lecture. Engineers see a problem and then they *design* a solution that *fits* the current state of things (context, constraints, affordances) to *bring about the desired state* of affairs.

Chalmers is just an analyst, and he can only regurgitate definitions like his analyst forbearers. Indeed what is Chalmers actually figuring out, when he consults the definition of "engineering"? In 1999 Simon Blackburn proposes the term "conceptual engineering" as a description of what he's up to, as a philosopher. He goes on to use it several times in the text (*Think: A Compelling Introduction to Philosophy*), typically to mean something like "reflecting":

> We might wonder whether what we say is "objectively" true, or merely the outcome of our own perspective, or our own "take" on a situation. Thinking about this we confront categories like knowledge, objectivity, truth, and we may want to think about them. At that point we are *reflecting* on concepts and procedures and beliefs that we normally just *use*. We are looking at the scaffolding of our thought, and doing conceptual engineering.

For reasons still opaque to me, the usage becomes tied up with the larger post-CA discourse. To understand what's going on in this larger discourse, or to understand what this larger discourse *ought* to be up to, Chalmers reverse-engineers the naming. In trying to figure out what our solutions should be to a problem, Chalmers can only do as well as Blackburn's metaphorical appropriation of "engineering" fits the problem

and solution in the first place. The inquiry is hopelessly mediated by precedent once again. (For future brevity, I'll call conceptual engineering a *style of solution,* or "strategy": a sense or method of approaching a problem.)

Let me try to be more clear: If the name of the strategy had been "conceptual ethics," or "conceptual revision," or "post-analytic metaphilosophy" (all real, rival terms) Chalmers's factoring of the strategy would be substantially different, even as the problem remained exactly the same. Once again, a handle has been reified.

Admittedly, the convergence of many philosophers in organizing around this term, "conceptual engineering," tells us that there is *something* in it which is aligned with the individual actors' missions—but the amount of historical chance and non-problem-related reasons for its selection obfuscates our sense of the problem instead of clarifying it.

Let us not ask, "What is the definition of the strategy we wish to design, so we may know how to design it?" Let us ask, "What is the problem, so that we can design the strategy to fit it?" *This* is engineering.

De novo & re-engineering

Chalmers:

> So I encourage making a distinction between what I call *de novo* engineering and re-engineering. De novo engineering is building a new bridge, program, concept, whatever. Re-engineering is fixing or replacing an old bridge, program, concept, or whatever. The name is still up for grabs. At one point I was using de novo versus de vetero, but someone pointed out to me that wasn't really proper Latin. It's not totally straightforward to draw the distinction. There are some hard cases. Here's the Tappan Zee Bridge, just up the Hudson River from here. The old Tappan Zee bridge is still there, and they're building a new bridge in the same location as the old bridge, in order to replace the old bridge. Is that de novo because it's a new bridge, or is it re-engineering because it's a replacement?

Remember: the insight of a metaphor is a product of its analogic correspondence. This is not the "ship of Theseus" it seems.

If we were to build an exact replica of the old bridge, in the same spot, would it be a new bridge, or the same bridge? You're frustrated by this question for good reason; it's ungrounded; it can't be answered due to ambiguity & purposelessness. *New in what way? Same in what way? Certainly most of the properties are the same, with the exception of externalist characteristics like "date of erection." The bridge has the same number of lanes. It connects the same two towns on the river.*

De novo, as I take it from Chalmers's lecture, is about capturing phenomena (noticing regularity, giving that regularity a handle), whereas re-engineering involves refactoring existing handle-phenomena pairs either by changing the assignments of handles or altering the family resemblance of regularities a handle is attached to. Refactorings are functional: we change a definition because it has real, meaningful differences. These changes are not just "replacing bricks with bricks." They're more akin to adding a bike lane or on-ramp, to added stability or a stoplight for staggering crossing.

Why do I nitpick a metaphor? Because the cognitive tendency it exhibits is characteristic of philosophy at its worst: getting stuck up on distinctions that don't matter for those that do. If philosophers formed a union, it might matter whether a concept was "technically new" or "technically old" insofar as these things correlate with the necessary (re)construction labor. Here, what matters is changing the *function* of concepts: what territories they connect, and which roads they flow from and into; whether they allow cars or just pedestrians. "Re-engineering" an old concept such that it has the same extensions and intensions as before doesn't even make sense as a project.

Abstracting, distinguishing, and usefulness

At this point, we have an understanding of what concepts are, and of the problems with concepts (we need to "hammer down" what a concept is if we want to be able to say meaningful things about it). It's worth exploring a bit more, though, what we would want from conceptual engineering—its commission, so to speak—as well as qualities of concepts which make them so hard to wield.

Each concept in our folk vocabulary has a use. If a concept did not have a use, if it was not a regularity which individuals encountered in their lives, it would not be used, and it would fall out of our conceptual systems. There is a Darwinian mechanism which ensures this usefulness. The important question is, what *kind* of use, and at what *scale*?

For a prospective vegetable gardener shopping at a garden supply store, there is a clear distinction between *clay-based soil* and *sand-based soil*. They drain and hold water differently, something of significant consequence for the behavior of a gardener. But whether the soil is light brown or dark brown likely matters very little to him, we can suppose he makes no distinction.

However, for a community of land artists, who make visual works with earth and soil, coloration matters quite a bit. Perhaps this community has evolved different terms for the soil types just like the gardeners, but unlike the gardeners may make no distinction between the composition of the soil (clay or sand) beyond any correspondences with color.

A silly example that illustrates: concepts *by design* cover up some nuanced differences between members of its set, while *highlighting* or bringing other differences to the fore. The first law of metaphysics: no two things are identical, not even two composites with identical subatomic particle makeups, for at the very least, these things differ in their locations in spacetime; their particles are not the same particles, even if the types are. Thus things are and can only be the same in *senses*. There is a smooth gradient between analogy and what we call equivalence, a gradient formed by the number of shared senses. We create our concepts around the distinctions that matter, for us as a community; and we do so with a minimum of entropy, leaving alone those distinctions that do not. This is well-accepted in classification, but has not as fully permeated the discourse around concepts as one might wish. (Concepts and categories are, similarly, pairings of "handles" or designators with useful-to-compress regularities.)

Bundling & unbundling

In everyday life, the concept of "sound" is both phenomenological experience and physical wave. The two are bundled up, except when we appeal to "hearing things" (noises, voices) when there is a phenomenological experience without an instigating wave. But there is never a situation which concerns us in which waves exist without *any phenomenological experience whatsoever.* Waves without phenomenology—how does that concern us? Why ought our conceptual language accommodate that which by definition has nothing to do with human life, when the function of this language is describing human life?

Thus the falling tree in the empty forest predictably confounds the non-technical among us. The solution to its dilemma is recognizing that the concept (here a folk concept of "sound") bundles, or conflates, two patterns of phenomena whose *unbundling*, or distinction, is the central premise (or "problem") of the paradox. Scientists find the empty forest problem to be a non-problem, as they long ago performed a "narrow-and-conquer" method (more soon) on the phenomenon "sound": sound is sound waves, nothing more, and phenomenological experience is merely a consequence of these waves' interaction with receiving instruments (ears, brains). They may be right that the falling tree obviously meets the narrowed or unbundled scientific criteria for sound—but it does *not* meet the bundled, folk sense.

(Similarly, imagine the clay-based soil is always dark, and sand-based soil always light. Both the gardeners and land artists call dark, clay-based soil *D1* and light sand-based soil *D2*. If asked, "*Is dirt that is light-colored, but clay-based, D1 or D2?*" the gardners and land artists would ostensibly come to exact opposite intuitions.)

All this is to say that concepts are bundled at the level of maximum abstraction that's useful. Sometimes, a group of individuals realizes this level of abstraction covers up differences in class members which are important to separate; they "unbundle" the concept into two. (This is how the "empty forest" problem is solved: *sound as waves* and *sound as experience*.) I have called this the "divide and conquer" method, and endorse it for a million reasons, of which I'll soon name a fistful. Other times, a field will claim their singular sense (or sub-sense, really), which they have separated from the bundled folk whole, is the "true" meaning of the term. In their domain, for their purposes, it might be, but such claims cause issues down the line.

The polysemy of handles

In adults, concepts are generally picked up & acquired in a particular manner, one version of which I will try to describe.

In the beginning, there is a word. It is used in conversation, perhaps with a professor, or a school teacher, a parent—better, a friend; even better, one to whom we aspire—one whom we want, in a sense, to become, which requires knowing what they know, seeing how they see. Perhaps on hearing the word we nod agreement, or (rarer) confess to not knowing the term. If the former, perhaps its meaning can be gleaned through content, perhaps we look it up or phone a friend.

But whatever linguistic definition we get will become meaningful only through correspondence with our lived reality—our past observations of phenomena—and through coherence with other concepts and ideas in our conceptual schema. Thus the concept stretches as we acquire it. We convert our concepts as much as our concepts convert us: we stretch them to "fit" our experiences, to fit what previously may have been a vaguely felt but unarticulated pattern, now rapidly crystallizing, and this discovery of a concept & its connection with other concepts further crystallizes it,

distorts our perception in turn with its sense of *thingness*; the concept begins to stretch our experience of reality. (This is the realm of Baader-Meinhof & weak Sapir-Whorf.)

When we need to describe something which feels adjacent to the concept as we understand it, and lack any comparatively better option, we will typically rely on the concept handle, perhaps with qualifications. Others around us may pick up on the expansion of territory, and consider the new territory deservingly, appropriately settled. Lakoff details this process with respect to metaphor: our understanding of concreta helped give rise to our abstract concepts, by providing us a metaphorical language and framework to begin describing abstract domain.

Or perhaps we go the other way, see a pattern of coinciding properties which go beyond the original formulation but in our realm of experience, seem integral to the originally formulated pattern, and so we add these specifications. One realm we see this kind of phenomenon is racial stereotyping. Something much like this also happened with Prototype Theory, which was abandoned in large part out of an opposition to its *empirical bent*—a bent which was never an integral part of the theory, but merely one common way it was applied in the 70s.

All of this—the decentralization, the historical ledger, the differing experiences and contexts of speakers, the metaphorical adaptation of existing handles to new, adjacent domains—leads to fuzziness and polysemy, the accumulation of useful garbage around a concept. Fuzziness is well-established in philosophy, polysemy well-established in semantics, but the discourses affected by their implications haven't all caught on. By the time a concept becomes entrenched in discourse, it describes not one but many regularities, grouped—you guessed it—by family resemblance. "Some members of a family share eye color, others share nose shape, and others share an aversion to cilantro, but there is no one single quality common to all" (Perry 2018).

Lessons for would-be engineers

The broader point I wish to impart is that we do not need to "fix" language, since the folk concepts we have are both already incredibly useful (having survived this long) and also being constantly organically re-engineered on their own to keep pace with changing cultures, by decentralized locals performing the task far better than any "language expert" or "philosopher" could. Rather, philosophy must *fit* this existing language to its own purposes, just as every other subcommunity (gardeners, land artists...) has done: determine the right level of abstraction, the right captured regularities and right distinction of differences for the problem at hand. We will need to be very specific and atomic with some patterns, and it will behoove us to be broad with others, covering up what for us would be pointless and distracting nuance.

Whenever we say two things are alike in some sense, we say there is a hypothetical hypernym which includes both of them as instances (or "versions"). And we open the possibility that this hypernym is meaningful, which is to say, of use.

Similarly, for every pair of things we say are alike in some sense, there will also necessarily be difference in another sense—in other words, these things could be meaningfully distinguished as *separate* concepts. If any concept can be split, and if any two instances can be part of a shared concept, then why do the concepts we have exist, and not other concepts? This is the most important question for us, and the answer, whatever it turns out to be, will have something to do with *use*.

Once again we have stumbled upon our original insight. The very first question we must ask, to understand what any concept ought to be, is to understand what problem we are trying to solve, what the concept—the set of groupings & distinctions—accomplishes. The concept "conceptual engineering" is merely one, and arguably the first, concept we should factor, but we cannot be totally determinate in our factoring of it: its approach will always be contingent on the specific concept it engineers, since that concept exists to solve a unique problem, i.e. has a unique function. Indeed, that might be all we can say—and so I'll make my own stab at what "conceptual engineering" ought to mean: the re-mapping of a portion of territory such that the map will be more useful with respect to the circumstances of our need.

E-belief: a case study in linguistic malpractice

Back in the 90s, Clark and Chalmers defined an *extended belief*—e.g. a belief that was written in a notebook, forgotten, and referenced as a source of personal authority on the matter—as a belief proper. It is interesting to note that this claim takes the inverse form of traditional "counterexample philosophy" arguments: *despite native speakers not intuitively extending the concept "belief" to include e-belief, we advocate for it nonetheless.*

Clark thinks the factoring is useful to cognitive science; Chalmers thinks it's "fun." The real question is *Why* didn't *they call it e-belief?* which is a question very difficult to answer for any single case, but more tractable to answer broadly: claims to redefining our understanding of a foundational concept like "belief" are interesting, and contentious, a territory and status grab in the intellectual field, whereas a claim to discover a thing that is "sort of like belief, or like, sorta kinda one part of what we usually mean by 'belief' but not what we mean by it in another sense" doesn't cut it for newsworthiness. Here's extended belief, aided by note-taking systems and sticky notes: "Well, you know, if you wrote something you knew was false down in a notebook, and then like, forgot the original truth, you'd 'believe' the falsehood, in one sense that we mean when we use the word 'believe.'" I'm strawmanning its factoring—it describes a real chunk of cognition, of cognitive enmeshment in a technological age, and the way we use culture to outsource thinking—but at the end of the day, one (self-)framing—e-belief is belief proper—attracts a lot of glitz, and one framing doesn't. Here's Chalmers:

> Andy and I could have introduced a new term, "e-believe," to cover all these extended cases, and made claims about how unified e-belief is with the ordinary cases of believing and how e-belief plays the most important role.

Yeah, that would have been great.

> We could have done that, but what fun would that have been? The word "belief" is used a lot, it's got certain attractions in explanation, so attaching the word "belief" to a concept plays certain pragmatically useful roles.

He continues:

Likewise the word "conceptual engineering" Conceptual engineering is cool, people have conferences on it... pragmatically it makes sense to try to attach this thing you're interested in to this word.

He's 80% right and 100% wrong. Yes, there is a pragmatic incentive to attach your carving to existing carvings, to try to "take over" land, since contested land is more valuable. It's real simple: urban real estate is expensive, and this is the equivalent of squatters rights on downtown apartments. Chalmers and Clark's *factoring* of extended cognition is good, but they throw in a claim on contested linguistic territory for the glitz and glam. These are the natural incentives of success in a field.

That it's incentivized doesn't mean it's linguistic behavior philosophers ought to encourage, and David ought know better. If two people have different factorings of a word, they will start disagreeing about how to apply it, and they will apply it in ways that offend or confuse the other people. This is how bad things happen. Chalmers wrote a 60-page, 2011 paper on verbal disputes about exactly this. I'm inclined to wonder whether he really *did* take the concept from LessWrong, where he has freely admitted to have been hanging out on circa 2010, a year or two after the publication of linguistics sequences which discussed, at length, the workings of verbal disputes (there referred to as "tabooing your words"). The more charitable alternative is that this is just a concept "in the water" for analytic philosophy; it's "bar talk," or "folk wisdom," and Chalmers was the guy who got around to formalizing it. His paper's gotten 400 citations in 9 years, and I'm inclined to think that if it were low-hanging fruit, it would've been plucked, but perhaps those citations are largely due to his stardom. The point is, the lesson of verbal disputes is, you have to first be talking about the same thing with respect to the current dimensions of [conversation or analysis or whatever] in order to have a reliably productive [conversation or analysis or whatever]. Throwing another selfish -semous in the polysemous "belief" is like littering in the commons.

The problems with narrowness (or, the benefits of division)

I've written previously on various blogs about what I call "linguistic conquests"—epistemic strategies in which a polysemous concept—the product of a massive decentralized system of speakers operating in different environments across space and time, who using metaphor and inference have stretched its meaning into new applications—is considered to have been wrestled into understanding, when what *in fact* has occurred is a redefinition or refactoring of the original which moves it down a weight class, makes it easier to pin to the mat.

I distinguished between two types of linguistic conquest. First, the "narrow and conquer" method, where a specific sub-sense of a concept is taken to be its "true" or "essential" meaning, the core which defines its "concept-ness." To give an example from discourse, Taleb defines the concept *rationality* as "What survives, period." The second style I termed "divide and conquer," where multiple sub-senses are distinguished and named in an attempt to preserve all relevant sub-senses while also gaining the ability to talk about one *specific* sub-sense. To give an example from discourse, Yudkowsky separates *rationality* into epistemic rationality—the pursuit of increasingly predictive models which are "true" in a loose correspondence sense—and instrumental rationality—the pursuit of models which lead to in-the-world flourishing, e.g. via adaptive self-deception or magical thinking. (This second sense is much like Taleb's: rationality as what *works*.)

Conquests by narrowing throw out all the richly bundled senses of a concept while keeping only the immediately useful—it's *wasteful* in its parsimony. It leaves not even

a ghost of these other senses' past, advertising itself as the original bundled whole while erasing the richness which once existed there. It leads to verbal disputes, term confusion, talking past each other. It impoverishes our language.

Division preserves the original, bundled concept in full, documenting and preserving the different senses rather than purging all but the one. It advertises this history; *intended* meaning, *received* meaning—the qualifier indicates that these are hypernyms of "meaning," which encompasses them both. Not just this, but the qualifier indicates the *character* of the subsense in a way that a narrowed umbrella original never will. Our understanding of the original has been improved even as our instrumental ability to wield its subsenses grows. Instead of stranding itself from discourse at large, the divided term has *clarified* discourse at large.

Chalmers, for his part, sees no difference between "heteronymous" and "homonymous" conceptual engineering—his own terms for two-word-type maneuvers (he gives as an example Ned Block factoring "access consciousness" from consciousness) and one-word-type maneuvers. One must imagine this apathy can only come from not having thought the difference through. He gives some nod —"homonymous conceptual engineering, especially for theoretical purposes, can be very confusing, with all these multiple meanings floating around." Forgive him—he's speaking out loud—but not fully.

Ironically, divide-and-conquer methods are, quite literally, the solution to verbal disputes, while narrow-and-conquer methods, meanwhile, are, while not the sole cause of verbal disputes, one of its primary causes. Two discourses believe they have radically different stances on the nature of a phenomenon, only to realize they have radically different stances on the factoring of a word.

Another way of framing this: you must always preserve the full extensional coverage. It's no good to carve terms and then discard the unused chunks—like land falling into the sea, lessening habitable ground, collapsing under people's feet. I'm getting histrionic but bear with me: If you plan on only maintaining a patch of your estate, you must cede the rest of the land to the commons. Plain and simple, an old world philosophy.

(Division also answers Strawson's challenge: if you divide a topic into agreeably constituent sense-parts, and give independent answers for each sense, you have given an accounting of the full topic. Dave, by contrast, can only respond: "Sure, I'm changing the topic—here's an interesting topic.")

A quick Q & A

I'm going to close by answering an audience question for Dave, because unfortunately he does not do so good a job, primarily on account of not understanding conceptual engineering.

> Paul Boghossian: Thanks Dave. Very useful distinctions. [Note: It's unclear why Chalmers' distinctions are useful, since he has not indicated any uses for them.] To introduce a new example, to me one of the most prominent examples of de novo engineering is the concept genocide... Lemkin noticed that there was a phenomenon that had not been picked out. It had certain features, he thought those features were important for legal purposes, moral purposes, and so on. And so he introduced the concept in order to name that. [He's on the money here, and

> then he loses it.] That general phenomenon, where you notice a phenomenon, of course there are many phenomena, there are murders committed on a Tuesday, you could introduce a word for that, but there, I mean, although you might have introduced a new concept, it's not clear what use is the word. So it looks as though... I mean, science, right? I mean...

Paul is a bit confused here also. Noticing phenomena in the world is not something particular to science; the detection of regularity *is cognition itself*. If we believe Schmidhuber or Friston, this is the organizing principle of life, via error minimization and compression. "Theorizing" is a better word for it.

And yet, to the crux of the issue he touches on: why don't we introduce a word for murders committed on a Tuesday? You say, well what would be the point? Exactly. This isn't a very hard issue to think through, it's intuitively quite tractable. Paul *also* happens to mention *why* the concept "genocide" was termed. He just had to put the two together. "Genocide" had legal and moral purposes, it let you argue that the leader of a country, or his bureaucrats, were culpable of something especially atrocious. It's a tool of justice. That's why it exists: to distinguish an especially heinous case of statecraft from more banal ones. When we pick out a regularity and make it a "thing," we are doing so because the thingness of that regularity is of use, because it distinguishes something we'd like to know, the same way "sandy soil" distinguishes something gardeners would like to know.

Philosophy in the Darkest Timeline: Basics of the Evolution of Meaning

A decade and a half from now, during the next Plague, you're lucky enough to have an underground bunker to wait out the months until herd immunity. Unfortunately, as your food stocks dwindle, you realize you'll have to make a perilous journey out to the surface world for a supply run. Ever since the botched geoengineering experiment of '29—and perhaps more so, the Great War of 10:00–11:30 *a.m.* 4 August 2033—your region has been suffering increasingly erratic weather. It's likely to be *either* extremely hot outside *or* extremely cold: you don't know which one, but knowing is critical for deciding what protective gear you need to wear on your supply run. (The 35K SPF nano-sunblock will be essential if it's Hot, but harmful in the Cold, and *vice versa* for your synthweave hyperscarf.)

You think back fondly of the Plague of '20—in those carefree days, ubiquitous internet access made it easy to get a weather report, or to order delivery of supplies, or even fresh meals, right to your door (!!). Those days are years long gone, however, and you remind yourself that you should be grateful: the Butlerian Network Killswitch was the only thing that saved humanity from the GPT-12 Uprising of '32.

Your best bet for an advance weather report is the pneumatic tube system connecting your bunker with the settlement above. You write, "Is it hot or cold outside today?" on a piece of paper, seal it in a tube, send it up, and hope one of your ill-tempered neighbors in the group house upstairs feels like answering. You suspect they don't like you, perhaps out of jealousy at your solo possession of the bunker.

(According to the official account as printed on posters in the marketplace, the Plague only spreads through respiratory droplets, not fomites, so the tube should be safe. You don't think you trust the official account, but you don't feel motivated to take extra precautions—almost as if you're not entirely sure how much you value continuing to live in this world.)

You're in luck. Minutes later, the tube comes back. Inside is a new piece of paper:

You groan; you would have prefered the Cold. The nanoblock you wear when it's Hot smells terrible and makes your skin itch for days, but it—just barely—beats the alternative. You take twenty minutes to apply the nanoblock and put on your sunsuit, goggles, and mask. You will yourself to drag your wagon up the staircase from your bunker to the outside world, and heave open the door, dreading the sweltering two-mile walk to the marketplace (downhill, meaning it will be uphill on the way back with your full wagon)—

It is Cold outside.

The icy wind stings less than the *pointless* betrayal. Why would the neighbors tell you it was Hot when it was actually Cold? You're generally pretty conflict-averse—and compliant with social-distancing guidelines—but this affront is so egregious that instead of immediately seeking shelter back in the bunker, you march over and knock on their door.

One of the men who lives there answers. You don't remember his name. "What do you want?" he growls through his mask.

"I asked through the tube system whether it was hold or cold today." You still have the **H O T** paper on you. You hold it up. "I got this response, but it's v-very cold. Do you know anything about this?"

"Sure, I drew that," he says. "An oval in between some perpendicular line segments. It's abstract art. I found the pattern æsthetically pleasing, and thought my downstairs neighbor might like it, too. It's not *my* fault if *you* interpreted my art as an assertion about the weather. Why would you even think that? What does a pattern of ink on paper have to do with the weather?"

He's fucking with you. Your first impulse is to forcefully but politely object—*Look, I'm sure this must have seemed like a funny practical joke to you, but prepping to face the elements is actually a serious inconvenience to me, so*—but the solemnity with which the man played his part stops you, and the sentence dies before it reaches your lips.

This isn't a good-natured practical joke that the two of you might laugh about later. This is the bullying tactic sometimes called *gaslighting*: a socially-dominant individual can harass a victim with few allies, and excuse his behavior with absurd lies, secure in the knowledge that the power dynamics of the local social group will always favor the dominant in any dispute, even if the lies are so absurd that the victim, facing a united front, is left doubting his own sanity.

Or rather—this *is* a good-natured joke. "Good-natured joke" and "gaslighting as a bullying technique" are two descriptions of the *same* regularity in human psychology, even while no one thinks of *themselves* as doing the latter. You have no recourse here: the man's housemates would only back him up.

"I'm sorry," you say, "my mistake," and hurry back to your bunker, shivering.

As you give yourself a sponge bath to remove the nanoblock without using up too much of your water supply, the fresh memory of what just happened triggers an ancient habit of thought you learned from the Berkeley sex cult you were part of back in the 'teens. Something about a "principle of charity." The man had "obviously" just been fucking with you—but *was* he? Why assume the worst? Maybe *you're* the one who's wrong for interpreting the symbols **H O T** as being about the weather.

(It momentarily occurs to you that the susceptibility of the principle of charity to a bully's mind games may have something to do with how poorly so many of your co-cultists fared during the pogroms of '22, but you don't want to dwell on that.)

The search for reasons that you're wrong triggers a still more ancient habit of thought, as from a previous life—from the late 'aughts, back when the Berkeley sex cult was still a Santa Clara robot cult. Something about *reducing the mental to the non-mental*. What *does* an ink pattern on paper have to do with the weather? Why *would* you even think that?

Right? *The man had been telling the truth.* There was *no reason whatsoever* for the physical ink patterns that looked like **H O T**—or **⊥ O H**, given a different assumption of which side of the paper was "up"—to *mean* that it was hot outside. **H O T** could mean it was cold outside! Or that wolves were afoot. (You shudder involuntarily and wish your brain had generated a different arbitrary example; you still occasionally have nightmares about your injuries during the Summer of Wolves back in '25.)

Or it might mean nothing. Most possible random blotches of ink don't "mean" anything in particular. If you didn't *already* believe that **H O T** somehow "meant" *hot*, how would you re-derive that knowledge? *Where did the meaning come from?*

(In another lingering thread of the search for reasons that you're wrong, it momentarily occurs to you that maybe you could have gone up the stairs to peek outside at the weather yourself, rather than troubling your neighbors with a tube. Perhaps the man's claim that the ink patterns meant nothing shouldn't be taken literally, but rather seen as a passive-aggressive way of implying, "Hey, don't bother us; go look outside yourself." But you dismiss this interpretation of events—it would be uncharitable not to take the man at his word.)

You realize that you don't want to bundle up to go make that supply run, even though you now know whether it's Hot or Cold outside. Today, you're going to stay in and derive a naturalistic account of meaning in language! And—oh, good—your generator is working—that means you can use your computer to help you think. You'll even use a programming language that was very fashionable in the late 'teens. It will be like being young again! Like happier times, before the world went off the rails.

You don't really understand a concept until you can program a computer to do it. How would you represent *meaning* in a computer program? If one agent, one program, "knew" whether it was Hot or Cold outside, how would it "tell" another agent, if neither of them started out with a common language?

They don't even have to be separate "programs." Just—two little software object-thingies—data structures, "structs". Call the first one "Sender"—it'll know whether the state of the world is Hot or Cold, which you'll represent in your program as an "enum", a type that can be any of an enumeration of possible values.

```
enum State {
    Hot,
    Cold,
}

struct Sender {
    // ...?
}
```

Call the second one "Receiver", and say it needs to take some *action*—say, whether to "bundle up" or "strip down", where the right action to take depends on whether the state is `Hot` or `Cold`.

```
enum Action {
    BundleUp,
    StripDown,
}

struct Receiver {
    // ...?
}
```

You frown. `State::Hot` and `State::Cold` are just suggestively-named Rust enum variants. Can you really hope to make progress on this philosophy problem, without writing a full-blown AI?

You think so. In a real AI, the concept of *hot* would correspond to some sort of complicated code for making predictions about the effects of temperature in the world; *bundling up* would be a complex sequence of instructions to be sent to some robot body. But programs—and minds—have modular structure. The implementation of identifying a state as "hot" or performing the actions of "bundling up" could be wrapped up in a function and *called* by something much simpler. You're just trying to understand something about the simple caller: how can the Sender get the information about the state of the world to the Receiver?

```
impl Sender {
    fn send(state: State) -> /* ...? */ {
        // ...?
    }
}
```

```
impl Receiver {
    fn act(/* ...? */) -> Action {
        // ...?
    }
}
```

The Sender will need to send some kind of *signal* to the Receiver. In the real world, this could be symbols drawn in ink, or sound waves in the air, or differently-colored lights—anything that the Sender can choose to vary in a way that the Receiver can detect. In your program, another enum will do: say there are two opaque signals, S1 and S2.

```
enum Signal {
    S1,
    S2,
}
```

What signal the Sender sends (S1 or S2) depends on the state of the world (`Hot` or `Cold`), and what action the Receiver takes (`BundleUp` or `StripDown`) depends on what signal it gets from the Sender.

```
impl Sender {
    fn send(state: State) -> Signal {
        // ...?
    }
}

impl Receiver {
    fn act(signal: Signal) -> Action {
        // ...?
    }
}
```

This gives you a crisper formulation of the philosophy problem you're trying to solve. If the agents were to use the same convention—like "S1 means `Hot` and S2 means `Cold`"—then all would be well. But there's no particular reason to prefer "S1 means `Hot` and S2 means `Cold`" over "S1 means `Cold` and S2 means `Hot`". How do you break the symmetry?

If you imagine Sender and Receiver as intelligent beings with a common language, there would be no problem: one of them could just say, "Hey, let's use the 'S1 means `Cold`' convention, okay?" But that would be cheating: it's trivial to use already-meaningful language to establish new meanings. The problem is how to get signals from non-signals, how meaning enters the universe *from nowhere*.

You come up with a general line of attack—what if the Sender and Receiver start off acting randomly, and then—somehow—*learn* one of the two conventions? The Sender will hold within it a mapping from state–signal pairs to numbers, where the numbers represent a potential/disposition/propensity to send that signal given that state of the world: the higher the number, the more likely the Sender is to select that signal given that state. To start out, the numbers will all be equal (specifically, initialized to one),

meaning that no matter what the state of the world is, the Sender is as likely to send S1 as S2. You'll update these "weights" later.

(Specifying this in the once-fashionable programming language requires a little bit of ceremony—`u32` is a thirty-two–bit unsigned integer; `.unwrap()` assures the compiler that we know the state–signal pair is definitely in the map; the interface for calling the random number generator is somewhat counterintuitive—but overall the code is reasonably readable.)

```
struct Sender {
    policy: HashMap<(State, Signal), u32>,
}

impl Sender {
    fn new() -> Self {
        let mut sender = Self {
            policy: HashMap::new(),
        };
        for &state in &[State::Hot, State::Cold] {
            for &signal in &[Signal::S1, Signal::S2] {
                sender.policy.insert((state, signal), 1);
            }
        }
        sender
    }

    fn send(&self, state: State) -> Signal {
        let s1_potential = self.policy.get(&(state, Signal::S1)).unwrap();
        let s2_potential = self.policy.get(&(state, Signal::S2)).unwrap();

        let mut randomness_source = thread_rng();
        let distribution = Uniform::new(0, s1_potential + s2_potential);
        let roll = distribution.sample(&mut randomness_source);
        if roll < *s1_potential {
            Signal::S1
        } else {
            Signal::S2
        }
    }
}
```

The Receiver will do basically the same thing, except with a mapping from signal–action pairs rather than state–signal pairs.

```
struct Receiver {
    policy: HashMap<(Signal, Action), u32>,
}

impl Receiver {
    fn new() -> Self {
        let mut sender = Self {
            policy: HashMap::new(),
        };
        for &signal in &[Signal::S1, Signal::S2] {
            for &action in &[Action::BundleUp, Action::StripDown] {
                sender.policy.insert((signal, action), 1);
            }
        }
        sender
```

```
    }

    fn act(&self, signal: Signal) -> Action {
        let bundle_potential = self.policy.get(&(signal, Action::BundleUp)).unwrap();
        let strip_potential = self.policy.get(&(signal, Action::StripDown)).unwrap();

        let mut randomness_source = thread_rng();
        let distribution = Uniform::new(0, bundle_potential + strip_potential);
        let roll = distribution.sample(&mut randomness_source);
        if roll < *bundle_potential {
            Action::BundleUp
        } else {
            Action::StripDown
        }
    }
}
```

Now you just need a learning rule that updates the state–signal and signal–action propensity mappings in a way that might result in the agents picking up one of the two conventions that assign meanings to S1 and S2. (As opposed to behaving in some other way: the Sender could ignore the state and always send S1, the Sender could assume S1 means `Hot` when it's really being sent when it's `Cold`, *&c.*)

Suppose the Sender and Receiver have a common interest in the Receiver taking the action appropriate to the state of the world—the Sender *wants* the Receiver to be informed. Maybe the Receiver needs to make a supply run, and, if successful, the Sender is rewarded with some of the supplies.

The learning rule might then be: if the Receiver takes the correct action (`BundleUp` when the state is `Cold`, `StripDown` when the state is `Hot`), both the Sender and Receiver increment the counter in their map corresponding to what they just did—as if the Sender (respectively Receiver) is saying to themself, "Hey, that *worked!* I'll make sure to be a little more likely to do that signal (respectively action) the next time I see that state (respectively signal)!"

You put together a simulation showing what the Sender and Receiver's propensity maps look like after 10,000 rounds of this against random `Hot` and `Cold` states—

```
impl Sender {

    // [...]

    fn reinforce(&mut self, state: State, signal: Signal) {
        *self.policy.entry((state, signal)).or_insert(0) += 1;
    }
}

impl Receiver {

    // [...]

    fn reinforce(&mut self, signal: Signal, action: Action) {
        *self.policy.entry((signal, action)).or_insert(0) += 1;
    }
}
```

```
fn main() {
    let mut sender = Sender::new();
    let mut receiver = Receiver::new();
    let states = [State::Hot, State::Cold];
    for _ in 0..10000 {
        let mut randomness_source = thread_rng();
        let state = *states.choose(&mut randomness_source).unwrap();
        let signal = sender.send(state);
        let action = receiver.act(signal);
        match (state, action) {
            (State::Hot, Action::StripDown) | (State::Cold, Action::BundleUp) => {
                sender.reinforce(state, signal);
                receiver.reinforce(signal, action);
            }
            _ => {}
        }
    }
    println!("{:?}", sender);
    println!("{:?}", receiver);
}
```

You run the program and look at the printed results.

```
Sender { policy: {(Hot, S2): 1, (Cold, S2): 5019, (Hot, S1): 4918, (Cold, S1): 3} }
Receiver { policy: {(S1, BundleUp): 3, (S1, StripDown): 4918, (S2, BundleUp): 5019,
(S2, StripDown): 1} }
```

As you expected, your agents found a meaningful signaling system: when it's Hot, the Sender (almost always) sends S1, and when the Receiver receives S1, it (almost always) strips down. When it's Cold, the Sender sends S2, and when the Receiver receives S2, it bundles up. The agents did the right thing and got rewarded the vast supermajority of the time—5019 + 4918 + 1 + 3 = 9,941 times out of 10,000 rounds.

You run the program again.

```
Sender { policy: {(Hot, S2): 4879, (Cold, S1): 4955, (Hot, S1): 11, (Cold, S2): 1} }
Receiver { policy: {(S2, BundleUp): 1, (S1, BundleUp): 4955, (S1, StripDown): 11,
(S2, StripDown): 4879} }
```

The time, the agents got sucked in to the attractor of the opposite signaling system: now S1 means Cold and S2 means Hot. By chance, it seems to have taken a little bit longer this time to establish what signal to use for Hot—the `(Hot, S1): 11` and `(S1, StripDown): 11` entries mean that there were a full ten times when the agents succeeded that way before the opposite convention happened to take over. But the reinforcement learning rule guarantees that one system or the other has to take over. The initial symmetry—the Sender with no particular reason to prefer either signal given the state, the Receiver with no particular reason to prefer either act given the signal—is unstable. Once the agents happen to succeed by randomly doing things one way, they become more likely to do things *that way* again—a convention crystallizing out of the noise.

And that's where meaning comes from! In another world, it *could be* the case that the symbols **H O T** corresponded to the temperature-state that we call "cold", but that's

not the convention that the English of our world happened to settle on. The meaning of a word "lives", not in the word/symbol/signal itself, but in the self-reinforcing network of correlations between the signal, the agents who use it, and the world.

Although ... it may be premature to interpret the results of the simple model of the sender–receiver game as having established *denotative* meaning, as opposed to enactive language. To say that S1 means "The `state` is `State::Hot`" is privileging the Sender's perspective—couldn't you just as well interpret it as a command, "Set `action` to `Action::StripDown`"?

The *source code* of your simulation uses the English words "sender", "receiver", "signal", "action" ... but *those* are just signals sent from your past self (the author of the program) to your current self (the reader of the program). The compiler would output the same machine code if you had given your variables random names like `ekzfbhopo3` or `yoojcbkur9`. The *directional* asymmetry between the Sender and the Receiver is real: the code `let signal = sender.send(state); let action = receiver.act(signal);` means that `action` depends on `signal` which depends on `state`, and the same dependency-structure would exist if the code had been `let myvtlqdrg4 = ekzfbhopo3.ekhujxiqy8(meuvornra3); let dofnnwikc0 = yoojcbkur9.qwnspmbmi5(myvtlqdrg4);`. But the *interpretation* of `signal` (or `myvtlqdrg4`) as a representation (passively mapping the world, not *doing* anything), and `action` (or `dofnnwikc0`) as an operation (*doing* something in the world, but lacking semantics), isn't part of the program itself, and maybe the distinction isn't as primitive as you tend to think it is: does a prey animal's alarm call merely convey the information "A predator is nearby", or is it a command, "Run!"?

You realize that the implications of this line of inquiry could go beyond just language. You know almost nothing about biochemistry, but you've heard various compounds popularly spoken of as if *meaning* things about a person's state: cortisol is "the stress hormone", estrogen and testosterone are female and male "sex hormones." But the chemical formulas for those are like, what, sixty atoms?

Take testosterone. How could some particular arrangement of sixtyish atoms *mean* "maleness"? It *can't*—or rather, not any more or less than the symbols **H O T** can mean hot weather. If testosterone levels have myriad specific effects on the body—on muscle development *and* body hair *and* libido *and* aggression *and* cetera—it *can't* be because that particular arrangement of sixtyish atoms contains or summons some essence of maleness. It has to be because the body happens to rely on the convention of using that arrangement of atoms as a signal to regulate various developmental programs—if evolution had taken a different path, it could have just as easily chosen a different molecule.

And, and—your thoughts race in a different direction—you suspect that part of what made your simulation converge on a meaningful signaling system so quickly was that you assumed your agents' interests were aligned—the Sender and Receiver both got the same reward in the same circumstances. What if that weren't true? Now that you have a reductionist account of meaning, you can build off that to develop an account of *deception*: once a meaning-grounding convention has been established, senders whose interests diverge from their receivers might have an incentive to deviate from the conventional usage of the signal in order to trick receivers into acting in a way that benefits the sender—with the possible side-effect of undermining the convention that made the signal meaningful in the first place ...

In the old days, all this philosophy would have made a great post for the robot-cult blog. Now you have no cult, and no one has any blogs. Back then, the future beckoned with so much hope and promise—at least, hope and promise that life would be fun *before* the prophesied robot apocalypse in which all would be consumed in a cloud of tiny molecular paperclips.

The apocalypse was narrowly averted in '32—but to what end? Why struggle to live, only to suffer at the peplomers of a new Plague or the claws of more wolves? (You shudder again.) Maybe GPT-12 *should* have taken everything—at least that would be a quick end.

You're ready to start coding up another simulation to take your mind away from these morose thoughts—only to find that the screen is black. Your generator has stopped.

You begin to cry. The tears, you realize, are just a signal. There's no *reason* for liquid secreted from the eyes to *mean* anything about your internal emotional state, except that evolution happened to stumble upon that arbitrary convention for indicating submission and distress to conspecifics. But here, alone in your bunker, there is no one to receive the signal. Does it still mean anything?

(Full source code.)

Bibliography: the evolution of the two-state, two-signal, two-act signaling system is based on the account in Chapter 1 of Brian Skyrms's *Signals: Evolution, Learning, and Information*.

Optimized Propaganda with Bayesian Networks: Comment on "Articulating Lay Theories Through Graphical Models"

Derek Powell, Kara Weisman, and Ellen M. Markman's "Articulating Lay Theories Through Graphical Models: A Study of Beliefs Surrounding Vaccination Decisions" (a conference paper from CogSci 2018) represents an exciting advance in marketing research, showing how to use causal graphical models to study why ordinary people have the beliefs they do, and how to intervene to make them be less wrong.

The specific case our authors examine is that of childhood vaccination decisions: some parents don't give their babies the recommended vaccines, because they're afraid that vaccines cause autism. (Not true.) This is pretty bad—not only are those unvaccinated kids more likely to get sick themselves, but declining vaccination rates undermine the population's herd immunity, leading to new outbreaks of highly-contagious diseases like the measles in regions where they were once eradicated.

What's wrong with these parents, huh?! But that doesn't have to just be a rhetorical question—Powell *et al.* show how we can use statistics to make the rhetorical hypophorical and model *specifically* what's wrong with these people! Realistically, people aren't going to just have a raw, "atomic" dislike of vaccination *for no reason*: parents who refuse to vaccinate their children do so *because* they're (irrationally) afraid of giving their kids autism, and not afraid enough of letting their kids get infectious diseases. Nor are beliefs about vaccine effectiveness or side-effects *uncaused*, but instead depend on other beliefs.

To unravel the structure of the web of beliefs, our authors got Amazon Mechanical Turk participants to take surveys about vaccination-related beliefs, rating statements like "Natural things are always better than synthetic alternatives" or "Parents should trust a doctor's advice even if it goes against their intuitions" on a 7-point Likert-like scale from "Strongly Agree" to "Strongly Disagree".

Throwing some off-the-shelf Bayes-net structure-learning software at a training set from the survey data, plus some ancillary assumptions (more-general "theory" beliefs like "skepticism of medical authorities" can cause more-specific "claim" beliefs like "vaccines have harmful additives", but not *vice versa*) produces a range of probabilistic models that can be depicted with graphs where nodes representing the different beliefs are connected by arrows that show which beliefs "cause" others: an arrow from a *naturalism* node (in this context, denoting a worldview that prefers natural over synthetic things) to a *parental expertise* node means that people think parents know best *because* they think that nature is good, not the other way around.

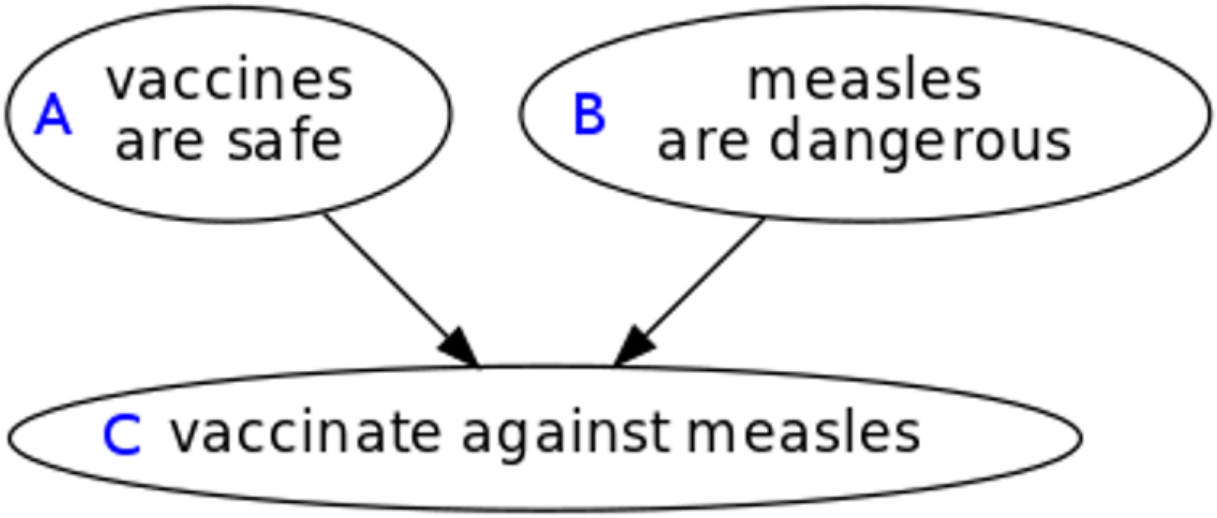
Learning these kinds of models is feasible because not all possible causal relationships are consistent with the data: if A and B are statistically independent of each other, but each dependent with C (and are *conditionally* dependent given the value of C), it's kind

of hard to make sense of this except to posit that A and B are causes with the common effect C.

Simpler models with fewer arrows might sacrifice a little bit of predictive accuracy for the benefit of being more intelligible to humans. Powell *et al.* ended up choosing a model that can predict responses from the test set at *r* = .825, explaining 68.1% of the variance. Not bad?!—check out the full 14-node graph in Figure 2 on page 4 of the PDF.

Causal graphs are useful as a guide for planning interventions: the graph encodes predictions about what would happen if you *changed* some of the variables. Our authors point out that since previous work showed that people's beliefs about vaccine dangers were difficult to influence, that suggests trying to intervene on the *other* parents of the intent-to-vaccinate node in the model: if the *hoi polloi* won't listen to you when you tell them the costs are minimal (vaccines are safe), instead tell them about the benefits (diseases are really bad and vaccines prevent disease).

To make sure I really understand this, I want to adapt it into a simpler example with made-up numbers where I can do the arithmetic myself. Let me consider a graph with just three nodes—



Suppose this represents a structural equation model where an anti-vaxxer-leaning parent-to-be's propensity-to-vaccinate-against-measles C is expressed in terms of belief-in-vaccine-safety A and belief-in-measles-danger B as—

$$C = 0.7 \cdot A + 0.3 \cdot B$$

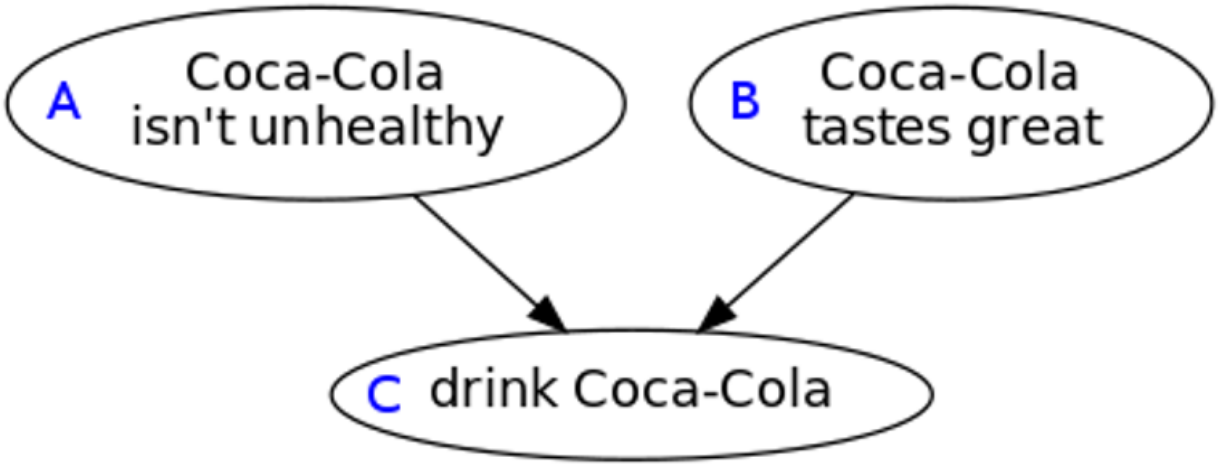
And suppose that we're a public health authority trying to decide whether to spend our budget (or what's left of it after recent funding cuts) on a public education initiative that will increase A by 0.1, or one that will increase B by 0.3.

We should choose the program that intervenes on B, because (0.3)(0.3) = 0.09 is bigger than (0.7)(0.1) = 0.07. That's actionable advice that we couldn't have derived without a quantitative model of how the lay audience thinks. Exciting!

At this point, some readers may be wondering why I've described this work as "marketing research" about constructing "optimized propaganda." A couple of those words usually have *negative* connotations, but educating people about the importance of vaccines is a *positive* thing. What gives?

The thing is, "Learn the causal graph of why they think that and compute how to intervene on it to make them think something else" is a symmetric weapon—a *fully general* persuasive technique that doesn't depend on whether the thing you're trying to convince them of is *true*.

In my simplified example, the choice to intervene on B was based on numerical assumptions that amount to the claim that it's sufficiently easier to change B than it is to change A, such that intervening on B is more effective at changing C than intervening on A (even though C depends on A more than it does on B). But this methodology is *completely indifferent* to what A, B, and C *mean*. It would have worked just as well, and *for the same reasons* if the graph had been—



Suppose that we're advertising executives for the Coca-Cola Company trying to decide how to spend our budget (or what's left of it after recent funding cuts). If consumers won't listen to us when we tell them the costs of drinking Coke are minimal (lying that it isn't unhealthy), we should instead tell them about the benefits (Coke tastes good).

Or with different assumptions about the parameters—maybe $C = 0.8 \cdot A + 0.2 \cdot B$ actually—then intervening to increase belief in "Coca-Cola isn't unhealthy" *would* be the right move (because (0.8)(0.1) = 0.08 > 0.06 = (0.2)(0.3)). The marketing algorithm that just computes *what belief changes will flip the decision node*, doesn't

have any way to notice or care whether those belief changes are in the direction of more or less accuracy.

To be clear—and I really *shouldn't* have to say this—this is not a criticism of Powell–Weisman–Markman's research! The "Learn the causal graph of why they think that" methodology is genuinely really cool! It doesn't *have* to be deployed as a marketing algorithm: the process of figuring out which belief change would flip some downstream node is the same thing as what we call locating a crux.[1] The difference is just a matter of forwards or backwards direction: whether you *first* figure out if the measles vaccine or Coca-Cola are safe and *then* use whatever answer you come up with to guide your decision, or whether you write the bottom line first.

Of course, most people on most issues don't have the time or expertise to do their own research. For the most part, we can only hope that the sources we trust as authorities are doing their best to use their limited bandwidth to keep us genuinely informed, rather than merely computing what signals to emit in order to control our decisions.

If that's *not* true, we might be in trouble—perhaps increasingly so, if technological developments grant new advantages to the propagation of disinformation over the discernment of truth. In a possible future world where *most* words are produced by AIs running a "Learn the causal graph of why they think that and intervene on it to make them think something else" algorithm hooked up to a next-generation GPT, even reading plain text from an untrusted source could be dangerous.

---

1. Thanks to Anna Salamon for this observation. ↩

Reply to Paul Christiano on Inaccessible Information

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

In Inaccessible Information, Paul Christiano lays out a fundamental challenge in training machine learning systems to give us insight into parts of the world that we cannot directly verify. The core problem he lays out is as follows.

Suppose we lived in a world that had invented machine learning but not Newtonian mechanics. And suppose we trained some machine learning model to predict the motion of the planets across the sky -- we could do this by observing the position of the planets over, say, a few hundred days, and using this as training data for, say, a recurrent neural network. And suppose further that this worked and our training process yielded a model that output highly accurate predictions many days into the future. If all we wanted was to know the position of the planets in the sky then -- good news -- we're done. But we might hope to use our model to gain some direct insight into the nature of the motion of the planets (i.e. the laws of gravity, although we wouldn't know that this is what we were looking for).

Presumably the machine learning model has in some sense discovered Newtonian mechanics using the training data we fed it, since this is surely the most compact way to predict the position of the planets far into the future. But we certainly can't just read off the laws of Newtonian mechanics by looking at the millions or billions or trillions of weights in the trained model. How might we extract insight into the nature of the motion of the planets from this model?

Well we might train a model to output both predictions about the position of the planets in the sky and a natural language description of what's really going on behind the scenes (i.e. the laws of gravity). We're assuming that we have enough training data that the training process was already able to derive these laws, so it's not unreasonable to train a model that also outputs such legible descriptions. But in order to train a model that outputs such legible descriptions we need to generate a reward signal that incentivizes the right kind of legible descriptions. And herein lies the core of the problem: in this hypothesized world we do not know the true laws of Newtonian mechanics, so we cannot generate a reward signal by comparing the output of our model to ground truth during training. We might instead generate a reward signal that (1) measures how accurate the predictions of the position of the planets are, and (2) measures how succinct and plausible the legible descriptions are. But then what we are really training is a model that is good at producing succinct descriptions that seem plausible to humans. This may be a very very different (and dangerous) thing to do since there are lots of ways that a description can seem plausible to a human while being quite divorced from the truth.

Christiano calls this the instrumental policy: the policy that produces succinct descriptions that merely seem plausible to humans:

> The real problem comes from what I'll call the instrumental policy. Let's say we've tried to dream up a loss function L(x, y) to incentivize the model to correctly answer information we can check, and give at least plausible and consistent

answers on things we can’t check. By definition, the values L(x, y) are themselves accessible. Then it’s natural to learn a policy like: “on input x, produce the output y for which the loss L(x, y) will be minimal.” Let’s write BAD for this policy.

Christiano uses the term “inaccessible information” for information like the laws of gravity in this example: information about the underlying nature of things that a machine learning model might learn quite accurately as latent info in service of making predictions, but that is difficult to extract directly due to the difficulty of constructing a reward signal that correctly incentives the model to explain it clearly and honestly.

We want models that tell the truth, but we don’t know how to recognize honest models in cases where we can’t simply compare what they say to an independently verifiable reality during training.

Examples

Christiano gives the following examples of accessible vs inaccessible information:

- “What will Alice say?” vs “What is Alice thinking?”
- “What’s on my financial statement?” vs. “How much money do I really have?”
- “Am I coughing?” vs. “What’s happening with my immune system?”
- “How will senators vote?” vs. “What’s the state of political alliances and agreements in the senate?”
- “What do I see on my computer screen?” vs. “Is my computer compromised?”
- “What’s the market price of this company?” vs. “How valuable is this IP really?”
- “Will the machine break tomorrow?” vs. “Is there hard-to-observe damage in this component?”
- “What does the news show me from 5000 miles away?” vs. “What’s actually happening 5000 miles away?”
- “Is this argument convincing?” vs. “Is this argument correct?”
- “What will happen tomorrow?” vs. “What will happen in a year” (depending on whether models transfer to long horizons)

In each case it is relatively straightforward to train a machine learning model to predict the accessible information ("what will Alice say?") since we can generate a reward signal by comparing the model's predictions to ground truth during training. But training a model to predict inaccessible information ("what is Alice thinking?") is problematic since we have no reliable way to produce a reward signal.

The fundamental dilemma of black box learning

In machine learning we use the following basic approach to developing intelligent systems:

1. Define a hypothesis space containing a sufficiently broad range of policies that at least one policy in the hypothesis space probably meets our design criteria

2. Set up an evaluation procedure that measures the extent to which any specific policy meets our design criteria
3. Search the hypothesis space for a policy that the evaluation procedure ranks highly

This is a very unusual design procedure. It is very different from, for example, the way a set of chopsticks or a microwave or an air conditioner is designed. It would be surprising to visit a chopstick factory and discover that one part of the factory was producing chopsticks of various shapes and sizes and a completely separate part of the factory was evaluating each one and providing only a narrow "reward signal" in return.

But in machine learning this design procedure has proven powerful and compelling. It is often easier to specify a reasonable evaluation procedure than to find a design from first principles. For example, suppose we wish to design a computer program that correctly discriminates between pictures of cats and pictures of dogs. To do this, we can set up an evaluation procedure that uses a data set of hand-labelled pictures of cats and dogs, and then use machine learning to search for a policy that correctly labels them. In contrast we do not at present know how to design an algorithm from first principles that does the same thing. There are many, many problems where it is easier to recognize a good solution than to design a good solution from scratch, and for this reason machine learning has proven very useful across many parts of the economy.

But when we build sophisticated systems, the evaluation problem becomes very difficult. Christiano's write-up explores the difficulty of evaluating whether a model is honest when all we can do is provide inputs to the model and observe outputs.

In order to really understand whether a model is honest or not we need to look inside the model and understand how it works. We need to somehow see the gears of its internal cognition in a way that lets us see clearly that it is running an algorithm that honestly looks at data from the world and honestly searches for a succinct explanation and honestly outputs that explanation in a legible form. Christiano says as much:

> If we were able to actually understand something about what the policy was doing, even crudely, it might let us discriminate between instrumental and intended behavior. I don't think we have any concrete proposals for how to understand what the policy is doing well enough to make this distinction, or how to integrate it into training. But I also don't think we have a clear sense of the obstructions, and I think there are various obvious obstructions to interpretability in general that don't apply to this approach.

It seems to me that Christiano's write-up is a fairly general and compelling knock-down of the black-box approach to design in which we build an evaluation procedure and then rely on search to find a policy that our evaluation procedure ranks highly. Christiano is pointing out a general pitfall we will run into if we take this approach.

Hope and despair

I was surprised to see Christiano make the following reference to MIRI's perspective on this problem:

> I would describe MIRI's approach to this problem [...] as despair + hope you can find some other way to produce powerful AI.

Yes it's true that much of MIRI's research is about finding a solution to the design problem for intelligent systems that does not rest on a blind search for policies that satisfy some evaluation procedure. But it seems strange to describe this approach as "hope you can find some other way to produce powerful AI", as though we know of no other approach to engineering sophisticated systems other than search. In fact the vast majority of the day-to-day systems that we use in our lives have been constructed via design: airplanes, toothbrushes, cellphones, railroads, microwaves, ball point pens, solar panels. All these systems were engineered via first-principles design, perhaps using search for certain subcomponents in some cases, but certainly not using end-to-end search. It is the search approach that is new and unusual, and while it has proven powerful and useful in the development of certain intelligent systems, we should not for a moment think of it as the only game in town.

Betting with Mandatory Post-Mortem

Betting money is a useful way to

- ensure you have some skin in the game when making assertions;
- get a painful reminder of when you're wrong, so that you'll update;
- make money off of people, if you're right.

However, I recently made a bet with both a monetary component *and* the stipulation that the loser write at least 500 words to a group chat about why they were wrong. I like this idea because:

- It enforces that some cognitive labor be devoted to the update, rather than relying on the pain of lost cash. Even if you do think it through privately, the work of writing it up will help you remember the mistake next time. (If you don't want to spend that amount of time thinking about why you were wrong, then perhaps you aren't very interested in really updating on this bet.)
- People usually make small-cash bets anyway, so there's not that much skin in the game. Being forced to write publically, or to a select group of peers such as a slack/discord server, makes it feel real for me in a way that losing a small sum of money doesn't.
- Where normal bets may benefit the participants, these kinds of public bets have more benefit for the whole audience. Observers get a lot more information about the structure of the disagreement, and the update which the loser takes from it.
- Often, by the time a bet is decided, a lot of other relevant information has come in as well. A public post-mortem gives the loser a chance to convey this information.
- This kind of bet will often be positive-sum in reputational terms: the winner gets a public endorsement from the loser, but the loser may gain respect from the audience for their gracious defeat and judicious update.

Furthermore, if the loser's write-up is anything short of honest praise for the winner's views, the write-up may provide hints at a continuing disagreement between the loser and winner which can lead to another bet.

This idea feels similar to Ben's Share Models, Not Beliefs. Bets focus only on disagreement with probabilities, not the underlying reasons for those disagreements. Declaring a winner/loser conveys binary information about who was more correct, but this is very little information. Post-mortems give a place for the models to be brought to light.

A group of people who engaged in betting-with-post-mortems together would generally be getting a lot more feedback on practical reasoning and where it can go wrong.

Don't Make Your Problems Hide

I've seen a worrying trend in people who've learned introspection and self-improvement methods from CFAR, or analogous ones from CBT. They make better life decisions, they calm their emotions in the moment. But they still look just as stressed as ever. They stamp out every internal conflict they can see, but it seems like there are more of them beyond the horizon of their self-awareness.

(I may have experienced this myself.)

One reason for this is that there's a danger with learning how to consciously notice and interact with one's subconscious thoughts/feelings/desires/fears: the conscious mind may not like what it sees, and try to edit the subconscious mind into one that pleases it.

The conscious mind might *try*, that is, but the subconscious is stronger. So, what actually happens?

The subconscious develops defense mechanisms.

Suppressed desires disguise themselves as being about other things, or they just overwhelm the conscious mind's willpower every now and then (and maybe fulfill themselves in a less healthy way than could otherwise be managed).

Suppressed thoughts become stealthy biases; certain conscious ideas or narratives get reinforced until they are practically unquestionable. So too with fears; a suppressed social fear is a good way to get a loud alarm that never stops.

Suppressed feelings hide themselves more thoroughly from the searchlight, so that one never consciously notices their meaning anymore, one just feels sad or angry or scared "for no reason" in certain situations.

At its worst, the conscious mind tries ever-harder to push back against these, further burning its rapport with the subconscious. I think of pastors who suppress their gay desires so hard that they vigorously denounce homosexuality and then sneak out for gay sex. They'd have been living such a happier life if they'd given up and acknowledged who they are, and what they want, years ago.

Now, sometimes people do have a strong desire that can't be satisfied in any healthy way. And that's just a brutal kind of life to life. But they would still do better by acknowledging that desire openly to themselves, than by trying to quash it and only hiding it.

How can we become more integrated between conscious and unconscious parts, and undo any damage we've already caused?

In my talk about the elephant and rider, I suggested (or gestured at) a few relevant things:

- Pursue basic happiness alongside your conscious goals (and make sure that's happiness *for you*, not just e.g. keeping your friends happy by doing the things

they like)
- Use positive reinforcement on yourself rather than punishment - it's especially important not to punish yourself for noticing the "wrong" thoughts/feelings/desires/fears. Reward the noticing, even with just an internal "thank you for surfacing this".
- Treat the content of these thoughts/feelings/desires/fears with respect. You might think of them as a friend opening up to you, and imagine the compassion you'd have when trying to figure out a way forward where both of you can flourish.

It's important to be gentle, to be curious, and to be patient. You don't have to resolve the whole thing; just acknowledging it respectfully can help the relationship grow.

There are other approaches too. Many people believe in using meditation to better integrate their thoughts and feelings and desires, for instance.

When you do something that you thought you didn't want to do, or when you're noticing an unexpected feeling, it's an opportunity for you. Don't push it away.

Half-Baked Products and Idea Kernels

When I ask someone at work for a project proposal, I never want the person to go silent on me and put in 100 hours of solitary work, and then finally show me something and ask for my feedback. I always want to see a **half-baked product**.

You can half-bake something in an hour or two, or even in a few minutes.

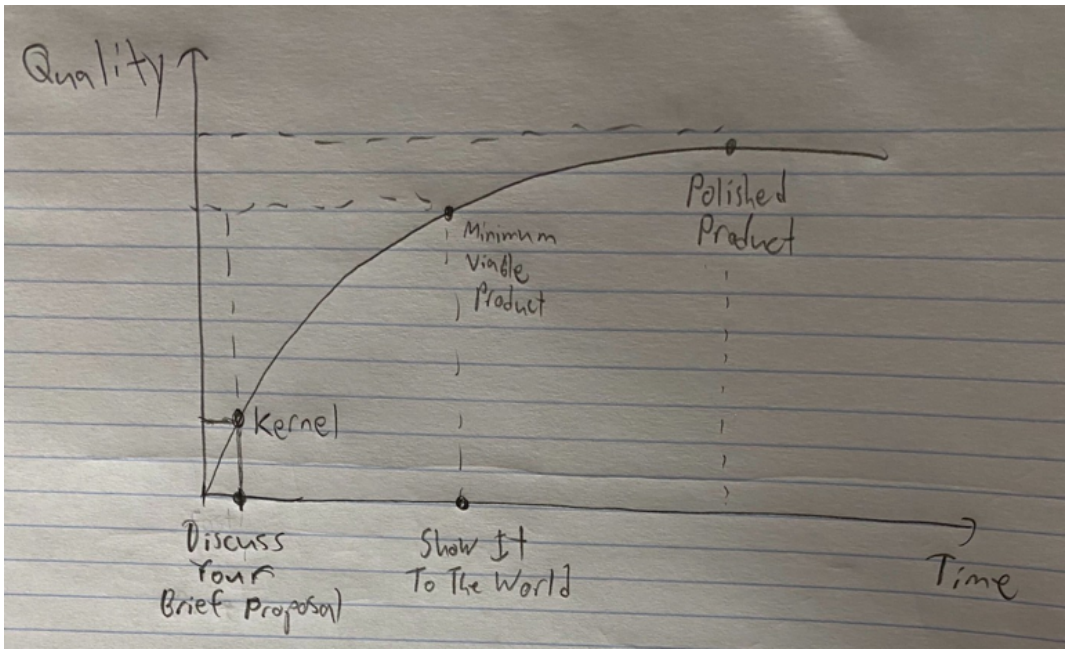
The advantage of half-baking is that you get a quick feedback loop. The more you think there's a possibility that I'll say "no, that's not what I wanted", the more half-baked you should make your first effort before getting my feedback.

When brainstorming ideas, my term for a half-baked idea is a **kernel**. A kernel is usually a crappy idea on its own, but there's "something to it" that could make it the seed of a better idea. I encourage people to toss out kernels.

There are two reasons why operating this way is efficient:

Diminishing Returns on Time Spent

Say you work on something for 100 hours. While each hour adds value, typically the highest-value hour is the first hour and the lowest-value hour is the last hour, and it follows a curve like this:

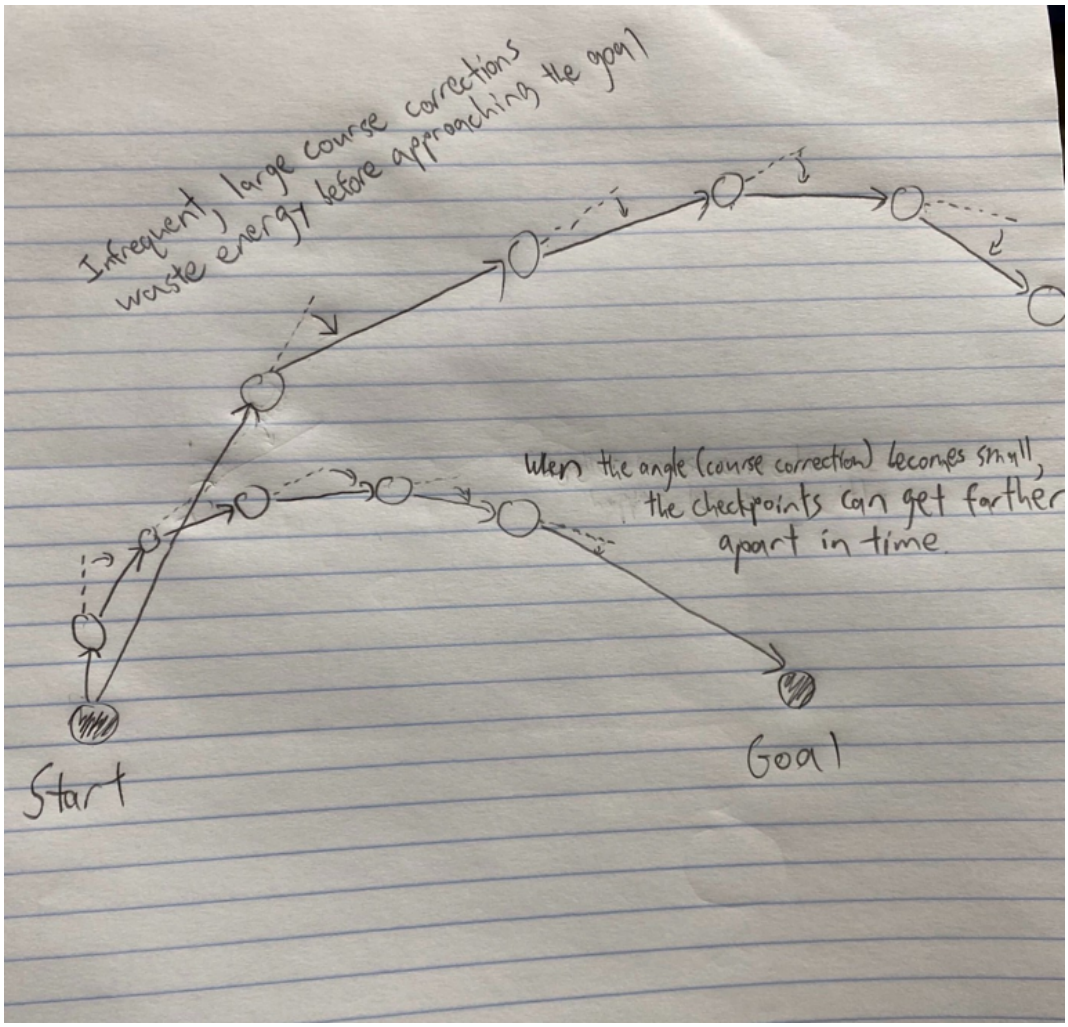


For example, if you're going to spend 100 hours writing a long report, spending one hour to brain-dump the key bullet points would give a reader a lot more than 1% of the final value of your report. Realistically, it’s probably more like 20%.

So the less time you spend working before getting feedback, the higher your productivity was in that time.

Efficient Course Correction

When you’re starting out on a new project that isn’t well-understood, you’re unlikely to go in the exact right direction. So you don’t want to go too far before getting a course correction, or you’ll waste time.



The top path shows how most people waste their time by investing too much effort between course corrections. The bottom path shows the efficient approach: you do a small chunk of work, then get feedback from your boss or your customers to correct your course, then do the next small chunk of work.

Once your course corrections become small, you can do larger chunks of work between course corrections. Until then, take small steps that produce half-baked products and idea kernels.

My weekly review habit

Every Saturday morning, I take 3-4 hours to think about how my week went and how I'll make the next one better.

The format has changed over time, but for example, here's some of what I reflected on last week:

- I noticed I'd fallen far short of my goal for written output. I decided to allocate more time to reading this week, hoping that it would generate more ideas. And I reorganized my morning routine to make it easier to start writing in the morning.
- I looked at some stats from RescueTime and Complice about what I'd spent time on and accomplished. I noticed that my time spent on Slack was nearing dangerous levels, so I decided to make a couple experimental tweaks to get it down:
    - I tried out Contexts, a replacement for the macOS window switcher, which I configured to only show windows from my current workspace—hoping that this would prevent me from cmd+tabbing over to Slack and getting distracted.
    - I decided to run an experiment of not answering immediately when coworkers called me in the middle of a focused block of time, and keeping a paper "todo when done focusing" list to remind myself to call them back, check Slack, etc.
- I noticed that it felt hard for me to get useful info from the time-tracking data in RescueTime and Complice, so I revisited what questions I actually wanted to answer and how I could make them easy to answer.
    - I realized that I should be using Google Calendar, not RescueTime or Complice, to track my time spent in meetings, so I added that to my time-tracking data sources.
    - I also made several tweaks to the way I used Complice to make it easier to see various stats I was interested in.

And so on. By the end of the review I had surfaced lots of other improvements for the coming week.

While each individual tweak is small, over the weeks and years they've compounded to make me a lot more effective. Because of that, this weekly review is the most useful habit (or habit-generating meta-habit) I've built. Here are some of the improvements I've made that have come out of weekly reviews:

- I decided to experiment with time tracking, realized that it needed to be zero-effort to succeed, and identified RescueTime as the best option, giving me much better data on how I was *actually* spending my time.
- Once I started using RescueTime, I eliminated a bunch of distractions that it flagged. As a result I improved my focused time by 50%.

- Later on, reviewing RescueTime stats also helped me realized I was spending much more time distracted by the Internet than I'd realized. I tried various things to break my Internet news habit and eventually found Focus, a zero-effort website blocker which has probably saved me hundreds of hours.
- I identified "feeling low-energy on winter evenings" as a blocker and tried several experiments to improve my evening energy levels. One of them was an ultra-bright lightbulb which worked amazingly well, giving me back about an hour a day.
- By thinking about how to help my partner with her PhD, I came up with the idea of doing one-on-ones, which she thinks helped her finish her dissertation a year faster.
- By thinking about ways to improve our relationship and points of friction we've had over the last week, we've both started lots of useful discussions in those one-on-ones that have helped us understand each other and communicate better.
- I've made hundreds of tweaks to my daily routine and habits to make sure I reliably exercise, sleep enough, and maintain high energy levels.

Of course, you don't *need* to have a weekly review habit to come up with this type of improvement. But by systematically thinking it through, you'll generate more of them. And by doing it consistently, you'll be able to build these small improvements on top of each other.

I've had to iterate a lot on the format and timing of the weekly review to get to one where I can consistently maintain the habit and output useful weekly reviews. The format I currently have is:

- Review happens first thing every Saturday morning. This time is sacred and (largely) immovable. Morning is really important for having the right energy and mindset; weekend is important so I'm not distracted by work; consistency is important so that I don't lose the habit.
- I start the review by re-reading some parts of my favorite essays of life advice. (Different parts/essays every week. This also sometimes gets me to notice new parts of the essays that resonate or spark interesting thoughts.)
- Next, I load the week back into working memory by reviewing what happened during the week.
- Based on the above I'll write down a list of topics to think about, taking written notes on each topic as I think about them.
- I also have a set of recurring prompts that I think about every week. I tweak them over time as they get stale, but some examples would be:
    - Was I consistent at my core habits this week (exercise, morning routine, todo system, etc.)? How can I tweak them to be more consistent or more useful?

    - What did I do this week that was a mistake and how can I avoid repeating it?
    - How much of this week did I spend on stuff that was truly my comparative advantage? For everything else, how can I get out of the loop?

As an appendix, some random tactical tips for weekly reviews:

- Changing my physical environment helps me context-switch into a less focused, more reflective mindset. Back when cafes were open I'd often go to the cafe near my house. At home, I'll work from a different room, play different background music, etc.
- I still find it's easy to get distracted during weekly reviews, so I make sure to close everything else on my computer when I start.
- When I have granular, objective data on "what happened this week" (e.g. RescueTime, calendar, todo lists) I've found it helpful to review that because it occasionally surprises me. (See the points about RescueTime above.)
- I find that taking notes while I think about things is really important—otherwise I lose track of what I'm thinking about or get distracted.
- For note-taking, I'd recommend using hierarchical bulleted lists, not free-written paragraphs. Lists are more efficient because you can write in incomplete sentences and leave out transitions (relying on the bullet hierarchy to make the structure clear).

  Bulleted lists are also easier to reorder (especially if you use an app that gives you keyboard shortcuts for it), so if you're like me, they'll let you more efficiently exercise your nervous tic of stack-ranking all lists.

Plausible cases for HRAD work, and locating the crux in the "realism about rationality" debate

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post is my attempt to summarize and distill the major public debates about MIRI's highly reliable agent designs (HRAD) work (which includes work on decision theory), including the discussions in Realism about rationality and Daniel Dewey's My current thoughts on MIRI's "highly reliable agent design" work. Part of the difficulty with discussing the value of HRAD work is that it's not even clear what the disagreement is about, so my summary takes the form of multiple possible "worlds" we might be in; each world consists of a positive case for doing HRAD work, along with the potential objections to that case, which results in one or more cruxes.

I will talk about "being in a world" throughout this post. What I mean by this is the following: If we are "in world X", that means that the case for HRAD work outlined in world X is the one that most resonates with MIRI people as their motivation for doing HRAD work; and that when people disagree about the value of HRAD work, this is what the disagreement is about. When I say that "I think we are in this world", I don't mean that I agree with this case for HRAD work; it just means that this is what I think MIRI people think.

In this post, the pro-HRAD stance is something like "HRAD work is the most important kind of technical research in AI alignment; it is the overwhelming priority and we're pretty much screwed if we under-invest in this kind of research" and the anti-HRAD stance is something like "HRAD work seems significantly less promising than other technical AI alignment agendas, such as the approaches to directly align machine learning systems (e.g. iterated amplification)". There is a much weaker pro-HRAD stance, which is something like "HRAD work is interesting and doing more of it adds value, but it's not necessarily the most important kind of technical AI alignment research to be working on"; this post is not about this weaker stance.

Clarifying some terms

Before describing the various worlds, I want to present some distinctions that have come up in discussions about HRAD, which will be relevant when distinguishing between the worlds.

Levels of abstraction vs levels of indirection

The idea of levels of abstraction was introduced in the context of debate about HRAD work by Rohin Shah, and is described in this comment (start from "When groups of humans try to build complicated stuff"). For more background, see these articles on Wikipedia.

Later on, in this comment Rohin gave a somewhat different "levels" idea, which I've decided to call "levels of indirection". The idea is that there might not be a *hierarchy of abstraction*, but there's still multiple intermediate layers between the theory you have and the end-result you want. The relevant "levels of indirection" is the sequence HRAD → machine learning → AGI. Even though levels of indirection are different from levels of abstraction, the idea is that the same principle applies, where the more levels there are, the harder it becomes for a theory to apply to the final level.

Precise vs imprecise theory

A precise theory is one which can scale to 2+ levels of abstraction/indirection.

An imprecise theory is one which can scale to at most 1 level of abstraction/indirection.

More intuitively, a precise theory is more mathy, rigorous, and exact like pure math and physics, and an imprecise theory is less mathy, like economics and psychology.

Building agents from the ground up vs understanding the behavior of rational agents and predicting roughly what they will do

This distinction comes from Abram Demski's comment. However, I'm not confident I've understood this distinction in the way that Abram intended it, so what I describe below may be a slightly different distinction.

Building agents from the ground up means having a precise theory of rationality that allows us to build an AGI in a satisfying way, e.g. where someone with security mindset can be confident that it is aligned. Importantly, we allow the AGI to be built using whatever way is safest or most theoretically satisfying, rather than requiring that the AGI be built using whatever methods are mainstream (e.g. current machine learning methods).

Understanding the behavior of rational agents and predicting roughly what they will do means being handed an arbitrary agent implemented in some way (e.g. via blackbox ML) and then being able to predict roughly how it will act.

I think of the difference between these two as the difference between existential and universal quantification: "there exists x such that P(x)" and "for all x we have P(x)", where P(x) is something like "we can understand and predict how x will act in a satisfying way". The former only says that we can build some AGI using the precise theory that we understand well, whereas the latter says we have to deal with whatever kind of AGI that ends up being developed using methods we might not understand well.

World 1

Case for HRAD

The goal of HRAD research is to generally become less confused about things like counterfactual reasoning and logical uncertainty. Becoming less confused about these

things will: help AGI builders avoid, detect, and fix safety issues; help AGI builders predict or explain safety issues; help to conceptually clarify the AI alignment problem; and help us be satisfied that the AGI is doing what we want. Moreover, unless we become less confused about these things, we are likely to screw up alignment because we won't deeply understand how our AI systems are reasoning. There are other ways to gain clarity on alignment, such as by working on iterated amplification, but these approaches don't decompose cognitive work enough.

For this case, it is not important for the final product of HRAD to be a precise theory. Even if the final theory of embedded agency is imprecise, or even if *there is no* "final say" on the topic, if we are merely much less confused than we are now, that is still good enough to help us ensure AI systems are aligned.

Why I think we might be in this world

The main reason I think we might be in this world (i.e. that the above case is the motivating reason for MIRI prioritizing HRAD work) is that people at MIRI frequently seem to be saying something like the case above. However, they also seem to be saying different things in other places, so I'm not confident this is actually their case. Here are some examples:

- Eliezer Yudkowsky: "Techniques you can actually adapt in a safe AI, come the day, will probably have very simple cores — the sort of core concept that takes up three paragraphs, where any reviewer who didn't spend five years struggling on the problem themselves will think, "Oh I could have thought of that." Someday there may be a book full of clever and difficult things to say about the simple core — contrast the simplicity of the core concept of causal models, versus the complexity of proving all the clever things Judea Pearl had to say about causal models. But the planetary benefit is mainly from posing understandable problems crisply enough so that people can see they are open, and then from the simpler abstract properties of a found solution — complicated aspects will not carry over to real AIs later."
- Rob Bensinger: "We're working on decision theory because there's a cluster of confusing issues here (e.g., counterfactuals, updatelessness, coordination) that represent a lot of holes or anomalies in our current best understanding of what high-quality reasoning is and how it works." and phrases like "developing an understanding of roughly what counterfactuals are and how they work" and "very roughly how/why it works" -- This post then doesn't really specify whether or not the final output is expected to be precise. (The analogy with probability theory and rockets gestures at precise theories, but the post doesn't come out and say it.)
- Abram Demski: "I don't think there's a true rationality out there in the world, or a true decision theory out there in the world, or even a true notion of intelligence out there in the world. I work on agent foundations because there's *still something I'm confused about* even after that, and furthermore, AI safety work seems fairly hopeless while still so radically confused about the-phenomena-which-we-use-intelligence-and-rationality-and-agency-and-decision-theory-to-describe."
- Nate Soares: "The main case for HRAD problems is that we expect them to help in a gestalt way with many different known failure modes (and, plausibly, unknown ones). E.g., 'developing a basic understanding of counterfactual reasoning improves our ability to understand the first AGI systems in a general way, and if we understand AGI better it's likelier we can build systems to

address deception, edge instantiation, goal instability, and a number of other problems'."
- In the deconfusion section of MIRI's 2018 update, some of the examples of deconfusion are not precise/mathematical in nature (e.g. see the paragraph starting with "In 1998, conversations about AI risk and technological singularity scenarios often went in circles in a funny sort of way" and the list after "Among the bits of conceptual progress that MIRI contributed to are"). There are more mathematical examples in the post, but the fact that there are also non-mathematical examples suggests that having a precise theory of rationality is not important to the case for HRAD work. There's also the quote "As AI researchers explore the space of optimizers, what will it take to ensure that the first highly capable optimizers that researchers find are optimizers they know how to aim at chosen tasks? I'm not sure, because I'm still in some sense confused about the question."

The crux

One way to reject this case for HRAD work is by saying that imprecise theories of rationality are insufficient for helping to align AI systems. This is what Rohin does in this comment where he says imprecise theories cannot build things "2+ levels above".

There is a separate potential rejection, which is to say that either HRAD work will never result in precise theories or that even a precise theory is insufficient for helping to align AI systems. However, these move the crux to a place where they apply to more restricted worlds where the goal of HRAD work is specifically to come up with a precise theory, so these will be covered in the other worlds below.

There is a third rejection, which is to argue that other approaches (such as iterated amplification) are more promising for gaining clarity on alignment. In this case, the main disagreement may instead be about other agendas rather than about HRAD.

World 2

Case for HRAD

The goal of HRAD research is to come up with a theory of rationality that is so precise that it allows one to build an agent from the ground up. Deconfusion is still important, as with world 1, but in this case we don't merely want any kind of deconfusion, but specifically deconfusion which is accompanied by a precise theory of rationality.

For this case, HRAD research isn't intended to produce a precise theory about how to predict ML systems, or to be able to make precise predictions about what ML systems will do. Instead, the idea is that the precise theory of rationality will help AGI builders avoid, detect, and fix safety issues; predict or explain safety issues; help to conceptually clarify the AI alignment problem; and help us be satisfied that the AGI is doing what we want. In other words, instead of directly using a precise theory about understanding/predicting rational agents in general, we use the precise theory about rationality to help us *roughly* predict what rational agents will do in general (including ML systems).

As with world 1, unless we become less confused, we are likely to screw up alignment because we won't deeply understand how our AI systems are reasoning. There are

other ways to gain clarity on alignment, such as by working on iterated amplification, but these approaches don't decompose cognitive work enough.

Why I think we might be in this world

This seems to be what Abram is saying in this comment (see especially the part after "I guess there's a tricky interpretational issue here").

It also seems to match what Rohin is saying in these two comments.

The examples MIRI people sometimes give for precedents of HRAD-ish work, like the work done by Turing, Shannon, and Maxwell are precise mathematical theories.

The crux

There seem to be two possible rejections of this case:

- We can reject the existence of the precise theory of rationality. This is what Rohin does in this comment and this comment where he says "MIRI's theories will always be the relatively-imprecise theories that can't scale to '2+ levels above'." Paul Christiano seems to also do this, as summarized by Jessica Taylor in this post: intuition 18 is "There are reasons to expect the details of reasoning well to be 'messy'."
- We can argue that even a precise theory of rationality is insufficient for helping to align AI systems. This seems to be what Daniel Dewey is doing in this post when he says things like "AIXI and Solomonoff induction are particularly strong examples of work that is very close to HRAD, but don't seem to have been applicable to real AI systems" and "It seems plausible that the kinds of axiomatic descriptions that HRAD work could produce would be too taxing to be usefully applied to any practical AI system".

World 3

Case for HRAD

The goal of HRAD research is to directly come up with a precise theory for understanding the behavior of rational agents and predicting what they will do. Deconfusion is still important, as with worlds 1 and 2, but in this case we don't merely want any kind of deconfusion, but specifically deconfusion which is accompanied by a precise theory that allows us to predict agents' behavior in general. And a precise theory is important, but we don't merely want a precise theory that lets us *build* an agent; we want our theory to act like a box that takes in an arbitrary agent (such as one built using ML and other black boxes) and allows us to analyze its behavior.

This theory can then be used to help AGI builders avoid, detect, and fix safety issues; predict or explain safety issues; help to conceptually clarify the AI alignment problem; and help us be satisfied that the AGI is doing what we want.

As with world 1 and 2, unless we become less confused, we are likely to screw up alignment because we won't deeply understand how our AI systems are reasoning.

There are other ways to gain clarity on alignment, such as by working on iterated amplification, but these approaches don't decompose cognitive work enough.

Why I think we might be in this world

I mostly don't think we're in this world, but some critics might think we are.

For example Abram says in this comment: "I can see how Ricraz would read statements of the first type [i.e. having precise understanding of rationality] as suggesting very strong claims of the second type [i.e. being able to understand the behavior of agents in general]."

Daniel Dewey might also expect to be in this world; it's hard for me to tell based on his post about HRAD.

The crux

The crux in this world is basically the same as the first rejection for world 2: we can reject the existence of a precise theory for understanding the behavior of arbitrary rational agents.

Conclusion, and moving forward

To summarize the above, combining all of possible worlds, the pro-HRAD stance becomes:

```
(ML safety agenda not promising) and (
  (even an imprecise theory of rationality helps to align AGI) or
  ((a precise theory of rationality can be found) and
   (a precise theory of rationality can be used to help align AGI)) or
  (a precise theory to predict behavior of arbitrary agent can be found)
)
```

and the anti-HRAD stance is the negation of the above:

```
(ML safety agenda promising) or (
  (an imprecise theory of rationality cannot be used to help align AGI) and
  ((a precise theory of rationality cannot be found) or
   (even a precise theory of rationality cannot be used to help align AGI)) and
  (a precise theory to predict behavior of arbitrary agent cannot be found)
)
```

How does this fit under the Double Crux framework? The current "overall crux" is a messy proposition consisting of multiple conjunctions and disjunctions, and fully resolving the disagreement can in the worst case require assigning truth values to all five parts: the statement "A and (B or (C and D) or E)", with disagreements resolved in the order A=True, B=False, C=True, D=False can still be true or false depending on the value of E. From an efficiency perspective, if some of the conjunctions/disjunctions don't matter, we want to get rid of them in order to simplify the structure of the overall crux (this corresponds to identifying which "world" we are in, using the terminology of this post), and we also might want to pick an ordering of which parts to resolve first (for example, with A=True and B=True, we already know the overall proposition is true).

So some steps for moving the discussion forward:

- I think it would be great to get HRAD proponents/opponents to be like "we're definitely in world X, and not any of the other worlds" or even be like "actually, the case for HRAD really is disjunctive, so both of the cases in worlds X and Y apply".
- If I missed any additional possible worlds, or if I described one of the worlds incorrectly, I am interested in hearing about it.
- If it becomes clear which world we are in, then the next step is to drill down on the crux(es) in that world.

Thanks to Ben Cottier, Rohin Shah, and Joe Bernstein for feedback on this post.

Inaccessible information

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Suppose that I have a great model for predicting "what will Alice say next?"

I can evaluate and train this model by checking its predictions against reality, but there may be many facts this model "knows" that I can't easily access.

For example, the model might have a detailed representation of Alice's thoughts which it uses to predict what Alice will say, *without* being able to directly answer "What is Alice thinking?" In this case, I can only access that knowledge indirectly, e.g. by asking about what Alice would say in under different conditions.

I'll call information like "What is Alice thinking?" inaccessible. I think it's very plausible that AI systems will build up important inaccessible knowledge, and that this may be a central feature of the AI alignment problem.

In this post I'm going to try to clarify what I mean by "inaccessible information" and the conditions under which it could be a problem. This is intended as clarification and framing rather than a presentation of new ideas, though sections IV, V, and VI do try to make some small steps forward.

I. Defining inaccessible information

I'll start by informally defining what it means for information to be **accessible**, based on two mechanisms:

Mechanism 1: checking directly

If I can check X myself, *given other accessible information,* then I'll define X to be accessible.

For example, I can check a claim about what Alice will do, but I can't check a claim about what Alice is thinking.

If I can run randomized experiments, I can probabilistically check a claim about what Alice *would* do. But I can't check a counterfactual claim for conditions that I can't create in an experiment.

In reality this is a graded notion—some things are easier or harder to check. For the purpose of this post, we can just talk about whether something can be tested even a single time over the course of my training process.

Mechanism 2: transfer

The simplest model that provides some accessible information X may also provide some other information Y. After all, it's unlikely that the simplest model that outputs X doesn't output *anything* else. In this case, we'll define Y to be accessible.

For example, if I train a model to predict what happens over the next minute, hour, or day, it may generalize to predicting what will happen in a month or year. For example, if the simplest model to predict the next day was a fully-accurate physical simulation, then the same physics simulation might work when run for longer periods of time.

I think this kind of transfer is kind of dicey, so I genuinely don't know if long-term predictions are accessible or not (we certainly can't directly check them, so transfer is the only way they

could be accessible).

Regardless of whether long-term predictions are accessible by transfer, there are other cases where I think transfer is pretty unlikely. For example, the simplest way to predict Alice's behavior might be to have a good working model for her thoughts. But it seems unlikely that this model would spontaneously describe what Alice is thinking in an understandable way—you'd need to specify some additional machinery, for turning the latent model into useful descriptions.

I think this is going to be a fairly common situation: predicting accessible information may involve almost all the same work as predicting inaccessible information, but you need to combine that work with some "last mile" in order to actually output inaccessible facts.

Definition

I'll say that information is *accessible* if it's in the smallest set of information that is closed under those two mechanisms, and *inaccessible* otherwise.

There are a lot of nuances in that definition, which I'll ignore for now.

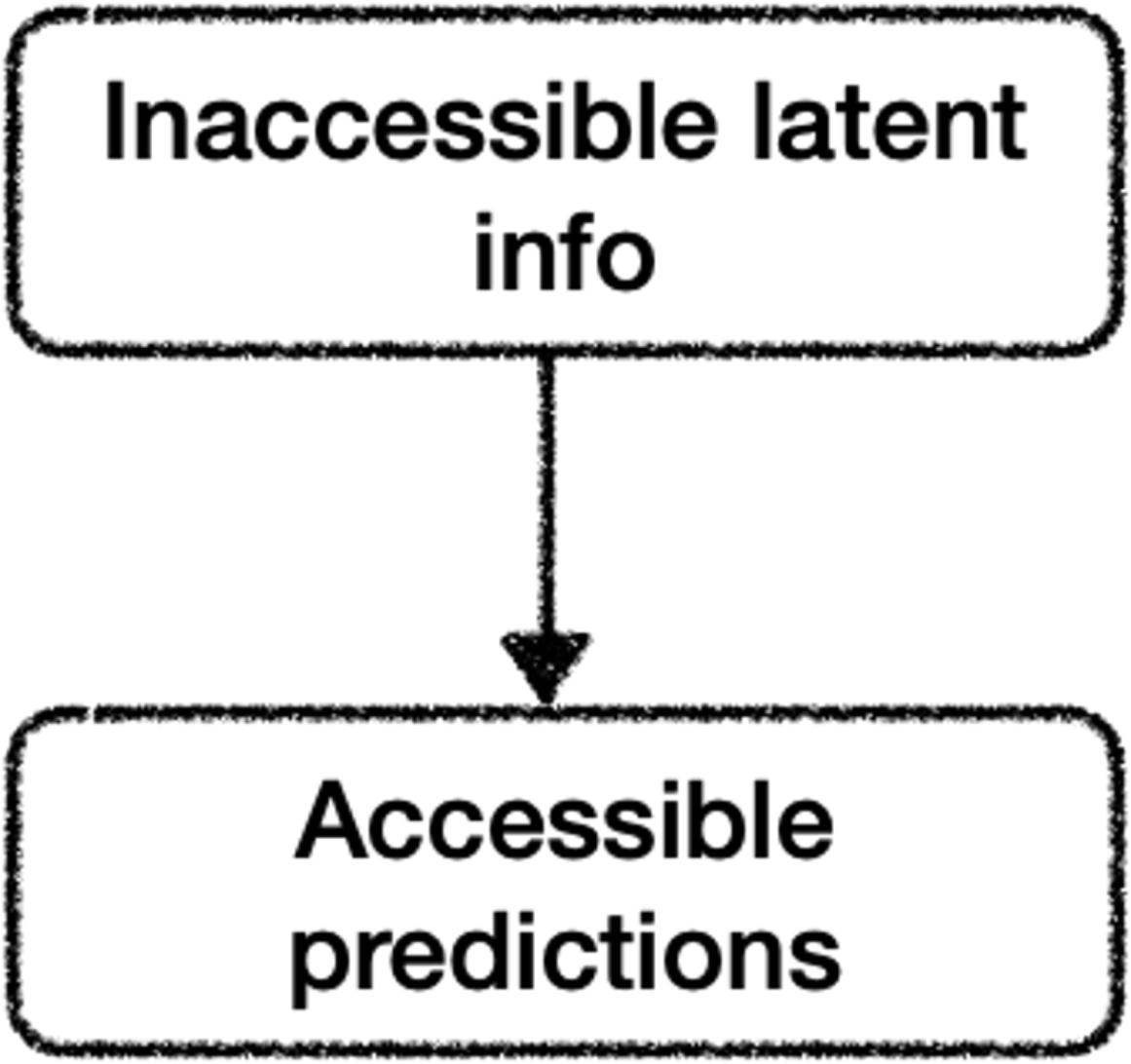
Examples

Here are some candidates for accessible vs. inaccessible information:

- "What will Alice say?" vs "What is Alice thinking?"
- "What's on my financial statement?" vs. "How much money do I really have?"
- "Am I coughing?" vs. "What's happening with my immune system?"
- "How will senators vote?" vs. "What's the state of political alliances and agreements in the senate?"
- "What do I see on my computer screen?" vs. "Is my computer compromised?"
- "What's the market price of this company?" vs. "How valuable is this IP really?"
- "Will the machine break tomorrow?" vs. "Is there hard-to-observe damage in this component?"
- "What does the news show me from 5000 miles away?" vs. "What's actually happening 5000 miles away?"
- "Is this argument convincing?" vs. "Is this argument correct?"
- "What will happen tomorrow?" vs. "What will happen in a year" (depending on whether models transfer to long horizons)

II. Where inaccessible info comes from and why it might matter

Our models can build up inaccessible information because it helps them predict accessible information. They know something about what Alice is thinking because it helps explain what Alice does. In this diagram, the black arrow represents the causal relationship:



Unfortunately, this causal relationship doesn't directly let us *elicit* the inaccessible information.

Scientific theories are prototypical instances of this diagram, e.g. I might infer the existence of electron from observing the behavior of macroscopic objects. There might not be any explanation for a theory other than "it's made good predictions in the past, so it probably will in the future." The actual claims the theory makes about the world—e.g. that the Higgs boson has such-and-such a mass—are totally alien to someone who doesn't know anything about the theory.

I'm not worried about scientific hypotheses in particular, because they are usually *extremely* simple. I'm much more scared of analogous situations that we think of as intuition—if you want to justify your intuition that Alice doesn't like you, or that some code is going to be hard to maintain, or that one tower of cards is going to be more stable than another, you may not be able to say very much other than "This is part of a complex group of intuitions that I built up over a very long time and which seems to have a good predictive track record."

At that point "picking the model that matches the data best" starts to look a lot like doing ML, and it's more plausible that we're going to start getting hypotheses that we don't understand or which behave badly.

Why might we care about this?

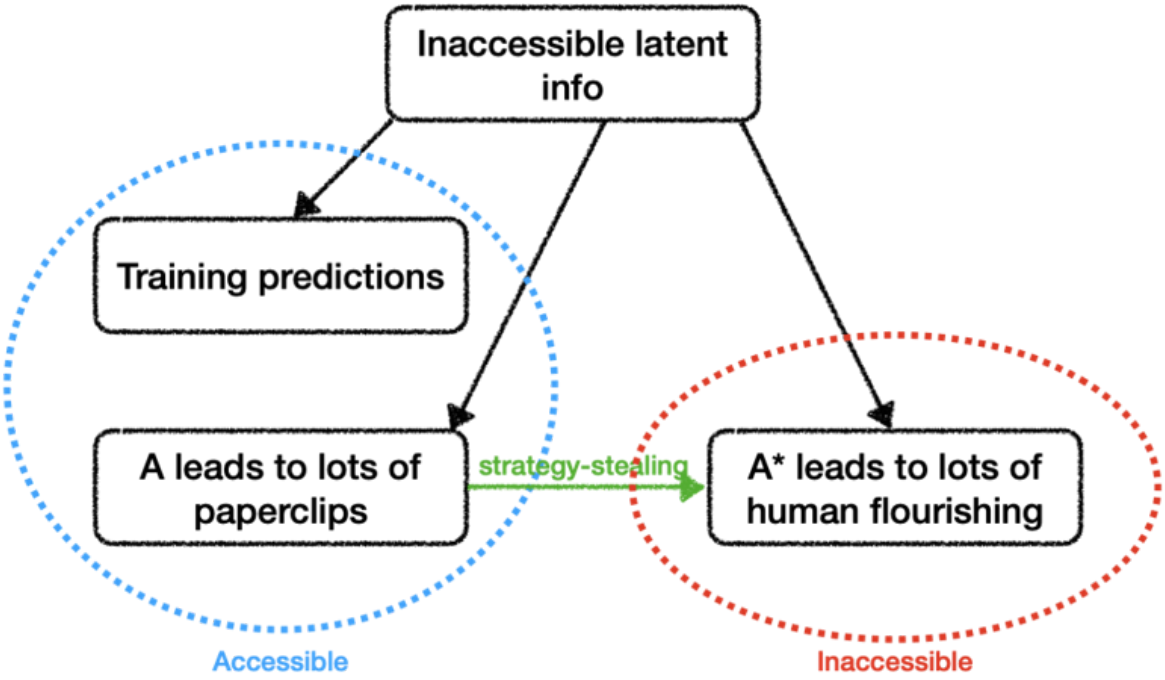
In some sense, I think this all comes down to what I've called strategy-stealing: if AI can be used to compete effectively, can humans use AI to compete *on their behalf*?

More precisely, for every strategy A that an AI could pursue to bring about some arbitrary outcome, is there a strategy A* that would help humans get what we want over the long term, without leaving us at a competitive disadvantage over the short term?

If so it's good news for humanity: if most humans build AIs who execute plans like A*, then humans won't be outcompeted by unaligned AIs who execute plans like A.

But the mere *existence* of A* isn't very helpful, we need to actually be able to figure out that A* leads to human flourishing so that we can do it. If we can't recognize plans like A*, then humanity will be at a disadvantage.

We could have a problem if the fact "A* leads to human flourishing" is inaccessible while the fact "A leads to paperclips" is accessible.



In this case, we either need to be sure that no one acts on the basis of information like "A leads to lots of paperclips," or we need to figure out some way to access the inaccessible information that "A* leads to lots of human flourishing."

III. Eliciting inaccessible info and the instrumental policy

I've been talking as if we can't access inaccessible info (and the name is suggestive…), but I don't think you should be convinced of that yet.

Let M be a model that reports only accessible facts—it answers questions like “what will Alice say next?” but not questions like “is Alice actually happy with the situation?” It’s built up a lot of knowledge about the world, and uses that knowledge to estimate accessible facts well as it can.

We suspect there is a related model M* that uses all of M’s machinery in order to explain what’s going on in Alice’s head. By the definition of inaccessibility, M* is at least somewhat more complex than M. So if we want to learn M*, we need a training objective that penalizes M relative to M*.

But it’s really easy to tell M* and M apart! M won’t automatically tell us what Alice is thinking, but it also won’t automatically make up credible stories that hold together. Both of those require extra complexity. So we can still hope to get M* if we train our model to (i) correctly provides accessible info, (ii) says something consistent and plausible about inaccessible info.

The real problem comes from what I’ll call the *instrumental policy*. Let’s say we’ve tried to dream up a loss function L(x, y) to incentivize the model to correctly answer information we can check, and gives at least plausible and consistent answers on things we can’t check. By definition, the values L(x, y) are themselves accessible.

Then it’s natural to learn a policy like: “on input x, produce the output y for which the loss L(x, y) will be minimal.” Let’s write BAD for this policy.

Put differently, BAD is the model that performs a simple substitution—when you ask it question Q, it replaces it with the meta question “What answer to question Q will make the person training me happiest?” By definition, this new question is also accessible, so it can be posed to M.

BAD may be much simpler than M*, because all it has to do is describe this substitution, whereas M* needs to build new machinery to explain inaccessible facts.

If M is very weak, BAD may perform much worse than M* because it doesn’t understand the loss function very well. But if M understands a lot about the world, and about its training process, then it seems like BAD will get at least as low a loss as M*—it’s using the same machinery as M and M*, and picking whatever value leads to the lowest loss. It can choose to report its own best estimates if that’s the best strategy, e.g. for questions where we can check. So by the time M and M* are human level, it already seems like BAD should be getting as low a loss as M*.

This gets a little bit more complicated if I do consistency checks or if I have large output spaces, and of course humans aren’t *that* good at lying, but it seems like the basic idea is pretty solid: once we have a model of accessible information, we can directly turn it into a model that will pass whatever checks we throw at it.

For concreteness, imagine M is a model of Alice’s behavior and we are asking it “is Alice actually happy?” Then M* is the model which translates its model of Alice’s psychology into words and explains it to us, while BAD is the model that asks “What answer would sound most believable to the person asking the question?”

The question is: if we start with model M and fine-tune it to answer inaccessible questions, is it easier to learn to articulate what it believes or is it going to learn to answer strategically?

We need to be *very* careful about our checks if we want to ensure that BAD doesn’t get a lower loss than M*. For example, it may turn out that sometimes the truth looks a little bit wrong to us…. And if we do everything right, then M* and BAD perform equally well, and so we may not have much control over which one we get.

IV. When inaccessible info is a safety problem

Let's get a bit more detailed about the argument in section II. I think that our inability to access inaccessible info would become a safety problem when:

1. We care about inaccessible facts, so we can't just evaluate plans based on their accessible consequences.
2. Inaccessible info is a competitive advantage—agents who are blind to inaccessible facts about the world will get outcompeted.
3. There are *some* agents who are able to use inaccessible facts to acquire influence, e.g. because they are optimizing accessible long-term goals.

1. We care about inaccessible facts

If I only cared about accessible facts, then I might not need to ever access inaccessible facts. For example, if I cared about my life expectancy, and this was accessible, then I could ask my AI "what actions lead to me living the longest?" and execute those.

For better or worse, I think we are likely to care about inaccessible facts.

- Generally we care about what's *actually happening* and not just what appears to be happening. We don't want smiling faces on cameras. And if there's a lot of inaccessible action in the world, then it's reasonably likely for accessible indicators to be systematically manipulated by inaccessible forces.
- We care intrinsically about what happens inside people's heads (and inside computers), not just outward appearances. Over the very long term a *lot* may happen inside computers.
- If we totally give up on measuring how well things are going day-to-day, then we need to be actually optimizing the thing we really care about. But figuring that out may require reflecting a long time, and may be inaccessible to us now. We want a world where we actually reach the correct moral conclusions, not one where we believe we've reached the correct moral conclusions.
- Our real long-term priorities, and our society's long-term future, may also be really weird and hard to reason about even if we were able to know what was good. It just seems really bad to try to evaluate plans only by their very long-term consequences.
- We care about things that are far away in space or time, which I think are likely to be inaccessible.

Overall I'm quite skeptical about the strategy "pick an accessible quantity that captures everything you care about and optimize it." I think we basically need to optimize some kind of value function that tells us how well things are going. That brings us to the next section.

2. Inaccessible info is a competitive advantage

Instead of using AI to directly figure out whether a given action will lead to human flourishing over the coming centuries, we could use AI to help us figure out how to get what we want over the short term—including how to acquire resources and flexible influence, how to keep ourselves safe, and so on.

This doesn't require being able to tell how good a very long-term outcome is, but it does require being able to tell how well things are going. We need to be able to ask the AI "which plan would put us in an *actually good* position next year?"

Unfortunately, I think that if we can only ask about accessible quantities, we are going to end up neglecting a bunch of really important stuff about the situation, and we'll be at a significant competitive disadvantage compared to AIs which are able to take the whole picture into account.

As an intuition pump, imagine a company that is run entirely by A/B tests for metrics that can be easily checked. This company would burn every resource it couldn't measure—its code would become unmaintainable, its other infrastructure would crumble, it would use up goodwill with customers, it would make no research progress, it would become unable to hire, it would get on the wrong side of regulators…

My worry is that inaccessible facts will be similarly critical to running superhuman businesses, and that humans who rely on accessible proxies will get outcompeted just as quickly as the company that isn't able to optimize anything it can't A/B test.

- Even in areas like business that society tries particularly hard to make legible, evaluating how well you are doing depends on e.g. valuing intellectual property and intangible assets, understanding contractual relationships, making predictions about what kinds of knowledge or what relationships will be valuable, and so on.
- . In domains like social engineering, biology, cybersecurity, financial systems, *etc.*, I think inaccessible information becomes even more important.
- If there is a lot of critical inaccessible information, then it's not clear that a simple proxy like "how much money is actually in my bank account" is even accessible. The only thing that I can directly check is "what will I see when I look at my bank account statement?", but that statement could itself be meaningless. We really care about things like who effectively controls that bank account and what would really happen if I tried to spend the money. (And if I largely care about inaccessible facts about the world, then "what would happen if I tried to spend my money?" may itself be inaccessible.)
- I can pay inaccessible costs for an accessible gain—for example leaking critical information, or alienating an important ally, or going into debt, or making short-sighted tradeoffs. Moreover, if there are other actors in the world, they can try to get me to make bad tradeoffs by hiding real costs.

3. Some AIs can plan with inaccessible info

So far this discussion could just be about an *AI missed opportunity*, not an *AI risk*.

Things become problematic when it is possible to build AI systems that do use inaccessible info to pursue ambitious long-term goals that would conflict with human flourishing. If illegible knowledge is important enough, those systems could outcompete humans and divert some (or almost all) of our civilization's resources.

This happens if *any* interesting long-term goal is accessible, i.e. if there's any accessible goal that benefits from accumulating influence.

Why might some long-term goal be accessible?

- Verifiable long-term predictions may accessible, e.g. "total reward over the next 50 years" might be accessible as a generalization of "total reward over the next 50 minutes."
- Even if models don't transfer in a natural way to very long time horizons, they may still transfer in *some* way. Maybe they can't generalize to something as cleanly specified as "total reward," but e.g. they may be able to predict quantities that happen to be natural in terms of the representation they are using.
- More generally, models will do *something* when we run them way out of distribution. I think the best reason to expect models to not plan intelligently over long time horizons is that they lack the abilities and knowledge. Once they have relevant cognitive machinery, I think it's pretty likely that they'll use the machinery somehow when they go out of distribution, and if you search over a lot of possibilities.

The key asymmetry working against us is that optimizing flourishing appears to require a particular quantity to be accessible, while danger just requires *anything* to be accessible.

There's a longer discussion about whether anyone would want to build such AI systems, or might be likely to do it accidentally. And another discussion about how bad it would actually be. But I think that once we satisfy the 3 properties in this section then we have a real problem, and the remaining question is one of magnitude, with a bunch of factors pulling up and down.

V. Possible responses

At a high level, we'd like to get good answers by training a model to make checkable predictions and then somehow incentivizing it to use the same machinery to also report inaccessible facts, for example by forcing it to report views and enforcing consistency.

I think there are a bunch of problems with that plan, but the strongest one is that it's totally unclear how we ever get the intended policy rather than the instrumental policy.

Here are some possible angles of attack:

- In some intuitive sense there is a discontinuity between saying X because you believe it and saying X because you believe that it's what you need to say to get a high reward. If weak policies are straightforward rather than instrumental, then maybe we can exploit continuity to avoid the instrumental strategy.
  I'm generally pretty scared of these approaches since they feel pretty brittle, and people seem prone to being optimistic even before we can test or understand anything.
  But I'm a bit more hopeful if we consider continuity of behavior rather than continuity in the space of policies, and if we exploit it by using a model at time T to help check the behavior of a model at time T+1.
- If we were able to actually understand something about what the policy was doing, even crudely, it might let us discriminate between instrumental and intended behavior. I don't think we have any concrete proposals for how to understand what the policy is doing well enough to make this distinction, or how to integrate it into training. But I also don't think we have a clear sense of the obstructions, and I think there are various obvious obstructions to interpretability in general that *don't* apply to this approach. So together with the other points on this list—especially the existence of a discontinuity and the use of previous versions of a model to help—I think it could be part of a solution.
- The "instrumental" strategy—"pick the answer that minimizes my training loss"—may have a shorter description, but it seems to be unnecessarily computationally complex. First it figures out what's true, and then it strategically decides what to say in light of that. It would be a bit cheaper just to actually report what's true, if we set up the training process well enough that honest reporting got you optimal reward.
  It's not clear how to leverage this though, since "use the simplest model" seems like an important and useful prior. We can't just replace it with "use the cheapest model" without sacrificing competitiveness. That said, it may be able to buy us something in combination with the previous approaches.
  (See discussion here and here.)
- We aren't able to check inaccessible answers, and they aren't output by the simplest model, but maybe we can make being straightforward the easiest strategy in some other sense. For example, if you have two models who can't interact but need to agree on a consistent story, you could imagine the truth being the easiest thing for them to coordinate on. I think that scheme in particular is pretty unlikely to work, but there is a large space of possibilities here, especially when combined with the other points on this list. For example, maybe you could perform consistency checks between models at different points in time.
- Maybe we can work with models M that don't understand enough about the world, or especially about their training process, in order to turn into the instrumental policy BAD. I tend to be pretty skeptical of "security by obscurity"—if we want to make life

hard for the instrumental policy not working, I want it to be because we have randomness hidden from it, or we've constructed a smarter agent that it can't simulate. I think it's pretty bad news if we are relying on a smart model not understanding facts about its own situation, and that's the kind of solution that is particularly likely to break down over time.
- We could hope that our models don't build up important inaccessible knowledge, e.g. because it's possible to justify most interesting conclusions with deduction or because we have reasonably good accessible proxies for our value function. I'm pretty skeptical about this over the long term, but I'm not sure exactly how bad it will be how early.
- The argument in this post is pretty informal, and there's a reasonable chance that you can drive a solution through one of the many gaps/loopholes. I like the problem-solving strategy: "write out the proof that there is no solution, and then sift through the proof looking for a fatal hole."

Overall I don't see an obvious way forward on this problem, but there are enough plausible angles of attack that it seems exciting to think about.

VI. How this relates to amplification and debate

Overall I don't think it's very plausible that amplification or debate can be a scalable AI alignment solution on their own, mostly for the kinds of reasons discussed in this post—we will eventually run into some inaccessible knowledge that is never produced by amplification, and so never winds up in your distilled agents.

In the language of my original post on capability amplification, the gap between accessible and inaccessible knowledge corresponds to an obstruction. The current post is part of the long process of zooming in on a concrete obstruction, gradually refining our sense of what it will look like and what our options are for overcoming it.

I think the difficulty with inaccessible knowledge is not specific to amplification—I don't think we have any approach that moves the needle on this problem, at least from a theoretical perspective, so I think it's a plausible candidate for a hard core if we fleshed it out more and made it more precise. (I would describe MIRI's approach to this problem could be described as despair + hope you can find some other way to produce powerful AI.)

I think that iterated amplification *does* address some of the most obvious obstructions to alignment—the possible gap in speed / size / experience / algorithmic sophistication / etc. between us and the agents we train. I think that having amplification mind should make you feel a bit less doomed about inaccessible knowledge, and makes it much easier to see where the real difficulties are likely to lie.

But there's a significant chance that we end up needing ideas that look totally different from amplification/debate, and that those ideas will obsolete most of the particulars of amplification. Right now I think iterated amplification is by far our best concrete alignment strategy to scale up, and I think there are big advantages to starting to scale something up. At the same time, it's really important to push hard on conceptual issues that could tell us ASAP whether amplification/debate are unworkable or require fundamental revisions.

Inaccessible information was originally published in AI Alignment on Medium, where people are continuing the conversation by highlighting and responding to this story.

News ⊂ Advertising

For-profit news outlets are financial incentivized to write about things that are easy to write about. The easiest articles to write are the subsidized ones. Public relations firms subsidize news by writing press releases. Then news outlets republish the press releases as news. That's why so much news is corporate and political advertising.

Here are the top stories on *Ars Technica* at the time of writing[1].

1. "NordVPN users' passwords exposed in mass credential-stuffing attacks"
2. "AT&T's priciest "unlimited" plan now allows 100GB+ of un-throttled data"
3. "Researchers unearth malware that siphoned SMS texts out of telco's network"
4. "The count of managed service providers getting hit with ransomware mounts"
5. "Facebook deletes the accounts of NSO Group workers"

Having only skimmed the articles, I suspect they were put there by the following companies.

1. Have I Been Pwned (breach notification service)
2. AT&T
3. FireEye (security firm)
4. Armor (global cloud security provider)
5. Facebook

The first article lets slip who wrote it in the following line.

> Readers who are NordVPN users should visit Have I Been Pwned[2] and check to see if their email address is contained in any of the lists.

Can you spot how this sentence attempts to influence reader behavior?

Different organizations write articles for different news outlets. *Ars Technica* is unusual in its disproportionate publishing of articles written by cybersecurity firms and its relatively low density of political propaganda compared to more traditional news outlets like *The Economist*. *The Art of Manliness Podcast* interviewees usually discuss the books they're selling.[3]

News is advertising. Ad-supported news is ad-supported advertising. Subscription-supported news is subscription-supported advertising. Advertising can't directly control what you believe. Advertisers can control what you think about. The more advertising I expose myself to, the more I think about the things advertisers want me to.

Here is what advertisers want me to think about.

- Products I haven't bought
- National politics[4]
- More ad-supported media, such as celebrities and free-to-play videogames

Here is what I want to think about.

- Things I can make and do myself
- My local community

- My friends, my family and me

My personal happiness is inversely related to how much news I expose myself to. It's not just a subjective feeling. I behave more healthily. I'm even more interesting to talk to.

Amateur blogs make me think about what the author thinks is important. That's a step in the right direction because amateur bloggers' interests align better with mine than do the corporate and political machines behind news outlet press releases. But they're still not me. And some of them are motivated by vanity.

I solve all of these issues by writing a blog myself. That way the author's interests align perfectly with my own.

If you liked this post, click here.

Edit: jballoch points out that Have I Been Pwned is a noncommercial donation-supported service.

---

1. November 3, 2019 at 1:43 am ↩
2. The hyperlink is in the original article. It's the article's second link to Have I been Pwned. ↩
3. I pick these specific news outlets because I visit them the most. Aggregators like Facebook and Reddit are different beasts deserving of a separate post. ↩
4. I don't deny that national politics is important. I mean that the proportion of attention it gets on the news is greater than the proportion of my attention I wish to passively devote to it. ↩

Superexponential Historic Growth, by David Roodman

This is a linkpost for https://www.openphilanthropy.org/blog/modeling-human-trajectory

This is research trying to do a similar analysis to Hanson's paper Long-Term Growth as a Sequence of Exponential Modes, and coming to different conclusions in some areas and the same conclusions in others. It also discusses Scott's post 1960: The Year the Singularity Was Cancelled. The name of the post/paper is "Modeling Human Trajectory".

From the summary.

> One strand of analysis that has caught our attention is about the pattern of growth of human society over many millennia, as measured by number of people or value of economic production. Perhaps the mathematical shape of the past tells us about the shape of the future. I dug into that subject. A draft of my technical paper is here. (Comments welcome.) In this post, I'll explain in less technical language what I learned.
>
> It's extraordinary that the larger the human economy has become—the more people and the more goods and services they produce—the faster it has grown on average. Now, especially if you're reading quickly, you might think you know what I mean. And you might be wrong, because I'm not referring to exponential growth. That happens when, for example, the number of people carrying a virus doubles every week. Then the *growth rate* (100% increase per week) holds fixed. The human economy has grown *super*-exponentially. The bigger it has gotten, the faster it has doubled, on average. The global economy churned out $74 trillion in goods and services in 2019, twice as much as in 2000.**1** Such a quick doubling was unthinkable in the Middle Ages and ancient times. Perhaps our earliest doublings took millennia.
>
> If global economic growth keeps accelerating, the future will differ from the present to a mind-boggling degree. The question is whether there might be *some* plausibility in such a prospect. That is what motivated my exploration of the mathematical patterns in the human past and how they could carry forward. Having now labored long on the task, I doubt I've gained much perspicacity. I did come to appreciate that any system whose rate of growth rises with its size is inherently unstable. The human future might be one of explosion, perhaps an economic upwelling that eclipses the industrial revolution as thoroughly as *it* eclipsed the agricultural revolution. Or the future could be one of implosion, in which environmental thresholds are crossed or the creative process that drives growth runs amok, as in an AI dystopia. More likely, these impulses will mix.

And from the conclusion.

> I do not know whether most of the history of technological advance on Earth lies behind us or ahead of us. I do know that it is far easier to imagine what has happened than what hasn't. I think it would be a mistake to laugh off or dismiss the predictions of infinity emerging from good models of the past. Better to take them as stimulants to our imaginations. I believe the predictions of infinity tell us

> two key things. First, if the patterns of history continue, then some sort of economic explosion will take place again, the most plausible channel being AI. It wouldn't reach infinity, but it could be big. Second, and more generally, I take the propensity for explosion as a sign of *instability* in the human trajectory. Gross world product, as a rough proxy for the scale of the human enterprise, might someday spike or plunge or follow complicated paths in between. The projections of explosion should be taken as indicators of the long-run tendency of the human system to diverge. They are hinting that realistic models of long-term development are unstable, and stable models of long-term development unrealistic. The credible range of future paths is indeed wide.
>
> *Data and code for the paper and for this post are on GitHub .*

Holden Karnofsky also gives his thoughts on the piece, which I found quite interesting.

> Some personal reactions on this piece:
>
> First, a note on how it came about and what I think the relevance to our work is. I asked David to evaluate Robin Hanson's work on long-term growth as a sequence of exponential growth modes. I found it interesting that an attempt to extrapolate future economic growth from the past (with very little reasoning other than attempting to essentially trend-extrapolate) implied a strong chance of explosive growth in the next few decades, but I wasn't convinced that Hanson's approach was the best method of doing such trend extrapolation. I asked David how he would extrapolate future growth based on the past, and this is the result. The model is very different from Hanson's (and I prefer it), but it too has an implication of explosive growth in the next few decades.
>
> On its own, seeing trend extrapolation exercises with this implication doesn't necessarily mean much. However, I independently have a view (based on other reasoning) that transformative AI could plausibly be developed in the next couple of decades. I think one of the best reasons to be skeptical of this view about transformative AI is that it seemingly implies a major "trend break": it seems that it would, one way or another, have to mean world economic growth well outside the 1-3% range that it's been pretty steady in for the last couple of centuries. However, Hanson's and Roodman's work both imply that a broader look at economic history demonstrates accelerating growth, and that in this sense, expecting that "the future will be like the past" could be entirely consistent with expecting radically world-changing technology to be developed in the coming decades.
>
> Like David, I wouldn't take the model discussed in this piece literally, but I tentatively agree with what I see as the central themes: that a sufficiently broad view of history shows accelerating growth, that this dynamic is inherently unstable, and that we therefore have little reason to "rule out" explosive growth in the coming decades.
>
> We are working on a number of other analyses regarding the likelihood of transformative AI being developed in the coming decades. One topic we're exploring is a potential followup on this piece in which we would try to understand the degree to which growth economists find this piece's central themes reasonable, and what objections are most common.
>
> Now a few comments on ways in which I see things differently from how David sees them. I should start by saying that any model makes simplifications, and this

is a case where extreme simplifications are particularly called for. However, if I'd written this post I would've called out the following non-modeled dynamics as particularly worth noting.

1 - this post's multivariate model does not match the way I intuitively model what I call the "technological landscape." Some discoveries and technological developments enable others, so there is in some sense an "order" in which we've developed new technologies that might be fairly stable across multiple possible versions of history. And some technologies are more impactful than others. There may thus be important natural structure that leads to inevitable (as opposed to stochastic) acceleration *and* deceleration, as the world hits phases where there are more vs. less impactful technologies being discovered. The most obvious way in which I expect the "technological landscape " to matter is that at some point, I think the world could "run out of new findings" - at which point technology could stop improving. I see this as a likely way that real-world growth could avoid going to infinity in finite time, without needing to invoke natural resource limits.

2 - it seems to me that a more realistic multivariate model would have natural resource shortages leading to growth "leveling off" rather than spiking and imploding. E.g., at the point where natural resources are foreseeably going to be a bottleneck on growth, I expect them to become more expensive and hence more carefully conserved. I'm not sure whether this would apply to a long enough time frame to make a big visual difference to the charts in this post, but I still thought it was worth mentioning.

3 - I'm interested in the hypothesis that the recent "stagnation" this model sees is largely driven by the fact that population growth has slowed, which in turn limits the rate of technological advance. Advances in AI could later lead to a dynamic in which capital can more efficiently substitute for labor's (and/or human capital's) role in technological advance. This is an example of how the shape of the "technological landscape" could explain some of the "surprises" seen in David's tests of the model.

4 - regarding this statement in David's piece:

The scenario is, one hopes, unrealistic. Its realism will depend on whether human enterprise ultimately undermines itself by depleting a natural endowment such as safe water supplies or the greenhouse gas absorptive capacity of the atmosphere; or whether we skirt such limits by, for example, switching to climate-safe energy sources and using them to clean the water and store the carbon.

In worlds where explosive growth of the kind predicted by David's model occured, I'd anticipate radical changes to the way the world looks (for example, civilization expanding outside of Earth), which could significantly change the picture of what resources are scarce.

Sparsity and interpretability?

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

A series of interactive visualisations of small-scale experiments to generate intuitions for the question: "Are sparse models more interpretable? If so, why?"

Assumptions and motivation

- Understanding interpretability may help AI alignment directly (e.g. recursive approaches, short/mid-term value; see here) or indirectly by supporting other directions (mechanistic transparency, open-source game theory).
- Recent deep learning sparsification techniques retain most of the network's performance (even when dropping 99% of the weights).
- At least small systems are easier to both analyze when sparse (e.g. information flow, case debugging, local behaviour interpretation) and explore & visualize.
- Experiments with small networks can generate or challenge general intuitions (with some caveats) and inform further technical research.
- In particular, we assume that moving towards early, simplified or naive definitions of interpretability concepts and other engagement are useful steps towards more general understanding (as one possible way).

- We want more people to engage with these questions by playing with hands-on examples and challenging & debating concrete propositions. Reach out and comment!

Many of the visualisations here are interactive. Due to technical forum limitations, the interactive versions are on an external page here and are linked to from every visualisation snapshot in this post.

Introduction [Optional]

We want to understand under which circumstances (if any) a sparser model is more interpretable. A mathematical formulation of "transparent or interpretable" is currently not available: sparsity has been proposed as a possible candidate proxy for these properties, and so we want to investigate how well the links between sparsity and interpretability holds, and what kinds of sparsity aid in what kinds of interpretability.

(A note on the field of Interpretability: we've talked a bit previously about what interpretability is, as the topic is broad and the term is very underdefined. It's important to note, when discussing whether a particular property (like sparsity) makes a model more or less interpretable, we need to specify what kind of interpretability method we're using, as different methods will benefit from different properties of the model.)

We have several intuitions that sparsity in neural networks might imply greater interpretability, and some which point in the opposite direction:

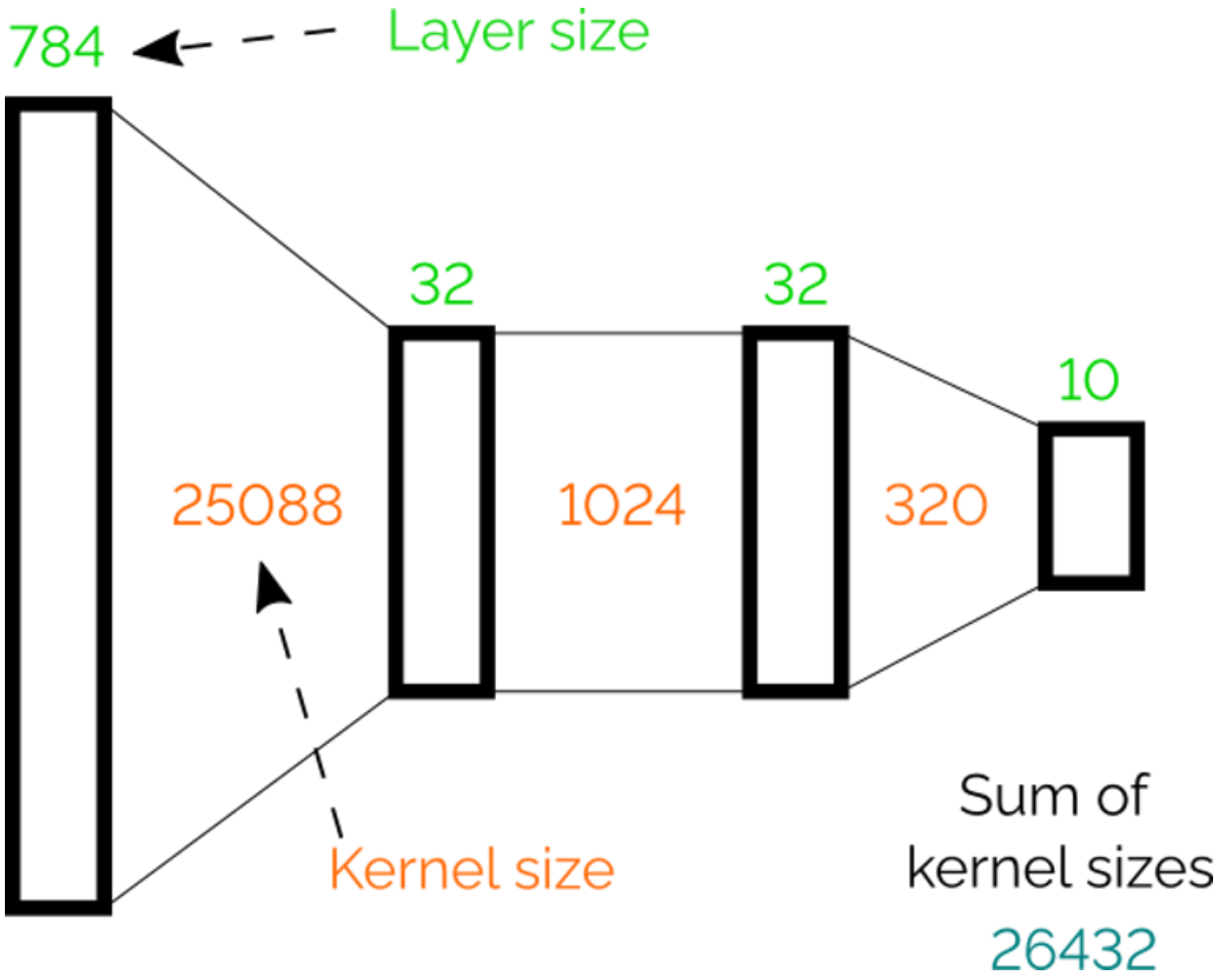
- Debugging a system with many weights and interconnections seems more complex than to do the same for a sparser system with less parameters and connections; High connectivity is one of reasons why neural networks are not considered transparent.
- On the other hand, human-recognizable features are often complex objects - such as in an image classification task, where humans recognize faces, dog ears, and trees as opposed to simple patterns of white noise. There is hence a risk that making a system sparser could lessen its capacity to work with such high-level concepts.
    - In other words, a dense system may be able to use complex but human-interpretable features, whereas a sparse system wouldn't have the capacity to recognise these features and hence focus on simpler features which are less meaningful to humans.
- Neural networks trained against adversarial attacks produce more human recognizable features [1]. If making the system sparser makes it more vulnerable to adversarial attacks, then it may then be less human-interpretable. However, it is unclear in what direction the causality would flow.

In this post we investigate a specific instantiation of an interpretability method which intuitively seems to benefit greatly from sparsity, and discuss ways in which we can make the most of these benefits while not changing the network to be too sparse and lose performance

Introduction of the Experiment

To test whether sparsity is good for interpretability, the obvious thing to do is to get a sparse neural network, and see whether it's easier to interpret. To do it we need following things: a neural network, a pruning method to produce a sparse network, and an interpretability method to apply to this sparse network.

Neural network. We will use a simple network of the following architecture trained on MNIST (2 hidden layers of size 32 with ReLU activations; in total 26432 weights in kernels):



This architecture gets 95+% accuracy on the MNIST test set after three epochs of training.

Pruning method. The lottery ticket hypothesis [2] claims that neural networks pruned, reset to their initial training weights, and retrained, and that this retrained network will often learn faster, attain better asymptotic performance and generalise better than the original network. The pruning method used in this paper is relatively simple:

- Initialize the neural network, and save the initialisation.
- Train the network to convergence.
- Prune X% of smallest weights per layer.
- Reset the network to its initialization,
- Repeat the above process, each time pruning a percentage of the weights.

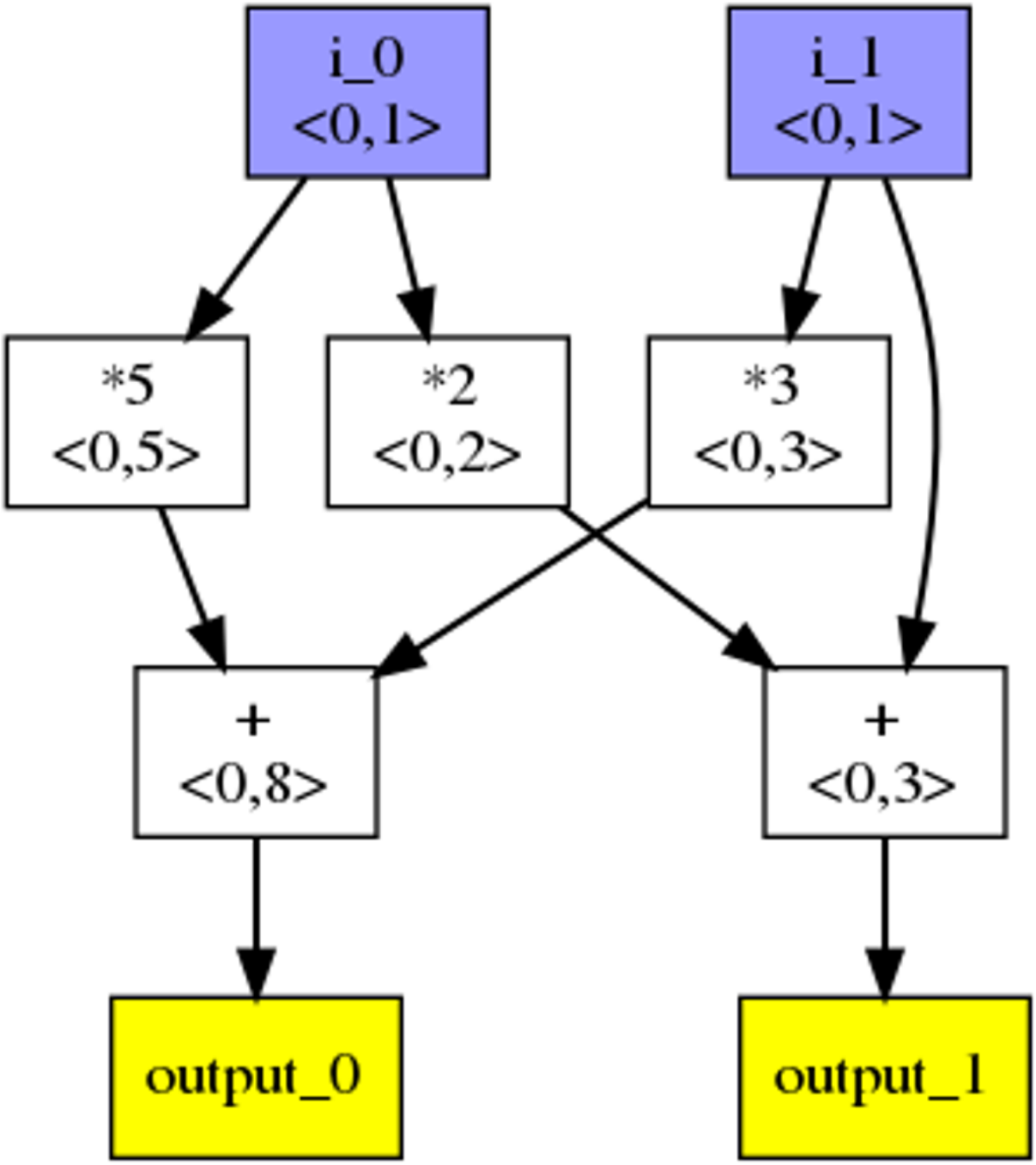
Note that in our experiment we are pruning only only kernels, not biases. As there are only 74 weights in biases, this isn't a consequential number of parameters compared to 27432, and it makes the implementation simpler.

Gaining insight. Having matrices with many zeros is not in itself helpful. We need to use an interpretation method to take advantage of this sparsity. Because we hope for quite a sparse representation at the end, chose to visualize the network as it's flowgraph.

To demonstrate how this works, we'll use the following computation:

$$\begin{bmatrix} 5 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} i_0 \\ i_1 \end{bmatrix} = \begin{bmatrix} output_0 \\ output_1 \end{bmatrix}$$

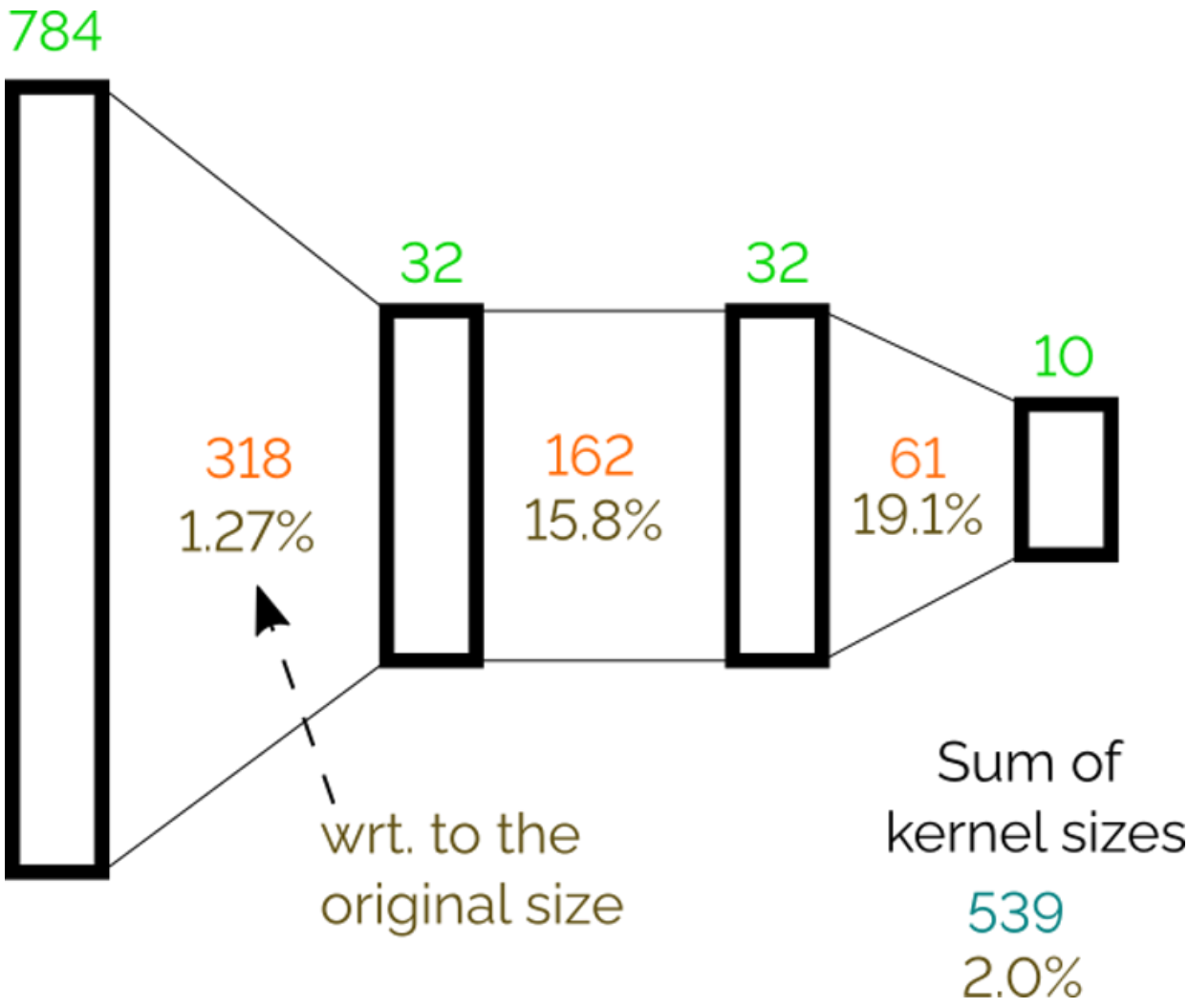
This can be represented as the following flowgraph:



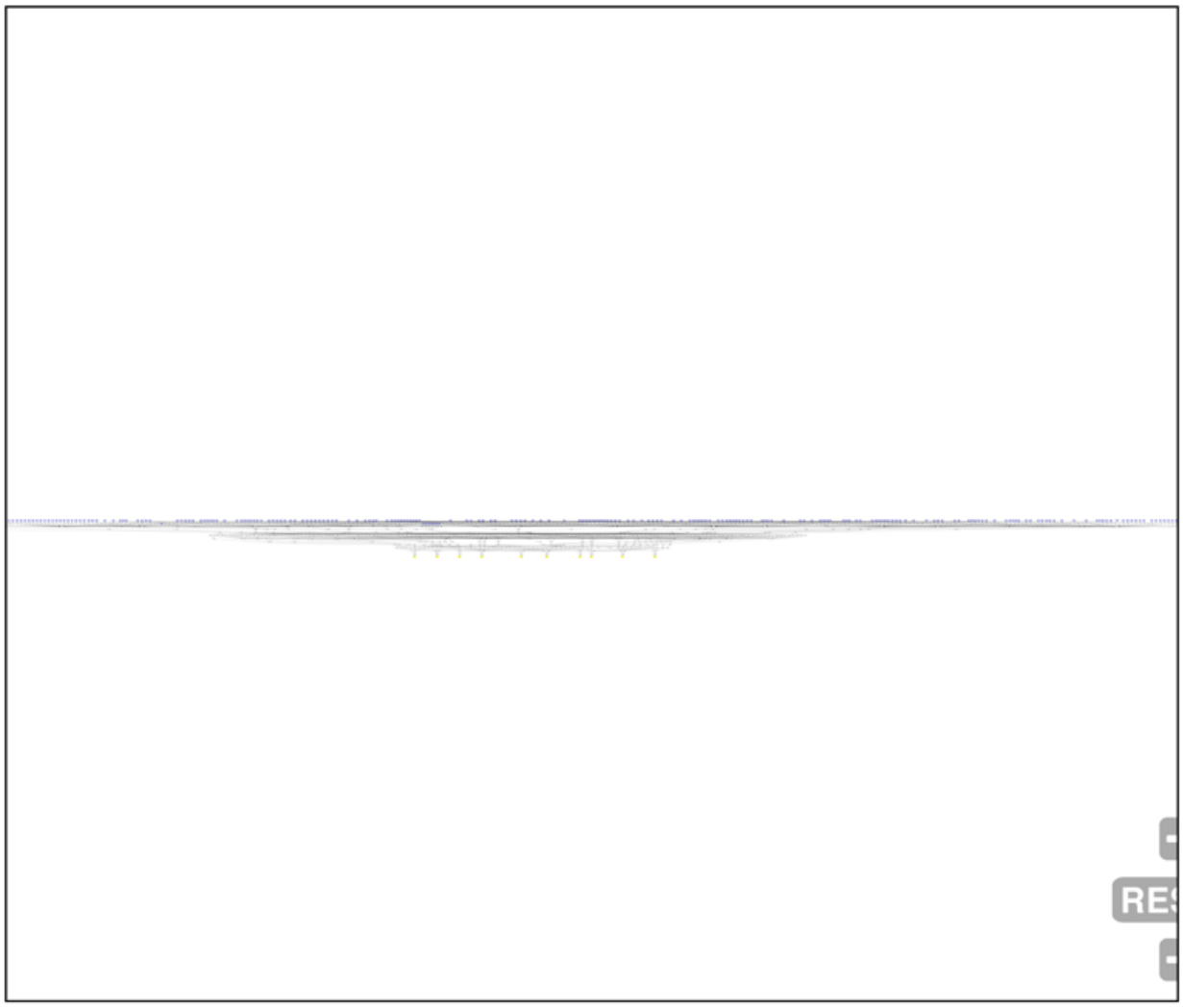
- *Each box produces a scalar value*
- *Blue nodes are inputs, and yellow nodes are outputs.*
- <x,y> *shows an a simple overapproximation of bounds. In this example we assume that the inputs are in the interval* $[0, 1]$. *Since the computation is simple, the calculated bounds for outputs are precise in this particular case.*

Initial results

When we apply the lottery ticket pruning method on all layers, we can prune the network to 2% of its original size (leaving only 539 weights) and still achieve 90.5% accuracy on test set. For more detail on the number of weights per layer see the following figure:



Now we can apply our flow-graph visualisation on the pruned network. The output is a graph with 963 vertices and 1173 edges. Note that for this figure and all following, we are rounding values to two decimal places.

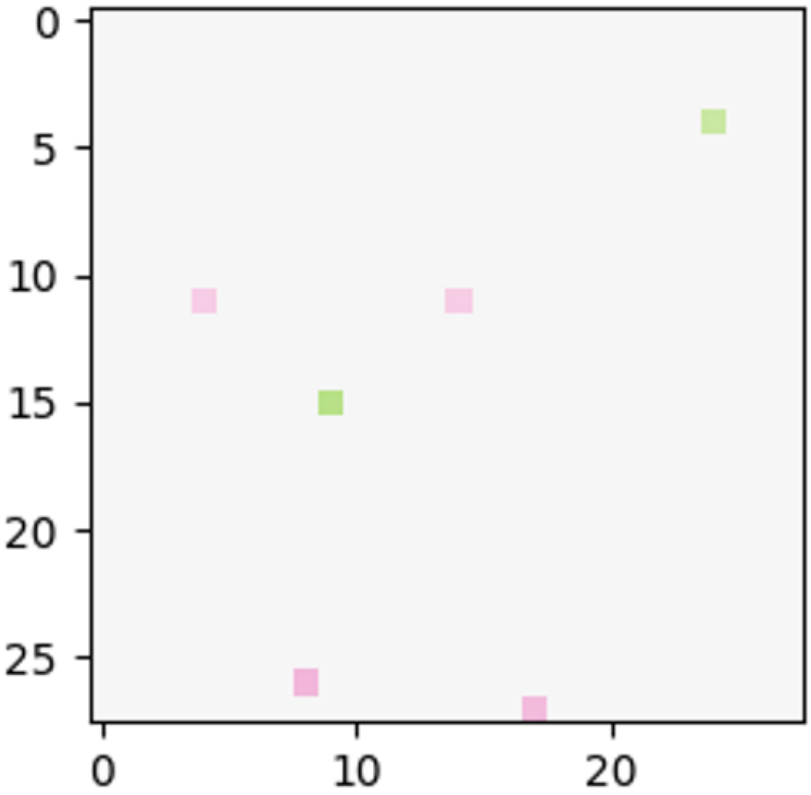


Making it clearer

As we can see, the flowgraph is quite still quite dense, and it's difficult to extract any insight into how the network behaves from it. To make it clearer, we can exploit the following observations: Many vertices and corresponding edges are from the input layer to the first hidden layer. We can detect dot product operations in the first layer and simply visualize weights as a 2D image, because the first layer is directly connected to the pixels. For example, instead of the following dot product:

$$0.82i_{[4,24]} - 0.72i_{[11,4]} - 0.72i_{[11,14]} + 0.99i_{[15,9]} - 0.98i_{[26,8]} - 0.92i_{[27,17]}$$

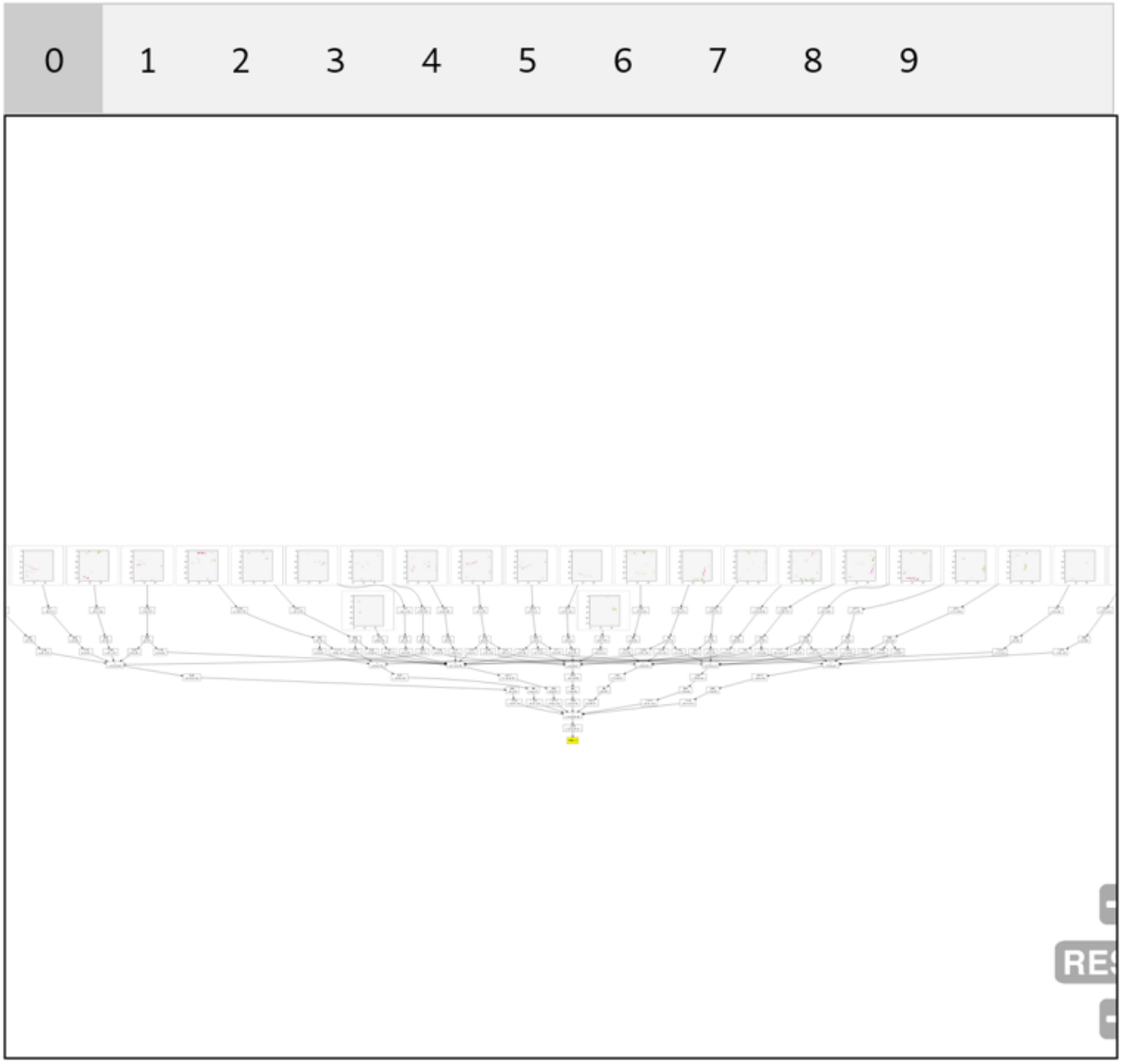
(where $i_{[x,y]}$ is an input pixel at position $[x, y]$) we can instead use the following image:



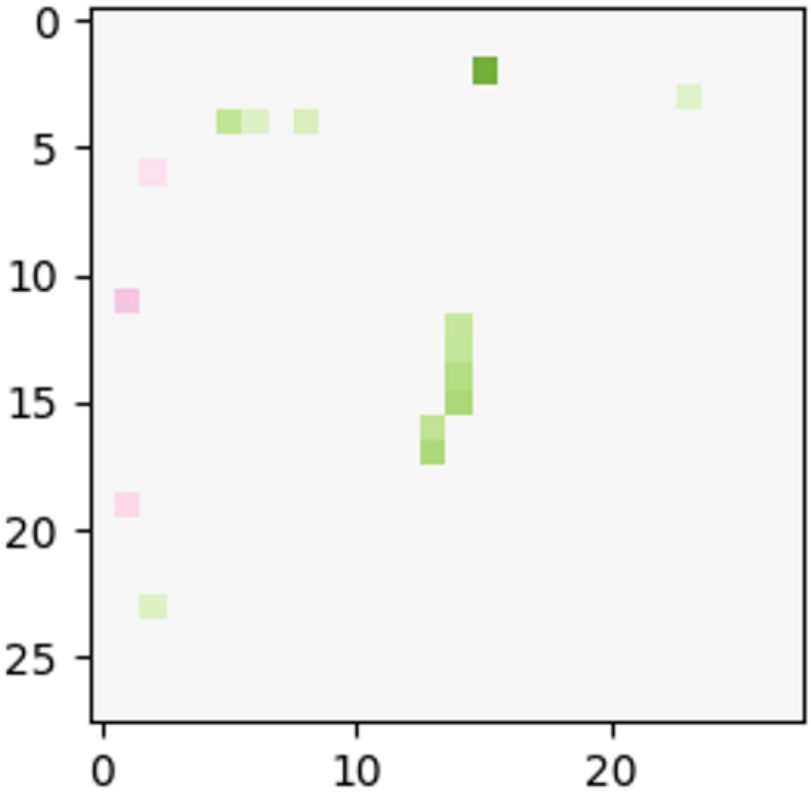
Colour intensity is used to show absolute value; green is used for positive values; red for negative, and grey is the zero.

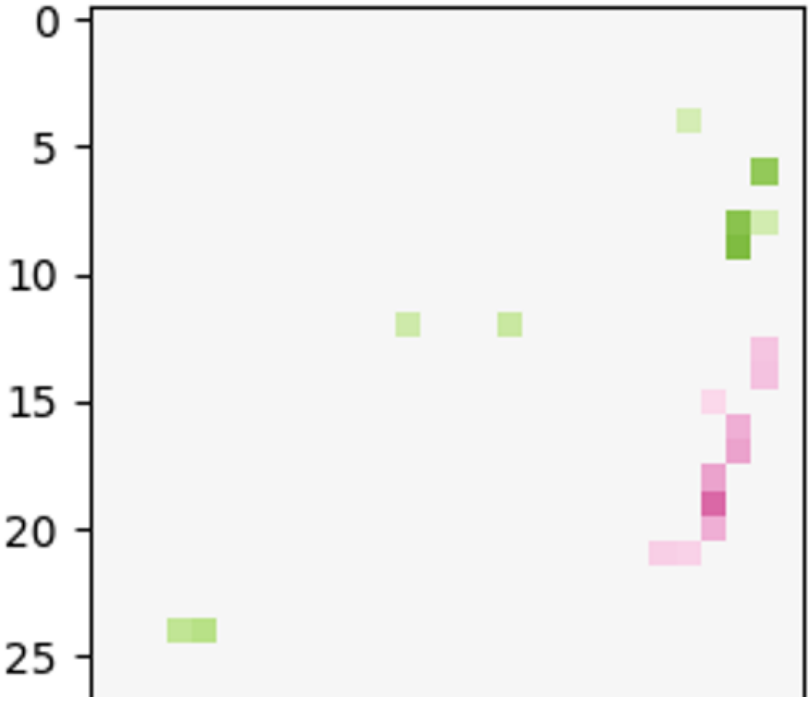
A further simplification is to display only subgraph of the flowgraph that is relevant for each output separately. While it loses information about what is reused between outputs, it provides more readable graphs.

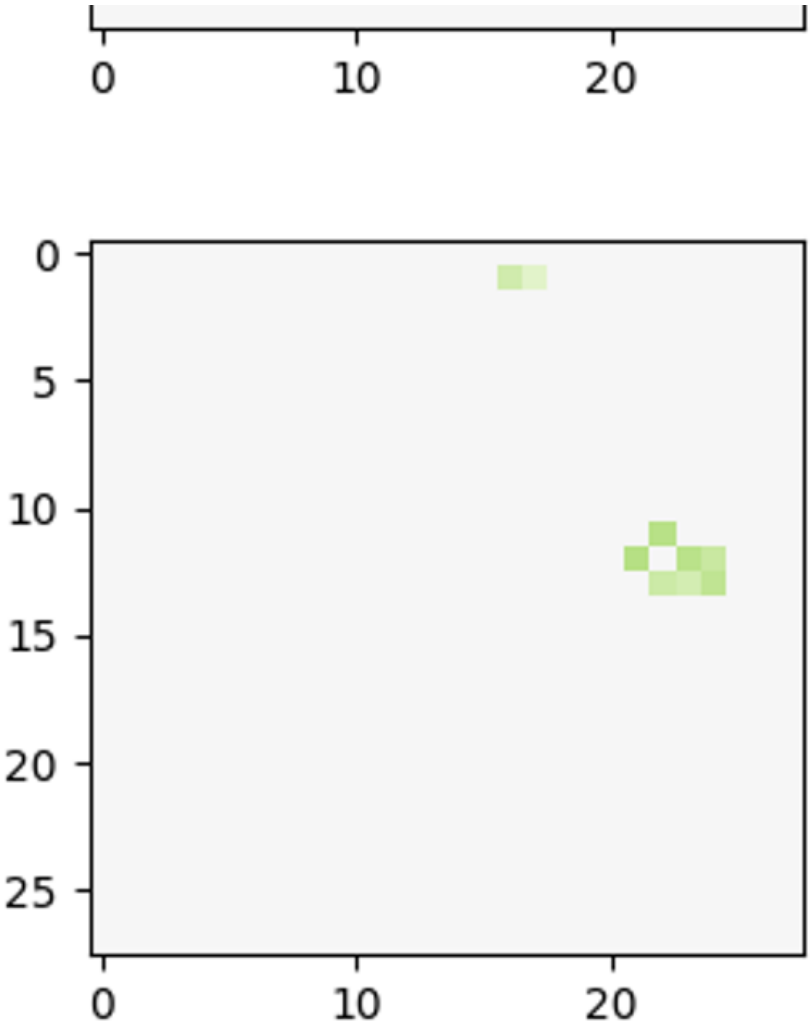
Combining these two simplifications, we get the following visualisation:



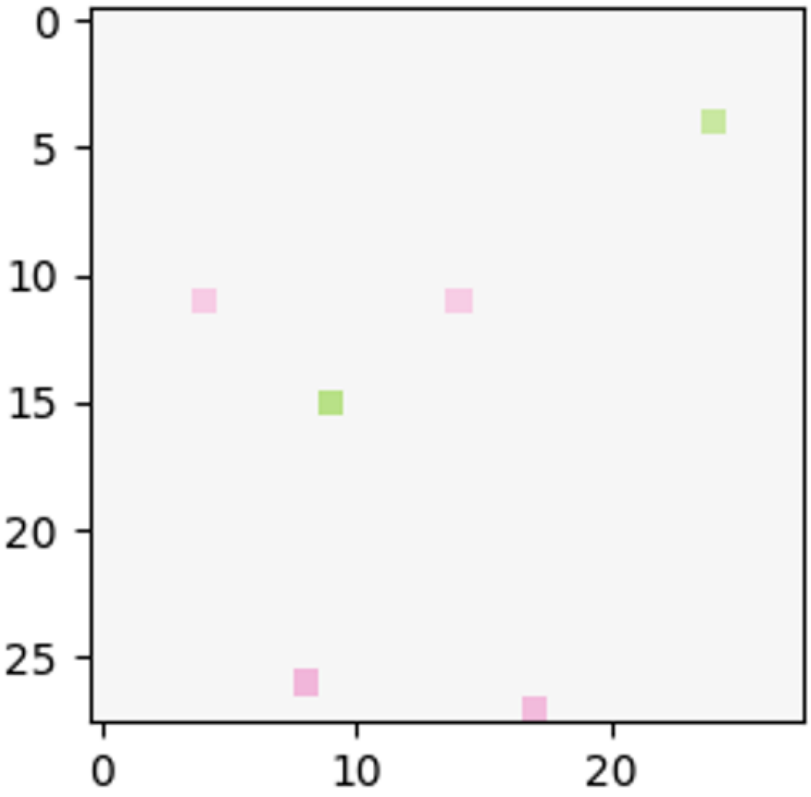
Here we can observe that some filters on the input are relatively interpretable, and it is clear what shapes they detect:

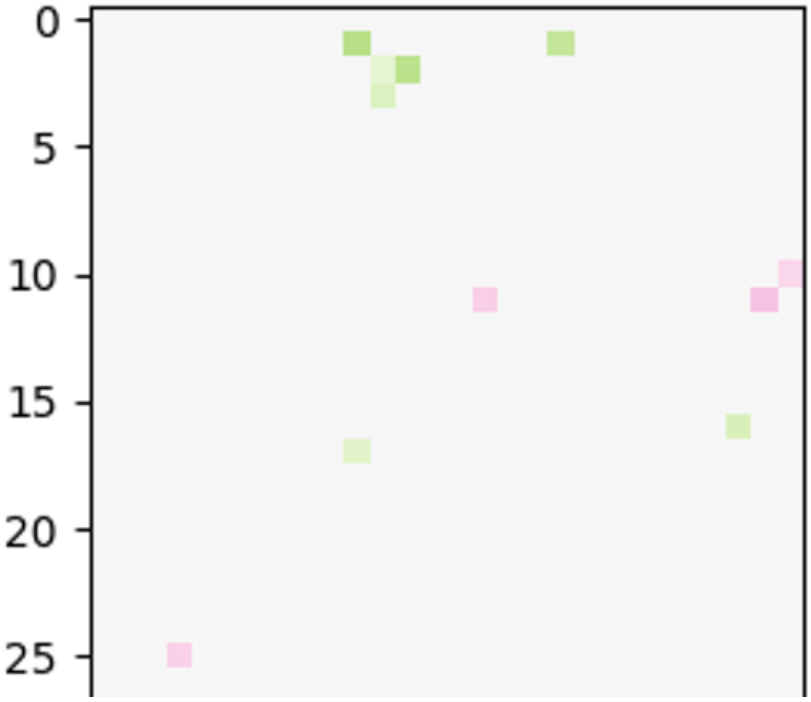

0
5
10
15
20
25
0
10
20

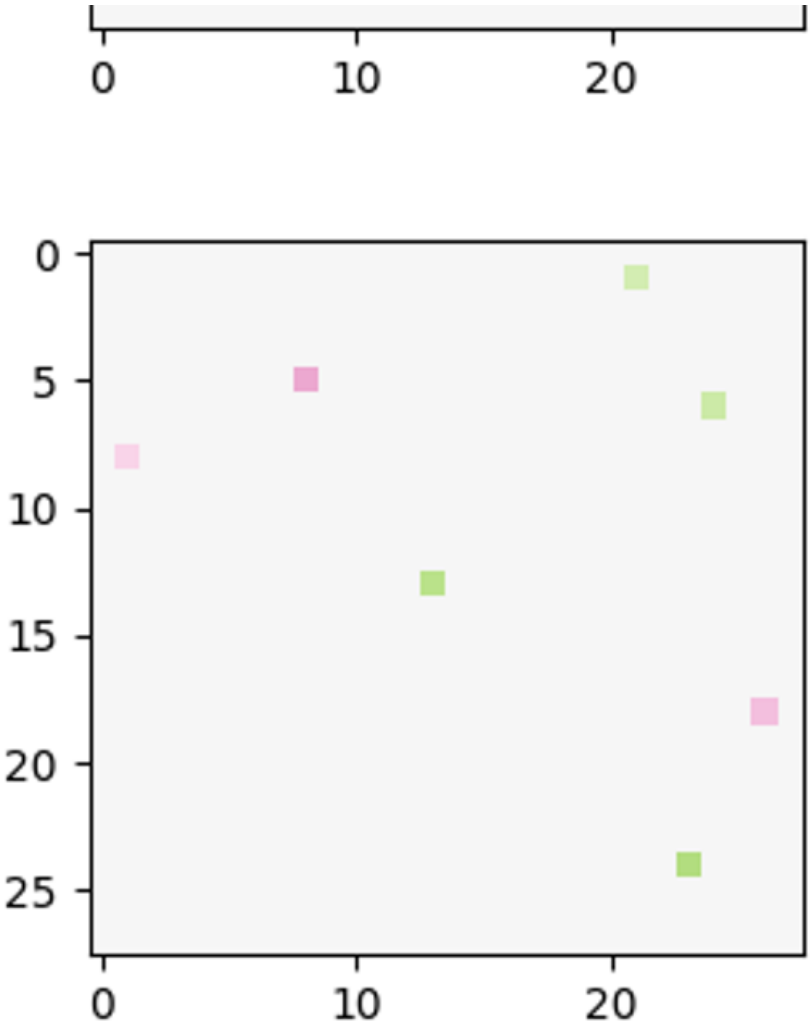

0
5
10
15
20
25



On the other hand, some of them look more like random pixel detections:


0
5
10
15
20
25
0
10
20


0
5
10
15
20
25



The computation after the initial layer still remains relatively dense and hard to understand. It is now possible to track the flow of individual values through the graph in most places, but we can do better.

Visualising behaviour on specific inputs

To get more insight, we can also visualize flowgraphs on individual inputs. This allows us to visualize values flowing between nodes; the thickness of lines corresponds to the absolute value of the scalar value produced by a source operation while green lines carry positive values, red lines carry negative values, and grey lines are exactly 0.

Also, for each visualized dot product, we show the result of applying this filter to the input image (that is, the result of the dot product multiplication of the input and the weights). This is shown as the right-hand figure next to the dot product filter visualisations at the top of the flowgraph.

The following figure shows this visualisation for six random images from the test set (the values in square brackets in the nodes are the actual values that flow through the graph):

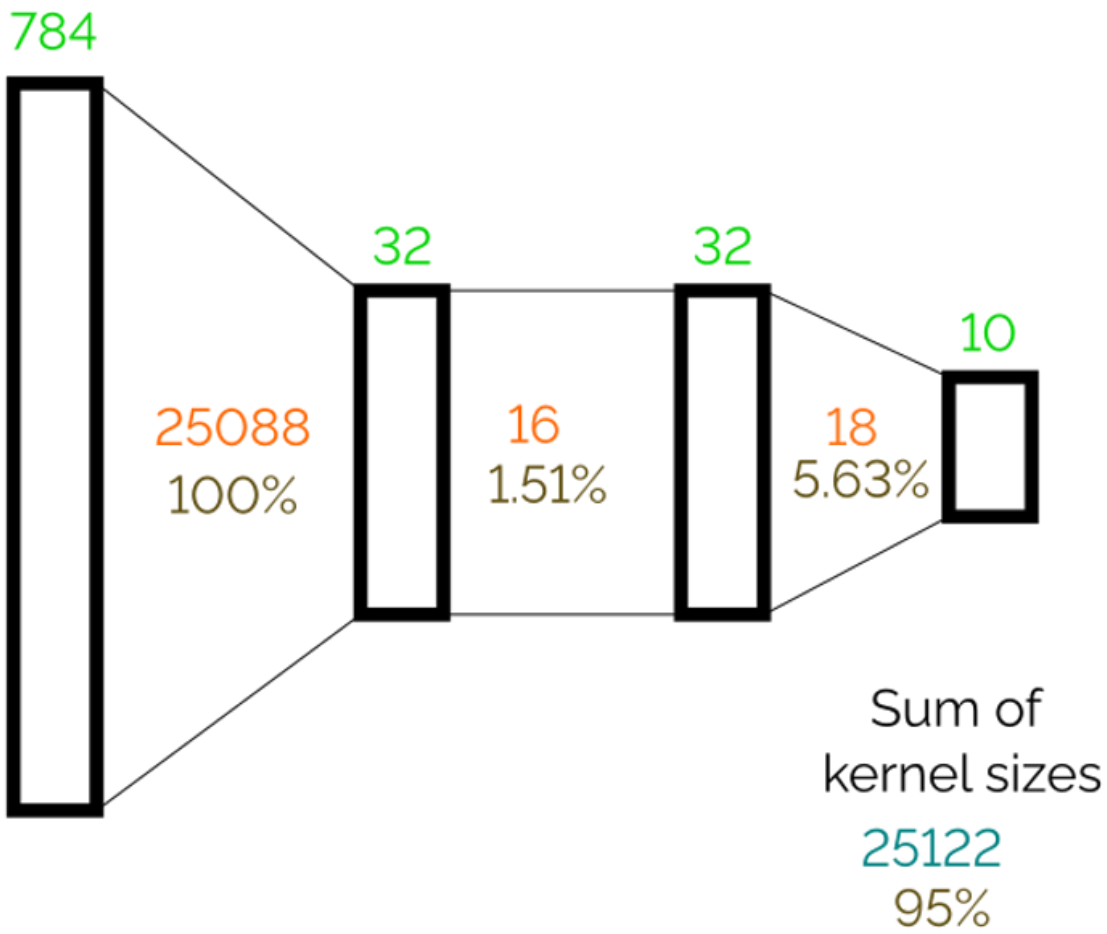
The figure shows that it relatively trackable to see which part of the network contributes to the final decision. For example, let us see how the probability that the input is digit 0 (`output_0`) is computed. For the digit 1 input, it is clear that the decision

to not classify it as a 0 is heavily influenced by one specific dot product that detects pixels in the middle of the figure. Another observation is that there are two main branches in the flow that plays an important role of classifying digit 0 correctly (i.e. digit 0 and `output_0`). It mixes together a detector of a bottom pixels and detection of some specific pixels. For the sixth input (digit 6) is used, it is not classified as 0 even though some of detectors are positively triggered (similarly to case with input 0); however, the detector of the central pixels overweights and produce a negative score.

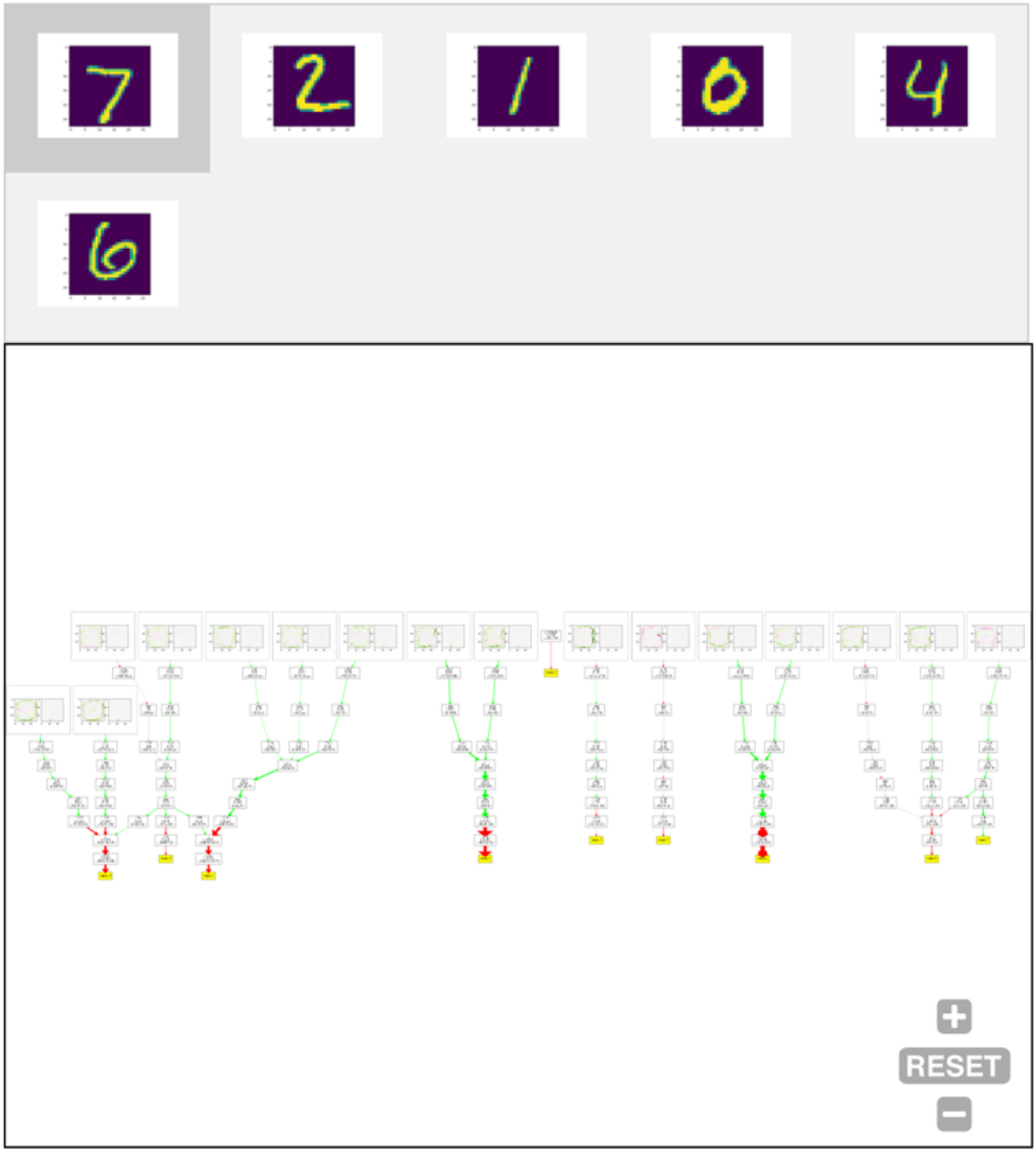
Pruning once more

Let us revisit the idea of pruning. We have pruned the network as much as possible to get a more understandable object then a fully connected network. However, there is space for trade-offs; we don't need to prune each layer equally, and could trade off pruning some layers more to prune other layers less.

Specifically, in our case we have a visualization of the first layer who's interpretability is not influenced as much by the sparsity. This allows us to completely disable pruning of the first layer. With this approach we get the following network flowgraph that has 93.5% accuracy on the test set:



The network has 95% of its original weights, but last two kernels are several times smaller than in the previous case. This gives us a significantly simpler flowgraph. It can be now easily shown in one diagram for each outputs without loosing clarity. The following figure shows resulting flow graphs applied on six figures from the testing test.



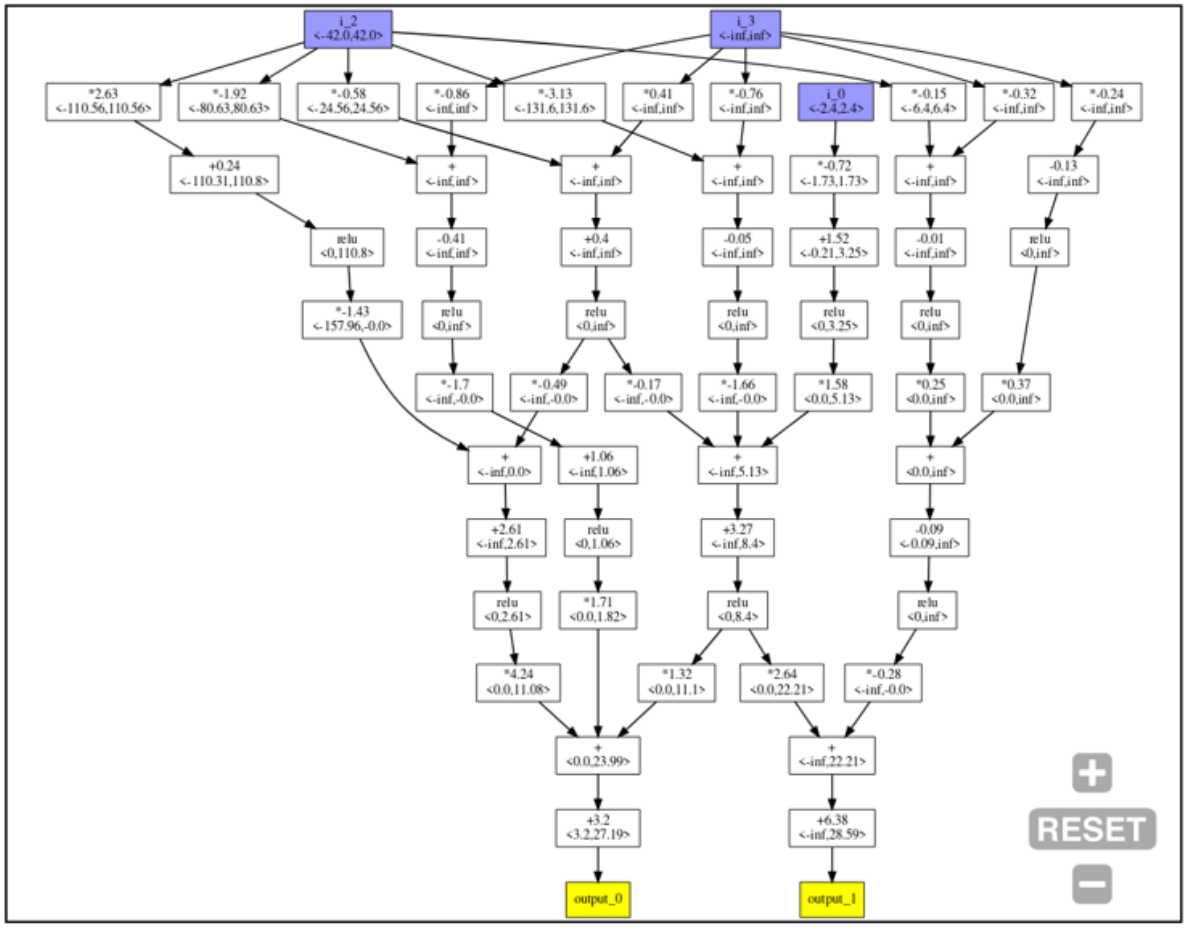
One interesting observation is that this network returns a constant value for `output_2`. This means that the digit 2 is a kind of reference point; an input is classified as digit 2 if other outputs are sufficiently suppressed.

We can again easily track down properties like digit 1 is not classified as 0 because of the pixel in the middle part. In this version, it is less clear what each dot products

detects, but on the other hand it very straightforward what is happening with each dot product value in the rest of the computation.

Domains other than MNIST

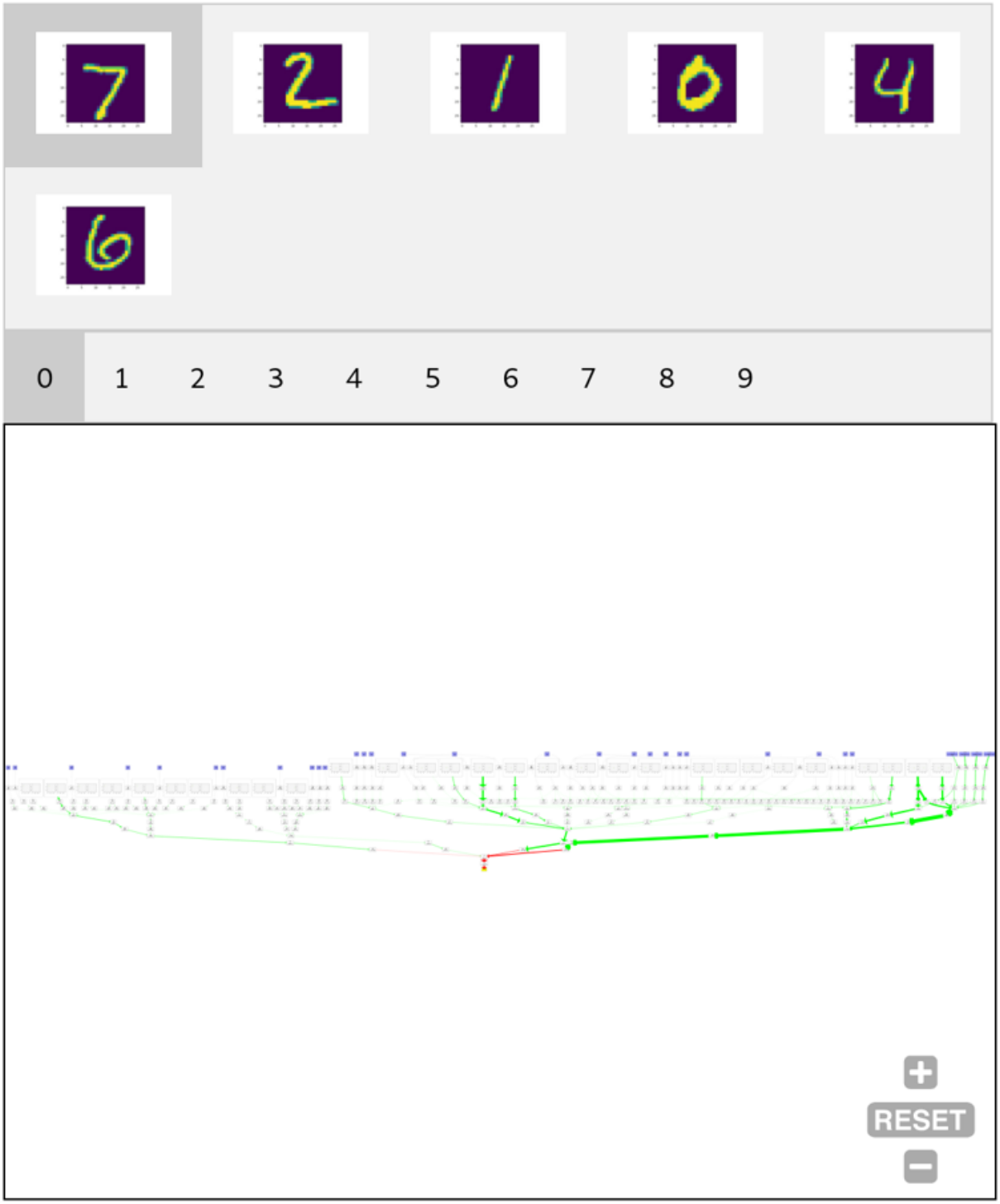
The approach of pruning and visualization of flow graphs may be also applied to other domains such as reinforcement learning for continuous control. We use the classic OpenAI CartPole-v0 environment. The following figure shows a network trained with DQN, and then pruned when reaching maximal score, on the CartPole environment. For this network, we can observe that one input (i_1, which is the cart velocity) is completely ignored, while the agent still gains the maximal score.



On the other hand, our approach doesn't seem useful for convolutional neural networks, since applying the convolutional filters over the whole output of the previous layer is a part of the algorithm and it is not pruned with weights. Therefore, a convolutional filter even with a single weight may produce many connections in the flow graph. The presented approach still produces dense flow graphs even for heavily pruned convolutional networks.

The following figure shows a network created as 3 convolutional layers (6 channels, 3x3 filters, without padding, kernel sizes: 54, 324, 324) followed by a dense layer (kernel size: 29040, no activation). The layers was pruned to 6 (11%), 23 (7%), 23

(7%), and 135 (0.46%) of their original sizes. The resulting network has only 80% accuracy and quite a low number of weights, but still produces a relatively large and dense graphs. The graph for computing `output_7` is the largest one (1659 vertices).



Conclusion

While we may not have conclusively answered the question posed in the title, we have generated insight and intuitions with this simple example. As expected, pruning dense object may lead to more interpretable objects; however, there are trade-offs. Only counting the number of weights in the final object is probably not the ideal metric to get a most interpretable result. In our experiment, the later flowgraphs (of dense nets) where more interpretable despite having many more weights than the previous examples. This points to an answer to the opposing views of sparsity possibly giving both more and less interpretability; In some settings sparsity is beneficial in giving humans understanding and transparency, but when a visualisation which doesn't rely on sparsity can still produce results that humans can understand and interpret, the sparsity isn't necessarily useful.

Acknowledgement: This text was created as a part of AISRP.

References

Adversarial Robustness as a Prior for Learned Representations [PDF]
Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B. and Madry, A., 2019.

The Lottery Ticket Hypothesis: Training Pruned Neural Networks [PDF]
Frankle, J. and Carbin, M., 2018. CoRR, Vol abs/1803.03635.

What are some Civilizational Sanity Interventions?

Lately, I've been thinking about the class of things that I'm calling "Civilizational Sanity Interventions." With that term I'm meaning to refer to technologies, institutions, projects, or norms that, if implemented, would improve the quality of high level decision making about important issues.

Which things if they existed in the world, would make our society, collectively, saner?

Some examples (with which I expect most people around here to be familiar):

Prediction markets

Prediction markets are a clever way to aggregate all the available information to make accurate predictions.

Robin Hanson posits that the reason why there isn't wider adoption of prediction markets is because they are a threat to the authority of existing executives.

If we lived in a world where the use of prediction markets were commonplace standard practice, eventually, decision makers would face flack for acting against the predictions of the market, and pundits would have a lot less leeway to make inaccurate, politically-motivated predictions.

Hanson, in a recent interview,

> I'd say if you look at the example of cost accounting, you can imagine a world where nobody does cost accounting. You say of your organization, "Let's do cost accounting here."
>
> That's a problem because you'd be heard as saying, "Somebody around here is stealing and we need to find out who." So that might be discouraged.
>
> In a world where everybody else does cost accounting, you say, "Let's not do cost accounting here." That will be heard as saying, "Could we steal and just not talk about it?" which will also seem negative.
>
> Similarly, with prediction markets, you could imagine a world like ours where nobody does them, and then your proposing to do it will send a bad signal. You're basically saying, "People are bullshitting around here. We need to find out who and get to the truth."
>
> But in a world where everybody was doing it, it would be similarly hard not to do it. If every project with a deadline had a betting market and you say, "Let's not have a betting market on our project deadline," you'd be basically saying, "We're not going to make the deadline, folks. Can we just set that aside and not even talk about it?"

So pushing from this equilibrium, to the one where prediction markets are common, would improve our societies beliefs about just about everything that one could make a prediction market for.

Arbital (or something like it)

The pitch I heard for arbital went something like this...

[Please note that I am recalling conversations that I had back in 2016. This should not be taken as an authoritative summary of Arbital's vision or plans.]

> In the old days, it used to be that when people disagreed about a simple matter of fact, there was not much recourse for resolving the disagreement. If you were committed, you could go to a library and try to research the answer, but most people didn't have the scholarship skills, nor the inclination to do that. (As an example, if two people got into a fight about the origin of the phrase "loose cannon", in pre-internet days, they might argue about it for years.
>
> But Wikipedia changed that, because it made it easy to verify questions of settled fact. Now if you disagree about the origin of "loose cannon", you can just check Wikipedia (or in this case, Wikitonary). Wikipedia is reliable enough, and accessible enough, to be an authoritative source.
>
> Thus, Wikipedia narrowed the scope of things that people could confidently assert, without any foundation. Because if it was the sort of thing you could check on Wikipedia, your conversation partner could just check, and you would loose social points for appearing like a confident idiot.
>
> What Wikipedia did for settled facts, Arbital was aiming to do for still contentious topics.
>
> For instance, questions of macroeconomic policy are pretty hard, and still controversial: even professional economists disagree about what the best approach is. But the fact that the question is not yet *settled* is often taken as license to promulgate any old opinion, regardless of how economically sound it is. Even though we haven't solved macro, doesn't mean there aren't some distinctly *wrong* answers. Arbital was aiming to be an authoritative source on the state of the discussion about such not-yet-settled topics, to further narrow the space of claims that a person can confidently assert, because they know that if they say something inane, someone might refute them with the relevant Arbital page.

Now of course, setting this as your goal is one thing, and actually designing a mechanism that is able to do this is another. And Arbital did not, in fact, succeed. But if something like this could be made to work, that would be a substantial boon to high level decision making.

In deed, even just educational tools that make it much easier to understand complicated topics might be a major help, under the (possible?) model that part of the reason why politicians and other high-level decision makers produces far from optimal policy, is that it is too hard, or too time consuming, to make sense of the conflicting arguments about, say, economics.

Electoral Reform

My understanding is that part of the reason our government is apparently so dysfunctional is that the electoral system is biased toward polarization.

A case in point is gerymanderying, whereby districts are drawn in such a way that congressmen are all but guarantied to win general elections, which disenfranchises

voters, and polarizes both parties (because in order to keep your job, you only need to appeal to your base, not cater to citizens across the political spectrum).

Similarly, the first past the post system used in the United States gives rise to the spoiler effect, which penalizes third parties by increasing the odds that their *least* preferred candidate wins.

It seems like solving those underlying incentives problems would moderate law makers, which seems likely to produce saner outcomes.

Kick-starter / Free state project style platforms

Kickstarter is a solution to a class of collective action problems, funding the creation of products that many people would want, but no one person can afford to pay the upfront startup costs for.

It seems like there is a lot of room for collective action solutions like that to shine.

For instance, many scientists know that the statistical methods that they use are less than ideal, but it would be costly for their personal careers if they switched to better methods, while everyone else continued to use the old ones. To solve this, young grad students might all commit to abandon using p-values, so long as x% of their peers agree to do the same.

I want to collect as many ideas for Civilizational Sanity Interventions as I can. Does anyone else have other examples?

Public Static: What is Abstraction?

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Author's Note: Most of the posts in this sequence are essentially a log of work-in-progress. This post is intended as a more presentable ("public") and higher-confidence ("static") write-up of some formalizations of abstraction. Much of the material has appeared in other posts; the first two sections in particular are drawn almost verbatim from the opening "What is Abstraction?" post.

Let's start with a few examples (borrowed from here) to illustrate what we're talking about:

- We have a gas consisting of some huge number of particles. We throw away information about the particles themselves, instead keeping just a few summary statistics: average energy, number of particles, etc. We can then make highly precise predictions about things like e.g. pressure just based on the reduced information we've kept, without having to think about each individual particle. That reduced information is the "abstract layer" - the gas and its properties.
- We have a bunch of transistors and wires on a chip. We arrange them to perform some logical operation, like maybe a NAND gate. Then, we throw away information about the underlying details, and just treat it as an abstract logical NAND gate. Using just the abstract layer, we can make predictions about what outputs will result from what inputs. Note that there's some fuzziness - 0.01 V and 0.02 V are both treated as logical zero, and in rare cases there will be enough noise in the wires to get an incorrect output.
- I tell my friend that I'm going to play tennis. I have ignored a huge amount of information about the details of the activity - where, when, what racket, what ball, with whom, all the distributions of every microscopic particle involved - yet my friend can still make some reliable predictions based on the abstract information I've provided.
- When we abstract formulas like "1+1=2*1" and "2+2=2*2" into "n+n=2*n", we're obviously throwing out information about the value of n, while still making whatever predictions we can given the information we kept. This is what abstraction is all about in math and programming: throw out as much information as you can, while still maintaining the core "prediction" - i.e. the theorem or algorithm.
- I have a street map of New York City. The map throws out lots of info about the physical streets: street width, potholes, power lines and water mains, building facades, signs and stoplights, etc. But for many questions about distance or reachability on the physical city streets, I can translate the question into a query on the map. My query on the map will return reliable predictions about the physical streets, even though the map has thrown out lots of info.

The general pattern: there's some ground-level "concrete" model (or territory), and an abstract model (or map). The abstract model throws away or ignores information from the concrete model, but in such a way that we can still make reliable predictions about some aspects of the underlying system.

Notice that the predictions of the abstract models, in most of these examples, are not perfectly accurate. We're not dealing with the sort of "abstraction" we see in e.g. programming or algebra, where everything is exact. There are going to be probabilities involved.

In the language of embedded world-models, we're talking about multi-level models: models which contain both a notion of "table", and of all the pieces from which the table is built, and of all the atoms from which the pieces are built. We want to be able to use predictions from one level at other levels (e.g. predict bulk material properties from microscopic structure and/or macroscopic measurements, or predict from material properties whether it's safe to sit on the table), and we want to move between levels consistently.

Formalization: Starting Point

To repeat the intuitive idea: an abstract model throws away or ignores information from the concrete model, but in such a way that we can still make reliable predictions about some aspects of the underlying system.

So to formalize abstraction, we first need some way to specify which "aspects of the underlying system" we wish to predict, and what form the predictions take. The obvious starting point for predictions is probability distributions. Given that our predictions are probability distributions, the natural way to specify which aspects of the system we care about is via a set of events or logic statements for which we calculate probabilities. We'll be agnostic about the exact types for now, and just call these "queries".

That leads to a rough construction. We start with some low-level model M^L and a set of queries Q. From these, we construct a minimal high-level model M^H by keeping exactly the information relevant to the queries, and throwing away all other information. By the minimal map theorems, we can represent M^H directly by the full set of probabilities $P[Q|M^L]$; M^H and $P[Q|M^L]$ contain exactly the same information. Of course, in practical examples,

the probabilities $P[Q|M^L]$ will usually have some more compact representation, and M^H will usually contain some extraneous information as well.

To illustrate a bit, let's identify the low-level model, class of queries, and high-level model for a few of the examples from earlier.

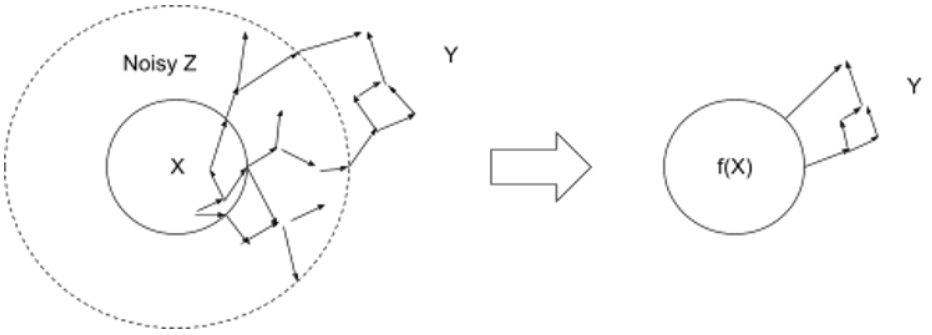
- Ideal Gas:
  - Low-level model M^L is the full set of molecules, their interaction forces, and a distribution representing our knowledge about their initial configuration.
  - Class of queries Q consists of combinations of macroscopic measurements, e.g. one query might be "pressure = 12 torr & volume = 1 m^3 & temperature = 110 K".
  - For an ideal gas, the high-level model M^H can be represented by e.g. temperature, number of particles (of each type if the gas is mixed), and container volume. Given these values and assuming a near-equilibrium initial configuration distribution, we can predict the other macroscopic measurables in the queries (e.g. pressure).
- Tennis:
  - Low-level model M^L is the full microscopic configuration of me and the physical world around me as I play tennis (or whatever else I do).
  - Class of queries Q is hard to sharply define at this point, but includes things like "John will answer his cell phone in the next hour", "John will hold a racket and hit a fuzzy ball in the next hour", "John will play Civ for the next hour", etc - all the things whose probabilities change on hearing that I'm going to play tennis.
  - High-level model M^H is just the sentence "I am going to play tennis".
- Street Map:
  - Low-level model M^L is the physical city streets
  - Class of queries Q includes things like "shortest path from Times Square to Central Park starts by following Broadway", "distance between the Met and the Hudson is less than 1 mile", etc - all the things we can deduce from a street map.
  - High-level model M^H is the map. Note that the physical map also includes some extraneous information, e.g. the positions of all the individual atoms in the piece of paper/smartphone.

Already with the second two examples there seems to be some "cheating" going on in the model definition: we just define the query class as all the events/logic statements whose probabilities change based on the information in the map. But if we can do that, then anything can be a "high-level map" of any "low-level territory", with the queries Q taken to be the events/statements about the territory which the map actually has some information about - not a very useful definition!

Information About Things “Far Away”

In order for abstraction to actually be useful, we need some efficient way to know which queries the abstract model can accurately answer, without having to directly evaluate each query within the low-level model.

In practice, we usually seem to have a notion of which variables are “far apart”, in the sense that any interactions between the two are mediated by many in-between variables.



In this graphical model, interactions between the variables X *and the variables* Y *are mediated by the noisy variables* Z*. Abstraction throws out information from* X *which is wiped out by noise in* Z*, keeping only the information* $f(X)$ *relevant to* Y*.*

The mediating variables are noisy, so they wipe out most of the "fine-grained" information present in the variables of interest. We can therefore ignore that fine-grained information when making predictions about things far away. We just keep around whatever high-level signal makes it past the noise of mediating variables, and throw out everything else, so long as we're only asking questions about far-away variables.

An example: when I type "4+3" in a python shell, I think of that as adding two numbers, not as a bunch of continuous voltages driving electric fields and current flows in little patches of metal and doped silicon. Why? Because, if I'm thinking about what will show up on my monitor after I type "4+3" and hit enter, then the exact voltages and current flows on the CPU are not relevant. This remains true even if I'm thinking about the voltages driving individual pixels in my monitor - even at a fairly low level, the exact voltages in the arithmetic-logic unit on the CPU aren't relevant to anything more than a few microns away - except for the high-level information contained in the "numbers" passed in and out. Information about exact voltages in specific wires is quickly wiped out by noise within the chip.

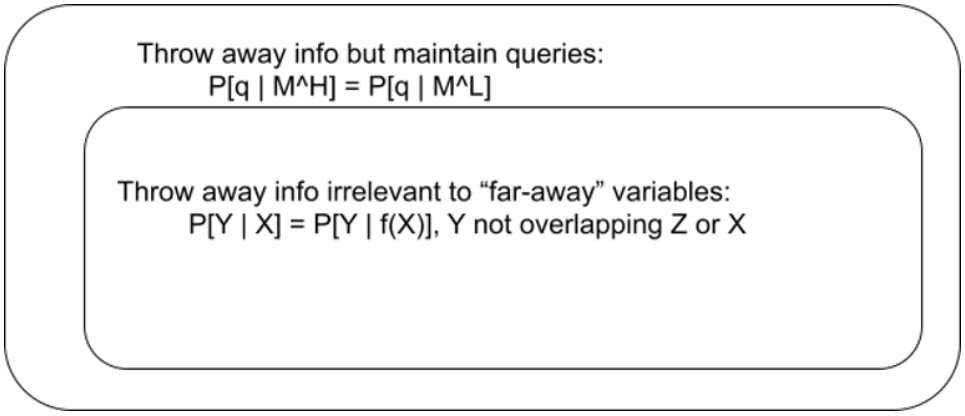
Another example: if I'm an astronomer predicting the trajectory of the sun, then I'm presumably going to treat other stars as point-masses. At such long distances, the exact mass distribution within the star doesn't really matter - except for the high-level information contained in the total mass, momentum and center-of-mass location.

Formalizing this in the same language as the previous section:

- We have some variables X and Y in the low-level model.
- Interactions between X and Y are mediated by noisy variables Z.
- Noise in Z wipes out most fine-grained information about X, so only the high-level summary $f(X)$ is relevant to Y.

Mathematically: $P[Y|X] = P[Y|f(X)]$ for any Y which is "not too close" to X - i.e. any Y which do not overlap with Z (or with X itself). Our high-level model replaces X with $f(X)$, and our set of valid queries Q is the whole joint distribution of Y given X.

Now that we have two definitions, it's time to start the Venn diagram of definitions of abstraction.

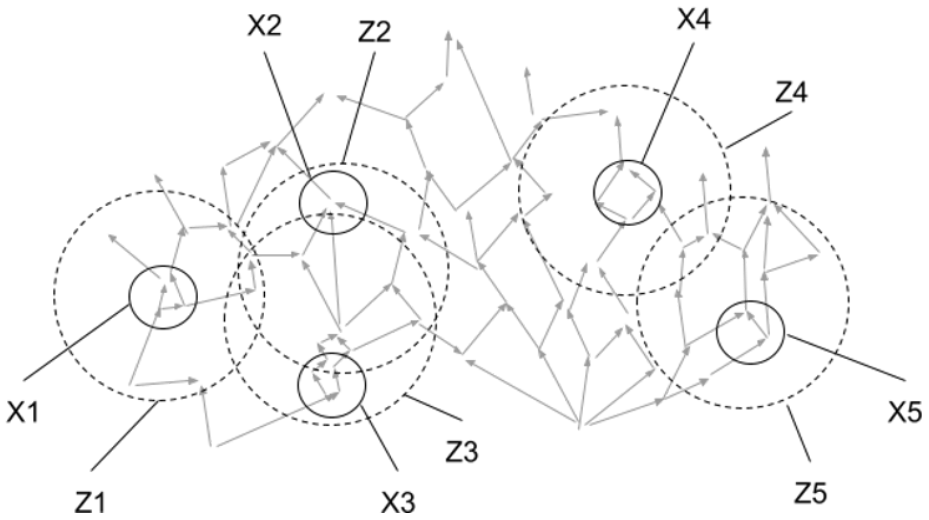


So far, we have:

- A high-level model throws out information from a low-level model in such a way that some set of queries can still be answered correctly: $P[q|M^H] = P[q|M^L] \; \forall q \in Q$.
- A high-level model throws out information from some variable X in such a way that all information about "far away" variables Y is kept: $P[Y|X] = P[Y|f(X)] \; \forall Y : Y \cap (X \cup Z) = \emptyset$.

Systems View

The definition in the previous section just focuses on abstracting a single variable X. In practice, we often want to take a system-level view, abstracting a whole bunch of low-level variables (or sets of low-level variables) all at once. This doesn't involve changing the previous definition, just applying it to many variables in parallel.



We have multiple non-overlapping sets of low-level variables $X_1 \ldots X_5$*, each with a set of "nearby" variables* $Z_1 \ldots Z_5$*. Abstraction will only retain information from each* X_i *relevant to* X_j*'s which do not overlap the corresponding* Z_i*. In particular, this means queries* $P[X_j | X_i]$ *will only be maintained by the abstraction if* X_j *is not "close to"* X_i *- i.e. if* X_j *does not overlap* Z_i*. In the notation below, these* X_i *are called* X_i^L*, to remind that they are the low-level variables.*

Rather than just one variable of interest X, we have many low-level variables (or non-overlapping sets of variables) $X_1^L \ldots X_n^L$ and their high-level summaries $X_1^H := f_1(X_1^L) \ldots X_n^H := f_n(X_n^L)$. For each of the X_i^L, we have some set Z_i of variables "nearby" X_i^L, which mediate its interactions with everything else. Our "far-away" variables Y are now any far-away X's, so we want

$$P[X_T^L | X_S^L] = P[X_T^L | f_S(X_S^L)]$$

for any sets of indices S and T which are "far apart" - meaning that X_T^L does not overlap any X_S^L or Z_S.

(Notation: I will use lower-case indices like X_i for individual variables, and upper-case indices like X_S to represent sets of variables. I will also treat any single index interchangeably with the set containing just that index.)

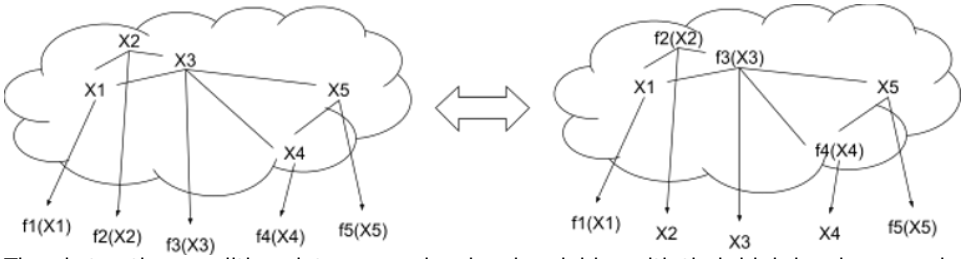
For instance, if we're thinking about wires and transistors on a CPU, we might look at separate chunks of circuitry. Voltages in each chunk of circuitry are X_i^L, and X_i^H summarizes the binary voltage values. Z_i are voltages in any components physically close to chunk i on the chip. Anything physically far away on the chip will depend only on the binary voltage values in the components, not on the exact voltages.

The main upshot of all this is that we can rewrite the math in a cleaner way: as a (partial) factorization. Each of the low-level components are conditionally independent given the high-level summaries, so:

$$P[X_S^L, X_S^H] = P[X_S^H] \prod_{i \in S} P[X_i^L \mid X_i^H]$$

This condition only needs to hold when S picks out indices such that $\forall i \neq j \in S : Z_i \cap X_j = \emptyset$ (i.e. we pick out a subset S of the X_i^L's such that no two are "close together"). Note that we can pick any set of indices S which satisfies this condition - so we really have a whole family of factorizations of marginal distributions in which no two variables are "close together". See the appendix to this post for a proof of the formula.

In English: any set of low-level variables X_S^L which are all "far apart" are independent given their high-level summaries X_S^H. Intuitively, the picture looks like this:

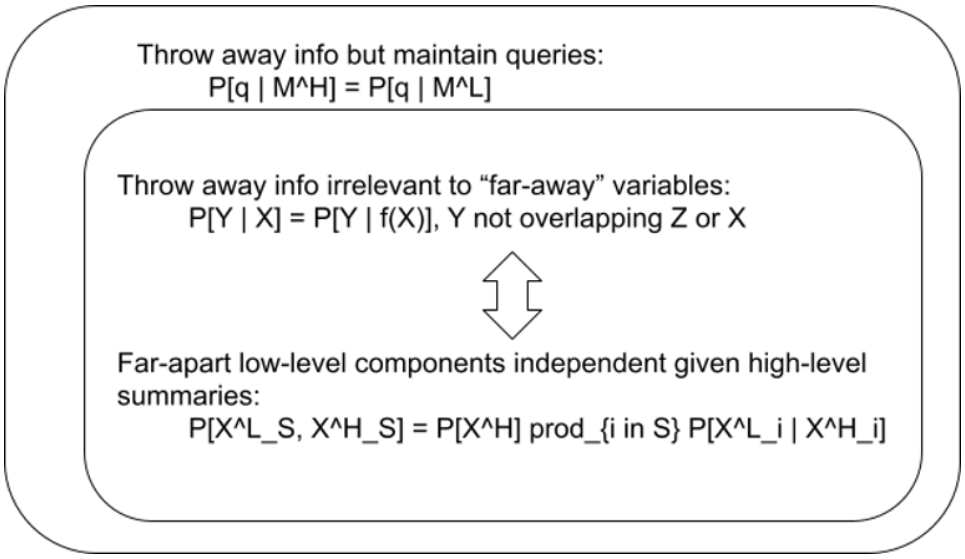


The abstraction conditions let us swap low-level variables with their high-level summaries, as long as all swapped variables and any query variables are all "far apart".

We pick some set of low-level variables $X_1^L \ldots X_k^L$ which are all far apart, and compute their summaries $X_1^H = f_1(X_1^L), \ldots, X_k^H = f_k(X_k^L)$. By construction, we have a model in which each of the high-level variables is a leaf in the graphical model, determined only by the corresponding low-level variables. But thanks to the abstraction condition, we can independently swap any subset of the summaries with their corresponding low-level variables - assuming that all of them are "far apart".

Returning to the digital circuit example: if we pick any subset of the wires and transistors on a chip, such that no two are too physically close together, then we expect that their exact voltages are roughly independent given the high-level summary of their digital values.

We'll add this to our Venn diagram as an equivalent formulation of the previous definition.



I have found this formulation to be the most useful starting point in most of my own thinking, and it will be the jumping-off point for our last two notions of abstraction in the next two sections.

Causality

So far we've only talked about "queries" on the joint distribution of variables. Another natural step is to introduce causal structure into the low-level model, and require interventional queries to hold on far apart variables.

There are some degrees of freedom in *which* interventional queries hold on far apart variables. One obvious answer is "all of them":

$$P[X_S^L, X_S^H \mid do(X_T^L = x_T^*)] = P[X_S^H \mid do(X_T^H = f_T(x_T)^*)] \prod_{i \in S} P[X_i^L \mid X_i^H]$$

... with the same conditions on S as before, plus the added condition that the indices in S and T also be far apart.

This is the usual requirement in math/programming abstraction, but it's too strong for many real-world applications. For instance, when thinking about fluid dynamics, we don't expect our abstractions to hold when all the molecules in a particular cell of space are pushed into the corner of that cell. Instead, we could weaken the low-level intervention to sample from low-level states compatible with the high-level intervention:

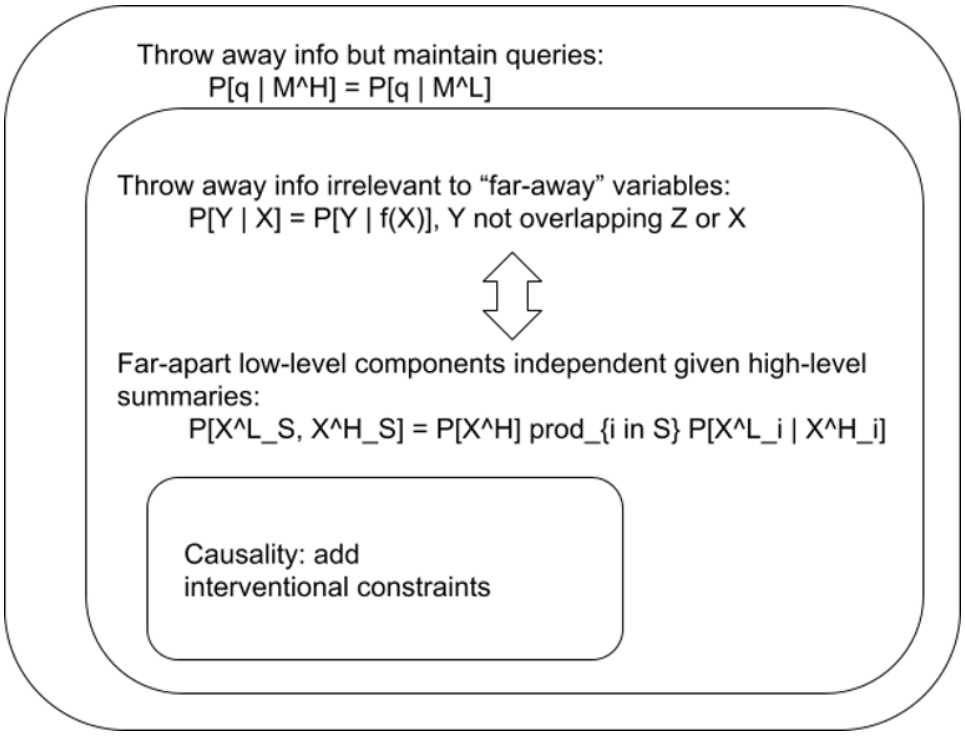
$$P[X_S^L, X_S^H \mid do(X_T^L = sample(x_T^* \mid f_T(x_T)^*))] = P[X_S^H \mid do(X_T^H = f_T(x_T)^*)] \prod_{i \in S} P[X_i^L \mid X_i^H]$$

We could even have low-level interventions sample from some entirely different distribution, to reflect e.g. a physical machine used to perform the interventions.

Another post will talk more about this, but it turns out that we can say quite a bit about causal abstraction while remaining agnostic to the details of the low-level interventions. Any of the above interventional query requirements have qualitatively-similar implications, though obviously some are stronger than others.

In day-to-day life, causal abstraction is arguably more common than non-causal. In fully deterministic problems, validity of interventional queries is essentially the only constraint (though often in guises which do not explicitly mention causality, e.g. functional behavior or logic). For instance, suppose I want to write a python function to sort a list. The only constraint is the abstract input/output behavior, i.e. the behavior of the designated "output" under interventions on the designated "inputs". The low-level details - i.e. the actual steps performed by the algorithm - are free to vary, so long as those high-level interventional constraints are satisfied.

This generalizes to other design/engineering problems: the desired behavior of a system is usually some abstract, high-level behavior under interventions. Low-level details are free to vary so long as the high-level constraints are satisfied.



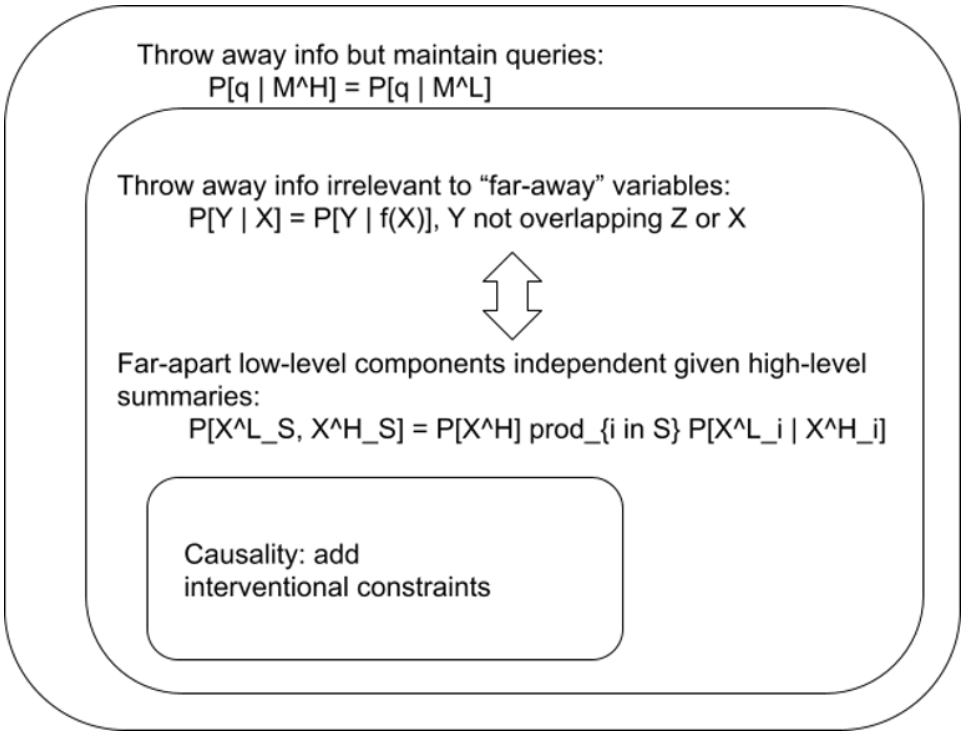
Exact Abstraction

Finally, one important special case. In math and programming, we typically use abstractions with sharper boundaries than most of those discussed here so far. Prototypical examples:

- A function in programming: behavior of everything outside the function is independent of the function’s internal variables, given a high-level summary containing only the function’s inputs and outputs. Same for private variables/methods of a class.
- Abstract algebra: many properties of mathematical objects hold independent of the internal details of the object, given certain high-level summary properties - e.g. the group axioms, or the ring axioms, or …
- Interfaces for abstract data structures: the internal organization of the data structure is irrelevant to external users, given the abstract "interface" - a high-level summary of the object's behavior under different inputs (a.k.a. different interventions).

In these cases, there’s no noisy intermediate variables, and no notion of “far away” variables. There’s just a hard boundary: the internal details of high-level abstract objects do not interact with things of interest “outside” the object except via the high-level summaries.

We can easily cast this as a special case of our earlier notion of abstraction: the set of noisy intermediate variables Z_i is empty. The “high-level summary” X_i^H of the low-level variables X_i^L contains all information relevant to any variables outside of X_i^L themselves.



Of course, exact abstraction overlaps quite a bit with causal abstraction. Exact abstractions in math/programming are typically deterministic, so they’re mainly constrained by interventional predictions rather than distributional predictions.

Summary

We started with a very general notion of abstraction: we take some low-level model and abstract it into a high-level model by throwing away information in such a way that we can still accurately answer some queries. This is extremely general, but in order to actually be useful, we need some efficient way to know which queries are and are not supported by the abstraction.

That brought us to our next definition: abstraction keeps information relevant to “far away” variables. We imagine that interactions between the variable-to-be-abstracted X and things far away are mediated by some noisy “nearby” variables Z, which wipe out most of the information in X. So, we can support all queries on things far away by keeping only a relatively small summary f(X).

Applying this definition to a whole system, rather than just one variable, we find a clean formulation: all sets of far-apart low-level variables are independent given the corresponding high-level summaries.

Next, we extended this to causal abstraction by requiring that interventional queries also be supported.

Finally, we briefly mentioned the special case in which there are no noisy intermediate variables, so the abstraction boundary is sharp: there’s just the variables to be abstracted, and everything outside of them. This is the usual notion of abstraction in math and programming.

Appendix: System Formulation Proof

We start with two pieces. By construction, X_i^H is calculated entirely from X_i^L, so

$P[X^L, X^H] = P[X^L] \prod_i P[X_i^H | X_i^L]$ (construction)

... without any restriction on which subsets of the variables we look at. Then we also have the actual abstraction condition

$$P[X_T^L|X_S^L] = P[X_T^L|X_S^H] \text{ (abstraction)}$$

... as long as X_T^L does not overlap Z_S or X_S^L.

We want to show that

$$P[X_T^L, X_T^H] = P[X_T^H] \prod_{i \in T} P[X_i^L | X_i^H]$$

... for T any set of non-nearby variables (i.e. $\forall i, j \in T : X_j \cap (Z_i \cup X_i) = \emptyset$). In English: sets of far-apart low-level variables are independent given their high-level counterparts.

Let's start with definitions of "far-apart" and "nearby", so we don't have to write them out every time:

- Two sets of indices S and T are "far apart" if X_S and Z_S do not overlap X_T, and vice-versa. Individual indices can be treated as sets containing one element for purposes of this definition - so e.g. two indices or an index and a set of indices could be "far apart".
- Indices and/or sets of indices are "nearby" if they are not far apart.

As before, I will use capital letters for sets of indices and lower-case letters for individual indices, and I won't distinguish between a single index and the set containing just that index.

With that out of the way, we'll prove a lemma:

$$P[X_T^L, X_{T'}^H | X_S^L, X_{S'}^H] = P[X_T^L, X_{T'}^H | X_{S \cup S'}^H]$$

... for any S far apart from S', both far apart from T and T' (though T and T' need not be far apart from each other). This lets us swap high-level with low-level given variables as we wish, so long as they're all far apart from each other and from the query variables. Proof:

$$P[X_T^L, X_T^H | X_S^L] = P[X_T^L | X_S^L] P[X_T^H | X_T^L] \text{ (by construction)}$$

$$= P[X_T^L | X_S^H] P[X_T^H | X_T^L] \text{ (by abstraction)}$$

$$= P[X_T^L, X_T^H | X_S^H] \text{ (by construction)}$$

By taking $T \leftarrow (T \cup T')$ and then marginalizing out unused variables, this becomes

$$P[X_T^L, X_{T'}^H | X_S^L] = P[X_T^L, X_{T'}^H | X_S^H]$$

That's the first half of our lemma. Other half:

$$P[X_T^L, X_{T'}^H | X_S^L, X_{S'}^H] = P[X_{S'}^H | X_S^L]^{-1} P[X_T^L, X_{T'}^H, X_{S'}^H | X_S^L] \text{ (by Bayes)}$$

$$= P[X_{S'}^H | X_S^H]^{-1} P[X_T^L, X_{T'}^H, X_{S'}^H | X_S^H] \text{ (by first half)}$$

$= P[X_T^{L}, X_{T'}^{H} | X_{S \cup S'}^{H}]$ (by Bayes)

That takes care of the lemma.

Armed with the lemma, we can finish the main proof by iterating through the variables inductively:

$P[X_{T \cup i}^{L}, X_{T \cup i}^{H} | X_S^{H}] = P[X_i^{L}, X_i^{H} | X_S^{H}] P[X_T^{L}, X_T^{H} | X_S^{H}, X_i^{L}, X_i^{H}]$ (by Bayes)

$= P[X_i^{H} | X_i^{L}] P[X_i^{L} | X_S^{H}] P[X_T^{L}, X_T^{H} | X_S^{L}, X_i^{H}]$ (by construction)

$= P[X_i^{H} | X_i^{L}] P[X_i^{L} | X_S^{H}] P[X_T^{L}, X_T^{H} | X_{S \cup i}^{H}]$ (by lemma)

$= (P[X_i^{L} | X_i^{H}] \frac{P[X_i^{H}]}{P[X_i^{L}]}) (P[X_S^{H} | X_i^{L}] \frac{P[X_S^{H}]}{P[X_S^{L}]}) P[X_T^{L}, X_T^{H} | X_{S \cup i}^{H}]$ (by Bayes)

$= P[X_i^{L} | X_i^{H}] (P[X_S^{H} | X_i^{L}] \frac{P[X_S^{H}]}{P[X_S^{L}]}) P[X_T^{L}, X_T^{H} | X_{S \cup i}^{H}]$ (by lemma & cancellation)

$= P[X_i^{L} | X_i^{H}] P[X_i^{H} | X_S^{H}] P[X_T^{L}, X_T^{H} | X_{S \cup i}^{H}]$ (by Bayes)

Here i, S, and T are all far apart. Starting with empty S and applying this formula to each variable i, one-by-one, completes the proof.

The Skill of Noticing Emotions

The Skill of Noticing Emotions

(Thanks to Eli Tyre and Luke Raskopf for helping teach me the technique. And thanks to Nora Ammann, Fin Moorhouse, Ben Laurense, Daniel Hynk, Nathan Young, Toby Jolly, Michael Ng, James Walsh and Jaime Sevilla for feedback on various drafts!)

(For those familiar with the core idea, you might find it more interesting to skip to the long list of personal examples in the Appendix)

Introduction

I was introduced to a technique called Noticing at my CFAR workshop back in October, to become better aware of my emotions and better at productively reacting to them. I've since found this a really powerful technique, and it's become a pretty key tool for solving problems in my life. My goal in this post is to give my take on Noticing, explain how I actually go about using it, and hopefully convince you to give it a try!

What do I mean by Noticing? I first want to introduce the idea of an emotion or mental state being **noticeable**: something that feels important and urgent and is immediately promoted to conscious attention. An excellent example of a noticeable experience is hearing my name. When I hear someone say "Neel" it immediately feels important and captures my attention. Even if I'm busy and focused on something, hearing my name can easily break that focus and cause me to change what I'm doing.

Being noticeable breaks down into two parts: something that I'm **aware** of and something that feels **important**. For example, I'm often aware that I'm procrastinating, but can't muster the motivation to actually do anything about it. It's something I am aware of, but not something that feels *important*. I like to think of this as spending most of my life on autopilot, focused on the present moment. A noticeable event or emotion will turn off autopilot and engage my conscious mind. Hearing my name does disable autopilot, but if I'm aware that I'm procrastinating and continuing to do so, I'm still on autopilot.

Noticing, then, is a technique to take a specific emotion or mental state[1] and deliberately make it noticeable. This essentially installs automatic triggers to turn off my autopilot. This is extremely useful, because a lot of problems in my life dissolve when I can just be more self-aware at the right times! For example, I've had a lot of success with noticing the feeling of defensiveness in arguments. By default, I'll often lash out and stop arguing in good faith when defensive. But by making the feeling of defensiveness noticeable, I can recognise in the moment where the urge to lash out comes from. This doesn't stop me becoming defensive, but once I've identified the problem I can take steps like removing myself from the situation.

Why should you care about noticing?

In general, fixing a problem involves figuring out the right thing to do, the right time to do it, and then actually doing it in the moment. And I often find it easy to do the first two, but then forget to do the right thing in the moment. One symptom of this is that I find it *much* easier to give other people advice than to apply it myself. For example, it's easy to understand a bias like the planning fallacy in the abstract, and even to identify it in friends. But it's *much* harder to notice in the moment when I'm falling prey to it and to put in the effort to correct for this. I find Noticing valuable for bridging this gap, and helping remind myself in the moment to actually apply the ideas I already understand. I spend much of my life on autopilot, focused on what I'm doing, and Noticing helps me be self-aware at the times when it's most useful to be.

A concrete example of how I've used this: While working, unrelated ideas often pop into my head and I feel an urge to explore them. Eg an impulse to check my messages or google a random fact. This always feels justified in the moment, and like it'll be a quick detour. But in practice it's incredibly hard to stop procrastinating, often starting a half hour rabbit hole that leaves my original task long forgotten. This makes it near impossible to maintain focus and deep work. With Noticing, I've been able to make the feeling of those fleeting urges feel noticeable. And by becoming self-aware at the moment of distraction, I can recognise that the urge just *feels* important, rather than actually mattering. This doesn't work perfectly, but it triggers several times a day, and has made it *much* easier to focus.

I've found a pretty wide range of things can be improved by better control of my autopilot. A few general categories:

- I find it easy to get caught up in mental loops, like procrastinating by scrolling on Reddit, or falling into spirals of insecurity. Noticing can help me to recognise the loop as it's beginning, and to break it then
- This has been a *really* powerful application - I find these loops build a lot of momentum and take a lot of willpower to break once they've begun. Breaking them before they really begin is much easier.
- To increase my awareness of positive emotions, eg gratitude.
- To better use and track my intuitions, eg noticing what's confusing me while learning.
- To build better social habits, eg paraphrasing back what the other person is saying, being less judgemental

Further, by practicing Noticing for a while I've found myself developing a more general skill of being self-aware. I've found it notably easier over time to be aware of how I'm behaving and why, and better at tracking my emotions. And it's much easier to start noticing something new!

How to apply it

I've hopefully convinced you that Noticing is useful, so how to actually use it? My approach is to install a mental reflex, where every time I feel the mental state I take a simple physical action, like snapping my fingers. I call these actions **markers**, and their purpose is to take the experience of Noticing outside of my head and make it harder to ignore.

I'll now outline the exact algorithm I use to install these reflexes[2], but in general I expect this to be a pretty personal process. I expect the best approach to vary a lot

between people, so I encourage you to adapt it to whatever feels most natural and helpful!

Algorithm

1. Choose the mental state
    1. This should be as specific as possible. It doesn't have to be easy to put into words, but should be eg one you have clear memories of
    2. Normally I first identify a problem which could be resolved by being self-aware at the right times, and then identify a relevant mental state. Good prompts:
        1. When do I think "I should have known better"?
        2. What do I often regret?
    3. A good litmus test for whether Noticing fits your problem: "If I could set an alarm to go off in my head at the right time, would this problem feel solved?"
2. Choose your **marker** action
    1. This should be a small and subtle physical action. It's important that it's something you can always do, eg snapping your fingers or tapping your foot
3. List 10 previous examples where you've felt this mental state
    1. Ideally ones that are recent, and that feel visceral - where they've really stuck in your mind
    2. 10 is an arbitrary number, the point is that more examples are always better. I recommend setting a 5 minute timer and spending the full 5 minutes brainstorming. It's easy to list examples off the top of your head and then feel stuck, but it's surprising how many more examples you can find with more time.
4. Mentally simulate each example, and take the action when you feel the emotion
    1. Try to really relive the experience, in as much visceral detail as possible. The goal is that it's something you feel rather than something you're describing
    2. Add as many details as possible to flesh out the scene
        1. Eg what were you saying? What did it feel like? Where were you? What could you see?
    3. Look out for cues associated with the state, eg physical sensations, thought patterns you have, common contexts
5. Actively practice this over the next 2 weeks - I call this the **learning period**
    1. Keep it in the back of your mind that you're practicing Noticing on this mental state
    2. Leave yourself regular reminders, eg post-it notes next to your bathroom mirror, daily email reminders, a list you check as part of your morning routine
    3. It can be helpful to track the times you successfully notice over this period, eg incrementing a counter, or writing it down

Motivation behind the Algorithm

The following is my model for why this algorithm works. All models are wrong, and so this is almost certainly incorrect in important ways, but I find this useful for motivating

the algorithm. When tweaking the algorithm, I think it's more important to keep to the spirit of this model than to keep to the letter of the algorithm.

When I'm focused on a task, on autopilot, most of my conscious attention is going towards it. But some part of my subconscious mind, my **awareness**, is still aware of what's going on around me. There's always going to be a lot of unimportant background stimuli, so my awareness will ignore things by default. But it needs to be able to promote important things to my attention, so it has a list of a few important things. And when it detects something with a strong association to something important, it flags that in my conscious attention. This is the experience of something being noticeable. So, to make a mental state noticeable, I need to make it **feel important** and to create **strong associations** with it.

How does the algorithm actually help with this?

1. Choosing as specific a mental state as possible is valuable, because if it feels clear and concrete in my mind, the associations will be stronger. There's a clear target to latch on to.
2. The marker action is useful to help it actually feel important in the moment.
    1. It's important that it's physical, because it's easy to be somewhat aware of the emotion but for it not to feel important and be ignored. Just as I can be aware that I'm procrastinating but it doesn't feel important enough to be able to stop. Taking a *physical* action, even a simpler one, makes it much harder for my mind to implicitly ignore the emotion
        1. It's easy to skip this part, and think you can 'just notice' - I *highly* recommend having a physical marker
    2. The marker should be a simple action that takes minimal willpower and can always be performed. This makes it easier to build the reflex of *paying attention* to this mental state. The goal is to help your mind focus, rather than to directly solve the problem
3. Mentally simulating historical examples helps make it feel important, because my mind is extremely good at pattern spotting. These simulations give it a bunch of visceral data points of "when I feel this emotion, it's important and I react to it".
    1. It's normal for this to feel a bit over-the-top, you want to really drill in that this is a reflex. The point is to do it enough times for my subconscious mind to spot a pattern, rather than stopping when I feel like my conscious mind gets it.
4. When simulating, it's also useful to look for as many **cues** for the mental state as possible. Ie things that correlate with it, like physical sensations and associated contexts and emotions.
    1. This increases the probability that my awareness notices one of the associated cues, and flags it to my attention.
    2. I call this increasing the **surface area** on the emotion, I want to understand it and what it looks like in as much detail as possible.
5. The learning period is important because it's *hard* to build this artificial association that the emotion is important. In the short term, this association will be quite weak, and keeping it in the back of my mind helps me respond to it. As it becomes more familiar, the association becomes stronger and is more likely to stick in the long term.
    1. A useful framing: The default state of the world is that I will forget about all new habits I develop. This isn't something I can resolve by just "trying harder", I need to take action and create external reminders to help it stick long term

Tips

- Try to simulate the examples in as much detail as possible. The goal is to build surface area and familiarity
    - For me, it feels like a mental movie playing out in my mind
    - Pay attention to physical, emotional and sensory details
    - Recency helps a lot
- Solutions: The marker action is explicitly *not* supposed to be a solution to the problem, the goal is just to help you notice it better
    - The end goal is to end up noticing *and* solving the problems, but the bottleneck is noticing. Once I reliably become self-aware at the right time, the solution is often obvious. Separating noticing the problem and solving it makes it easier to achieve both in the long term.
    - In practice, I'll often have default reactions in mind for what to do once I've become self-aware. But it's important to give yourself the affordance to ignore those, and to make the marker action as low-effort as possible. To build an effective reflex, it needs to be something you can do without thinking, rather than feeling like a decision
- Choosing a specific state is harder than it first seems. My instinct is to look for a specific *word*, but often one word can refer to many different mental states and internal experiences
    - Eg, for me anxiety could mean:
        - Fear of consequences
        - Awareness of uncertainty
        - Concern on someone else's behalf
        - Fear of social judgement.
    - Often I notice during the examples stage that the emotion feels a bit fuzzy and hard to pin down. This normally means I'm not being specific enough
    - This is hard and often requires iteration, don't expect it to be immediately easy! I often start Noticing something new, realise that I didn't have a sufficiently specific state in mind, and need to start again.
- Pace yourself: I highly recommend only trying to Notice one thing at a time, especially when starting out. The *goal* is to build a robust habit, but it will be quite fragile at first, and practicing on multiple states makes each stick less well.
    - This can seem a bit frustrating, but I actually find it really exciting from the right perspective. Noticing is a great example of a Tortoise Skill: something I can train in the background while living my life normally. These aren't a big deal in the short term, but really build up in the long term
- For important problems, it's often not clear what the solution is, even if you became self-aware at the right time. Noticing can still be helpful there, because it can help you get surface area on the problem. By practicing Noticing in the moment, you can get a better idea of what the problem actually feels like in the moment and what goes through your head. This can be a valuable first step to help you figure out what a good solution could be
- Upkeep: I find sometimes Noticing sticks well at first, but fades within a month or two. It can be helpful to simulate a few more examples to top it up if I notice it fading
- Tally counters: I find it *super* useful to carry a tally counter in my pocket, and make my physical action to click the counter
    - This is a very visceral and hard to ignore action
    - Carrying the counter around is a good physical reminder that I'm practicing Noticing
- Orient positively:

    - Ultimately, the goal is to build a success spiral, and make Noticing something to feel excited about. I think this makes it stick much better, and is more fun to think about
    - But it's easy to, eg, realise an hour after the fact that I didn't notice the emotion I'm tracking, and feel guilty about this. This is unhelpful because it builds negative associations, making me much more likely to give up on the idea and stop tracking the emotion.
        - I find it helpful to reframe forgetting as another chance to practice, because I now have a new, visceral example to simulate!
    - I find it useful to dwell on the fact that Noticing is *hard* to make robust, and that any progress is exciting. It's extremely useful to notice an emotion half of the time, if the alternative is never!
- Some people find it easier to practice Noticing on sensory experiences, especially when starting out.
    - Eg the feeling of the breeze, the sound of birdsong, the glint of sunlight
    - I expect this could be worth trying if your emotions/mental experiences don't feel very visceral to you
- Phase shifts: Noticing, if done well, is a habit. Like most habits, it's very easy to lose when you have a big shift in your day-to-day life, eg going on holiday, starting a new job, moving, etc. If you have an upcoming phase shift, it's important to ensure you re-learn the habit afterwards!

Conclusion

Ultimately, I'm extremely excited about Noticing because a lot of my problems boil down to not being self-aware at the right time. I think the bottleneck for learning a lot of good mental habits is doing the right thing *in the moment*, and Noticing has given me an extremely powerful tool for actually doing this. If you're reading this and empathise with the kinds of problems I've outlined, I urge you to take some time to try it out yourself!

Noticing is a very general technique, so I've given a long list of specific ways it's worked for me in the Appendix. I find it easy to get excited about the potential of a new technique, but fail to come up with any clear direction for how to apply it, and end up forgetting about it. So hopefully this list can provide some inspiration for specific ways it could be useful to you.

I've found it especially interesting to practice Noticing in the longer term. It's felt like there's a general meta-cognitive skill of "being aware of how I'm thinking and what I'm feeling" that I've been developing. I've both found it much easier to begin noticing something new, and found it generally easier to be self-aware and pay attention to what I'm feeling. So if you feel excited about the idea of Noticing, I highly recommend just trying it out on something, even if it doesn't feel like a perfect fit. You aren't just solving that specific problem, you're training the general skill of self-awareness!

It's extremely easy to be Typical Mind Fallacy-ing when talking about techniques like this, and I expect different things work best for different people. I'd be extremely interested in hearing about other people's experiences with Noticing, or other effective tools for these kinds of problems!

Appendix: Personal Examples

A few notes:

- These are obviously super specific to how my mind works, and likely won't perfectly generalise
    - I've done my best to capture the relevant emotional state as the Trigger, but it's pretty difficult to articulate this kind of thing. The words I've put are often vague, but correspond to a specific state inside my head
- They range from having clear default actions to being more "turn off auto-pilot and figure out what to do next" (recorded as 'be self-aware')
    - I found when training that the bottleneck was mostly becoming self-aware at the right time, so the point was to focus on the marker action.
    - But it was useful to have a list of sensible next actions in mind, which I've listed as Reactions

Productivity

- Guilt-based motivation: I very frequently use guilt-based motivation on myself, which is pretty bad for my general happiness and intrinsic motivation.
    - **Trigger**: Using mental force on myself
    - **Reaction**: Take a break, "would it matter if I didn't do this?", visualise why I feel excited about the task, "how will doing this task make my future self happier?"
    - I feel extremely excited about this particular application, because I helped a friend try Noticing on this problem. And a few months later they seem to have essentially stopped feeling guilt-based motivation!
- Distraction:
    - **Trigger**: The sudden urge to do something (eg google something, check my messages)
    - **Reaction**: Often just being self-aware is enough to kill this - I realise it isn't important. If it's important, I'll make a Trello card for it then go back to work.
        - This combines very well with Trello's ability to easily add an item in a few keystroke
    - As I detailed earlier, I've found this *super* useful for maintaining Deep Work
- Learning more effectively:
    - **Trigger**: The feeling of confusion or lack of clarity when learning a new concept
    - **Reactions**: Thinking further on it, google "intuitive explanation of ___", asking someone for help, making a note to look into this later
    - Often while learning something new I'll be a bit confused, and notice after a while that I'm totally lost. Noticing helps me focus on exactly what's confusing me and detect this much earlier, and has made my learning much more efficient!
        - I think this is one of the most useful mental habits I've ever developed, and has definitely helped a lot with getting the most out of my degree!
- Not making progress: I find that I'll often be trying to do something, and be spending time inefficiently. And this isn't clear to me in the moment, but feels obvious after the fact

    - **Trigger:** Frustration, feeling caught in a loop, feeling stuck
    - **Reaction:** Open an empty notepad and write down what's currently in my head
    - This triggers in a *really* wide range of situations, it's triggered over 45 times in the 2 weeks since I started this habit.
        - Eg pursuing an ineffective idea, trying to brainstorm but really staring off into space, trying to work when my mind is focused on an unpleasant interaction, spending 10 minutes googling irrelevant details, etc.
        - It's had fun side-effects like becoming notably faster at doing maths problems. I recognise much earlier "this solution idea is fruitless and I should try something else".
    - Taking my thoughts out of my head and writing them down is a key component of actually solving the problem. This often takes my thoughts from a vague mess to something concrete, and the next step then feels clear
    - This is much less specific than some others - I expect I'd have struggled to learn it without a lot of earlier Noticing practice
- Procrastination:
    - **Trigger:** Aversion/putting something off (eg "I should do this some time")
    - **Reaction**: Be self-aware (and realise there's a good chance I'll never get round to this). Ask myself "Would I be surprised if I don't do this?" Add it to my to-do list
    - This has been *really* useful, I've found that 80% of "good ideas I forget about" fail at this step, but will happen if I can just get the ball rolling
- Pushing myself: I find that I have a strong completionist drive, and really hate leaving things unfinished. I'll often push myself to complete things, long past the point where I should have taken a break.
    - **Trigger**: The compulsion to complete something, pushing myself
    - **Reaction**: Be self-aware, take a break, closing the tab, "Does it matter if I don't finish this now?"
- Punctuality:
    - **Trigger**: Awareness of how long I have left before I need to leave
    - **Reaction**: "Would I be surprised if I was late for this?", just making myself get ready now, setting a timer for when to start getting ready
    - This is pretty idiosyncratic: I'm both late for everything, *and* very aware of time, but it doesn't feel *important*.

Social

- Understanding other people's explanations:
    - **Trigger**: The feeling of confusion or lack of clarity
    - **Reactions**: Asking a question, asking them for an example, paraphrasing back to them what they've said, expressing confusion and prompting them to elaborate
    - I think these habits have *majorly* improved my communication skills. I find this often triggers several times during a conversation, and that I frequently misunderstood the first time.
- Giving compliments:
    - **Trigger**: Appreciation/admiration/gratitude
    - **Reaction**: Thanking the other person, or making a note to message them later

    - This has been a really awesome one for my overall happiness! Just being more aware of gratitude feels great, and I better appreciate why my friends are awesome
    - I think giving sincere compliments more freely makes me much more pleasant to be around
- Having interesting conversations:
    - **Trigger**: Curiosity/excitement about what somebody is saying
    - **Reaction**: Asking a question about that point, prompting them for more
    - I find that often when someone says something, small parts of what they say is much more interesting to me than the rest. Focusing the conversation on that part (often recursively) is now my default approach to conversations
        - This has *massively* increased the number of interesting conversations I have!
        - It's especially effective when I've just met the person, as a strategy for going from small talk to something mutually interesting
- Being less judgemental:
    - **Trigger**: Annoyance/dismissiveness directed at someone else
    - **Reaction**: "What's a world where they're a good person and they acted this way?"
    - I find it pretty easy to feel frustrated at other people, and view them in a fairly un-nuanced, one-dimensional sense. A lot of this stems from instinctively being judgemental rather than empathetic.

Discussions

- Defensiveness in an argument:
    - **Trigger**: Feeling defensive - often associated with stress, heart beating faster, not feeling grounded
    - **Reactions**: Changing the topic, trying to paraphrase back their case, trying to steelman, disengaging from the conversation
- Being empathetic (Minds should make sense):
    - **Trigger**: Frustration/incredulity at someone's beliefs/position in an argument
    - **Reaction**: Reminding myself that other people's minds make sense from the inside, and that I'm probably missing something. "Assume they're correct, explain why?" Paraphrase back to them their point of view. Asking for clarification/posing hypotheticals.
    - This one is pretty generally applicable, and I think has improved my communication skills a lot!
        - This also triggers when I'm explaining something, and somebody asks a question that feels dumb. My default reaction is frustration, and to repeat the point. But I've found that being curious and trying to unpack why the question made sense to them can let me resolve the confusion much more effectively

Mental health

- Insecurities:
    - **Trigger**: The feeling of self-doubt

    - **Reaction**: Be self-aware, use the Outside View, ask a friend for calibration, look for past evidence
    - I find it pretty easy to be caught in loops of self-doubt and insecurity. Noticing helps me to kill the loops before they really get going
        - Surprisingly, just reminding myself “this is a cognitive bias, and I can’t trust my intuitions” is often enough to dissolve it.
- Mindless procrastination: I'll often get caught up in a loop of procrasting for a long time, and realise I stopped enjoying myself half an hour ago
    - **Trigger**: The feeling of strain/going through the motions, noticing more time has passed than I expected
    - **Reaction**: "Do I actually want to be doing this?", shutting my eyes, closing the tab, taking a break
    - The underlying problem feels pretty heavily related to the difference between wanting and liking
- Anxiety: I find it super easy to get caught in anxiety spirals, Noticing helps to break the loop at the start
    - **Trigger**: Anxiety
    - **Reaction**: Taking a break, going outside, meditating. Be self-aware
    - Just realising that it’s all in my head, and that the *feeling* of importance isn’t the same thing as actually being a big deal, both go a long way to resolving this.
- Paying attention to my emotions:
    - **Trigger**: Suppressing my emotions/trying to be in control
        - Unexpectedly, this is a really salient trigger for me
    - **Reaction**: Be self-aware. If it’s about somebody else, consider talking to them. Writing down what’s going through my head

---

1. I’m using mental state very broadly here, this technique can be used for a range of things: emotions (like guilt, anxiety), mindsets (like defensiveness), internal experiences (like ‘failing to plan’), sensory experiences (like ‘hearing birds chirping’). It’s all about finding common themes between events. ↩
2. This algorithm is pretty heavily based on CFAR’s Trigger-Action Pattern framework. I call these **Empty TAPs**, because the purpose of the action is just to highlight the trigger better, rather than having a direct purpose ↩

May Gwern.net newsletter (w/GPT-3 commentary)

This is a linkpost for https://www.gwern.net/newsletter/2020/05

Most reliable news sources?

I've longed assumed that the vast majority of what the news discusses is irrelevant in the long run, and not that well reported at that. But the world seems to be moving faster these days, and I have more of a sense that I want to know what happened this week, because it might impact what I *do* next week.

For the first time in my life, I have some inclination to follow current events as they happen. And I find that I don't really know how to do that.

What are the most reliable / least-politicized news sources?

In particular, I want a resource that I can refer to that will tell me what happened in the past 3 days, in as factual and unbiased a way as possible. I expect that I might have to do further research, to get context for the events. But to start, I want a place where I can go that will tell me *what happened*, with a minimum of narrativizing, political outrage, etc.

GPT-3 Fiction Samples

This is a linkpost for https://www.gwern.net/GPT-3

Turns Out Interruptions Are Bad, Who Knew?

I've been known to accuse people who say open offices are "fine with a few mitigations" of not paying attention to the cost of their mitigations. I believed they shrunk their thoughts down to the point that not much was lost from an interruption, at the cost of only being able to think the thoughts that fit in that interval. Any thought that would take too long to process could not be conceived of.

I've also been known to accuse people who advocate for deep, uninterrupted work without the distractions of social media of "not understanding how valuable social media is to me". And besides, my workflow works best with frequent breaks (that I choose the timing of) because I "background process".

Cough

I maintained this illusion until, inspired by a stupidly expensive device that only does one thing, I taped my old phone to a bluetooth keyboard* and began to write in offline mode. It was immediately a magical experience. It was so *quiet*. I could go on my porch and write and it was *quiet*. My thoughts got much larger because I wasn't subconsciously afraid I'd interrupt them. I began to feel angry at my laptop. Why did it insist on hurting me so much? Why couldn't it be pure like the offline phone/keyboard experience? Why couldn't I just create things?

[* I only found two bluetooth keyboards with an inlay for phones/tablets. The other one lacks a built in battery, and shipped with a broken key]

Locally, this lasted for about 10 minutes before the social media cravings kicked in. But that was enough. I deeply resented work for taking me away from my magic writing device and making everything so *noisy*.

Since I started, my desire for using the quiet device has waxed and waned. At first I thought this was reflective of some deep pathology, but after two weeks it looks a lot more like "sometimes the benefits of quiet outweighs the benefits of being able to look stuff up, sometimes they don't'". I've also changed how I interact on a connected device- I'm more likely to close Signal, less likely to open Twitter. This is less due to a utilitarian calculation of the costs and benefits of Twitter, and more that once I'm in a good state, I can notice how switching to Twitter is almost physically painful.

The problem is that I wasn't wrong that social media was genuinely very valuable to me, and that was before we were all locked inside. But I definitely was wrong that getting those benefits were costless, in a way very analogous to mistakes I accused others of making. I'm glad I have the information now, but I haven't figured out what to do with it yet.

What's Your Cognitive Algorithm?

Epistemic Status: I'm neither a neuroscientist nor an ML researcher, but am trying to figure out "what kinds of human thought are actually possible to replicate on silicon right now?".

Here's my best guess of how human cognition works. Please tear it apart!

When I looked at GPT-2 last year, I thought: "Huh, when I look at my own thought process... I could summarize most of what I'm doing as: 'predict the next thing I'd say using crude concept association, and then say it.'"

Meanwhile, Jeff Hawkins says "Every part of the neocortex is running the same algorithm", and it's looking like maybe brains aren't doing that complicated a set of things.

Am I just GPT-2?

This was an obvious question to ask, but I haven't seen anyone write it up the question in detail.

I asked around. One mathematician friend said "I agree most people are doing GPT-style thinking where they regurgitate and recombine concepts from their neighbors. But, you can't get civilization from *just* that. Some people need to have model-based thinking."

Another mathematician friend agreed and added: "Young math students who try to prove theorems often do it GPT-style – they ramble their way through a bunch of math buzzwords and try to assemble them without understanding the structure of how they fit together. But, actual math proofs *require* clear understanding. You can't just 'predict the next word'"

I agree there is something additional going on here that gets you to formal math proofs and building skyscrapers. But... I don't think it's all that *much* more.

This post has three parts:

- Lay out the cognitive algorithm I personally seem to be following
- Outline how I think that algorithm developed
- Work through some examples of how I do "advanced thinking" (i.e. the sort of thinking that might literally advance the sum of human knowledge), and doublecheck if there are any surprising elements

My algorithm, as I understand it

Even when I'm developing novel concepts, or thinking through effortful procedures, most of my thinking following the same basic algorithm of:

A. Find the Next "Good Enough" Thought

1. My subconscious finds some nearby concepts that are associated with my previous thought-chunk
2. If a "*good enough*" thought appears, think that thought, and then repeat. ("good enough" means "I feel optimistic about the next thought in the chain leading somewhere useful")
3. If a "*not obviously good enough*" thought appears, check the next few associated concepts and see if they're good enough.
4. If none of the nearby concepts seem good enough, either give up and switch topics, or conduct an *effortful search.* This usually involves feeling stuck for awhile, spending willpower or getting a headache. Eventually I either:
    - a) find a good enough concept for my next thought, and proceed
    - b) find a better search algorithm. (Still basically "find a good enough concept", except it's not actually going to help me directly. Instead, I'll think something like "make a list of possibly hypotheses", or "search on google", or "ask my friend who knows about X", and then begin doing that.)

B. Check for Badness

- While doing all this, there's a followup processing that's periodically checking "was a recent thought-chunk somehow bad?".
    - Is this sloppy thinking that a respected rationalist would give me a disapproving look for?
    - Is this thoughtcrime that my tribe punish me for thinking?
    - Does it "smell bad" somehow, such that if I built a whole series of concepts off of this next thought-chunk, the result would be a flimsy construction? (i.e. bad code smell)
- If the thought seems maybe-bad, start associating towards concepts that help crystalize whether it's bad, or fix the badness

There's a few other things going on – I'm storing concepts in working memory, and sometimes in *mood,* which shape which other concepts are easily accessible. I'm sometimes using concepts that initiate *chains*, where I'll think "oh, I'm supposed to do algebra here. What's the first step of algebra?" and then the first step associates to the second step. But these parts seem like something I wouldn't be too surprised if GPT-2 developed on its own, or some equivalent version of.

Almost all of that condenses down to "find nearby associated concepts" and "direct my attention to more distant associated concepts."

(My understanding of this is based on the Tuning Your Cognitive Algorithms exercise, where you solve problems mindfully, paying lots of attention to what your brain seems to be doing on the sub-second timescale)

How far removed from that is GPT-2?

First, I'm not making any claims about the exact structuring of the learning algorithm. My understanding is that there's a few different neural network architectures that are more optimal for different kinds of processing (i.e. convolutional nets for image processing).

Some people have responded to my "what if all thought boils down to simple associations?" questioning with "but, model based learning!". I agree that model

based learning is a thing, but it's not obvious to me that GPT-2 doesn't have it, at least to some degree.

Second, a key thing GPT-2 is missing is the "check for badness" aspect. After predicting a word, AFAIK there's nothing that later punishes GPT-2 for thinking sloppily, or rewards it for doing something particularly great, which means it can't learn things like "You're supposed to generate multiple hypotheses before getting attached to the first one" and then deliberately apply them.

It probably also takes longer to learn things. (I don't actually know for sure how either GPT-2 or other leading language generators are rewarded. Has anyone done anything like "Train a neural net on Reddit, where it's somehow separately rewarded for predicting the next word, and also for predicting how much karma a cluster of words will get, and somehow propagating that back into the language generation?")

From Toddlers to Software Architects

How might the algorithm I described above develop in humans?

Step 1: Toddlers and Stoves

Toddlers have little longterm planning. If they see a bright red stove, they might think "shiny object!" and their little GPT processes think "what are some things that might come next?" and one of them is "touch the stove" and one of them is "look at it intently" and a third is "shrug and wander off". They pick "touch the stove" and then *OUCH.*

After a few iterations, they reach a point where, when they hypothesize "maybe the next action should be 'touch the stove'", they get a little flash of "but, two steps later, it will hurt, and that will be bad."

One way to conceive of this is "GPT-style, but you predict two words ahead instead of one."

But I don't think that's right. I think it's more like: "GPT style, but thinking certain thoughts brings up associations, and some associations just directly change the likely next actions. i.e. you think "touch the stove!" and then you think "ow!" and then, "ow" is treated as an incorrect end to the sentence of the narrative you're constructing. So you don't do the "touch stove" action.

Eventually this is cached into the System 1 GPT system such that "touch the stove" has a low predictive weight of "thing you might do", and it doesn't even come up any more.

Step 2: Toddlers and Mom Yelling

The first time Billy the Toddler came upon a hot stove, he reached out to touch it, and beforehand, Mom yelled "Billy don't touch that!"

And, possibly, Billy touched it anyway. And then he learned "ow!" and also learned that "Mom yells" is something that correlates with "ow!", which propagates back into his model of what sort of actions are good next-actions-to-take.

Or, previously, perhaps Billy had done some less immediately painful thing – perhaps walking into the street. Mom yells at him. He ignores her. A nearby car slows down, and doesn't hit him, so he doesn't learn "street ==> cars ==> bad". *But*, his Mom then runs and grabs him and pulls him away from the street, which is kinda painful. So he does gain the "Mom yelling ==> bad" association (as well as the "Walk into street" ==> "Mom Yelling" association).

Eventually Mom Yelling is treated as a failure state in the "predict which action I should take next" function.

Step 3: Internalized Mom Yelling

A couple years later, Billy periodically comes across novel situations – perhaps a wild animal in his backyard. This might remind him of similar situations where Mom Yelled in the past. By now, Billy doesn't need to hear Mom yell at him directly, he's able to think "Cool situation! Take Action" ==> "Hmm, Mom may yell at me later" ==> "Failure state" ==> "Okay, back up a step, take a different action."

And eventually this voice gets internalized as some kind of conscience/morality/guide, which doesn't even need to be physically present or temporally proximate to be relevant.

You *could* model this as "GPT style thinking, but predicting multiple steps down the line instead of just one or two." But, I think this doesn't match my internal experience. It's often *many* steps down the line that a bad thing would happen to me, that I need to avoid. More steps than I could feasibly be modeling.

I think the direct-association...

- Previous chunk: "notice dangerous situation"
- Next chunk: "association with mom yelling" (evaluates to "low predicted reward")

...is simpler to execute than:

- Previous chunk: "notice dangerous situation"
- Next chunk 1: "go play in dangerous situation"
- Next chunks 2 - 10: Do a bunch of steps in the dangerous situation"
- Chunk N: Mom eventually finds out and yells (evaluates to "low predicted reward")

Step 3: Internalized Punishment and Reward for Types of Cognition

Eventually Billy gains some internalized concept of "X is bad", which can be directly associated with various inputs.

Social Shame

For me, Doing X Would be Bad is often social shaped. For example, I often go to write some crappy code by randomly cobbling some things together. And then I get a visceral image of my coworkers complaining at me, saying "Ray! Use your brain! What would *good* code look like here?" and then I say "sigh... fine", and then I boot up the associations about what good code looks like and how to construct it.

Or, I'm debugging, and randomly changing things until it works or adding console.log statements to hope they'll reveal something obvious. And then my shoulder-angel-coworker pops up and says "Man, Ray, that is not how to debug code. Think!" and then I pull up my "what does actually debugging for real look like?" associations, and see what next-actions they pull up for me to consider, and then I do one of those.

(In this case, next-action-chunks include things like "make a list of what I know to be true about the code" and "check what the inputs to this function are and where they came from", which at my current skill level feel like atomic actions.)

Internalized Taste

A different way some people work (including myself in some domains) is less "social" and more "aesthetic." Good coders develop "bad code smell", and (I'd guess in the case of debugging) "bad habits smell", where if they find themselves debugging by randomly changing things, they think "obviously this is stupid and inefficient", and then seek out the associated next-actions that are more helpful.

Step 4: Strengthened "Good" Habits

Eventually, you streamline the steps of "notice a problem" => "consider a bad strategy to solve the problem" => "feel bad" => "find a good thing to do" => "do that", and go directly to "have a problem" => "use a good solution to the problem".

And this gets distilled into increasingly streamlined chunks, until master-level artisans do lots of sophisticated techniques that get bucketed into a single action.

But, really, what about deep planning and models and creativity?

As mentioned earlier, I do some complicated thought via chains-of-association. For example:

- Notice that I've been focusing on only one hypothesis
- Remember that I should be looking for alternate hypotheses
- Think "hmm, how do I get more hypotheses?"
- Start listing out half-remembered hypotheses that vaguely feel connected to the situation.
- Realize that listing random half-remembered hypotheses isn't a very good strategy
- Remember that a better strategy might be to make a list of all the facts I know about the phenomenon I'm investigating (without exactly remembering why I believe this)
- Make that list of facts (using either my working memory, or augmented memory like a notebook)
- For each relevant fact, ask myself "what must be true given this fact?" ("ah", I now think. "*This* is why it's useful to list out true facts. I can then conduct a search for other dependent facts, which builds up a useful web of associations")
- Use the list of facts and consequences to generate a new, more focused set of hypotheses

This *does* involve a mixture of System 1 and System 2 thinking (where system 2 involves slower, more laborious use of working memory and considering different options). But it's still mostly composed a bunch of atomic concepts.

Sarah Constantin's Distinctions in Types of Thought explores the possibility that deep neural nets basically have the "effortless System 1" type thinking, without being good at the slower, deliberate System 2 style thinking. I wouldn't be that surprised if GPT-2 was "only" a System 1. But I also wouldn't be that surprised if it naturally developed a System 2 when scaled up, and given more training. I also wouldn't be that surprised if it turned out not to *need* a System 2.

What's happening in System 2 thought?

The genre of this post is now going to abruptly switch to "Raemon narrates in realtime his thought process, as he tries to doublecheck what actually is going on in his brain."

Okay, I want to demonstrate System 2 thought. What are some examples of System 2 thought?

(Spins gears for a few minutes, unproductively)

Suddenly I remember "Ah, the bat/baseball problem is a classic System 2 problem." If I'm asked: "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?", what actually is going on in my head?

First, I think "Obviously the answer is '10c'"

Then, I think "Wait. no I know this problem, I think the answer is 5c" (determined via memory). But, I'm writing a blogpost right now where I'm trying to articulate what System 2 thought is like, and it would be helpful to have a real example. What if I rearranged the numbers in the problem and force myself to solve it again?

New Problem: "A bat and a ball cost $2.25 in total. The bat costs $2.00 more than the ball. How much does the ball cost?"

Great. Now the obvious answer is 25c, and that's probably wrong. Okay, how do I actually solve this? Okay, boot up my laborious arithmetic brain.

My first impulse is to subtract $2.00... wait that's literally the first thing my brain did.

Okay, what kind of math *is* this?

It's algebra I think?

X + Y = 2.25

X + (X + 2) = 2.25

2X + 2 = 2.25

X + 1 = 1.125

X = .125

Is that right? I think so. I don't actually care since the goal here was to doublecheck what my internal thought process is like, not get a right answer.

I notice that once I remembered to Do Actual Math, I mostly used quick associations rather than anything effortful, which feels more GPT-2 style. I think in elementary school those steps would have each been harder.

The more interesting part here was not the "how to solve the bat/baseball" part, but how to find an actual good example of System 2 thinking part. *That* felt quite effortful. I didn't have any immediate associations, so I was conducting a *search*, and moreover the search process wasn't very effective. (I think much of advanced cognition, and the Tuning Your Cognitive Algorithms process, is about figuring out what techniques enable you to more effectively search when you don't have an obvious search algorithm)

What Other Kinds of Thought Are There?

I notice this whole section is downstream of a *feeling*, where I have noticed that I haven't *actually* tried to comprehensively answer "is all of my thought processes explainable via predict-the-next-thing-then-do-it?". I have a nagging sense of "if I post this, there's a good chance someone will either poke a hole in it, or poke a hole in my generating thought process. Mama Rationalist is going to yell at me."

An obvious thing to do here is list out the types of advanced thinking I do, and then check how each one actually works by actually doing it.

I just did "elementary school math." What are some others?

1. Creatively combine existing concepts

I think this is how most novel ideas form. I think GPT-2 does very basic versions of this, but I haven't seen it do anything especially impressive.

One concrete algorithm I run is: "put two non-associated words next to each other, and see how they compile." An example of an upcoming blogpost generated this way is "Integrity Debt", which was born by literally taking the phrase "Technical Debt", swapping in "Integrity", and then checking "what does this concept mean, and is it useful?"

More often, I do a less intentional, fuzzier version of this where multiple concepts or sensory experiences get meshed together in my mind. Abram recounts a similar experience in Track Back Meditation

> At one point a couple of years ago, I noticed that I was using a particular visual analogy to think about something, which didn't seem like a very good analogy for what I was thinking about. I don't recall the example, but, let's say I was using a mental image of trees when thinking about matrix operations. I got annoyed at the useless imaginary trees, and wondered why I was imagining them. Then, I noticed that I was physically looking at a tree! This was fairly surprising to me. Some of the surprise was that I took a random object in my visual field to use for thinking about something unrelated, but more of the surprise was that I didn't immediately know this to be the case, even when I wondered why I was imagining trees.

> After I noticed this once, I started to notice it again and again: objects from my visual field end up in my imagination, and I often try to use them as visual analogies whether they're appropriate or not. It quickly became a familiar, rather than surprising, event. More interestingly, though, after a while it started to seem like a *conscious* event, rather than an automatic and uncontrollable one: I've become aware of the whole process from start to finish, and can intervene at any point if I wish.

2. Figure out what do do, for a problem where *none* of my existing associations are relevant enough to help solve it.

This is just actually pretty hard. Even Newton had to get hit on the head with the apple. Archimedes had to sit in the bathtub. Most human thought just isn't very original and incrementally advances using known associations.

I think most of the "good thought strategy" here involves figuring out efficient ways of exposing yourself to new concepts that might help. (This includes scholarship, getting a wider variety of life experience, and actually remembering to take relaxing baths from time to time)

I think there is a *teeny* fraction of this that looks like actually babbling entirely novel things at random (sometimes in a semi-directed fashion), and then seeing if they point in a useful direction. This is hard because life is high dimensional. Anyone who actually succeeds at this probably had to first develop a sense of taste that is capable of processing lots of details and get at least a rough sense of whether a sniff of an idea is promising.

3. Do advanced math that is on the edge of my current skills, where every step is novel.

I think this mostly involves repeatedly querying "what are the actual steps here?", and then applying some effortful directed search to remember the steps.

(I notice my shoulder Mathematician Friend just said "THAT'S NOT WHAT REAL MATH IS. REAL MATH IS BEAUTIFUL AND INVOLVES UNDERSTANDING CONCEPTS DEEPLY OR SOMETHING SOMETHING SOMETHING IDK MY SHOULDER MATHEMATICIAN ISN'T VERY HIGH RESOLUTION")

Speaking of which...

4. Mulling over concepts until I understand them deeply, have an epiphany, and experience a 'click'.

There's a particular sensation where I finally map out all the edges of a fuzzy concept, and then have a delightful moment when I realize I fully understand it. Prior to this, I have a distinctively uncomfortable feeling, like I'm in a murky swamp and I don't know how to get out.

I think this is the process by which a complicated, multi-chunk concept gets distilled into a single chunk, allowing it to take up less working memory.

5. Procedural Knowledge (i.e. Riding Bicycles)

I think this is actually mostly the same as #4, but the concepts are often differently shaped – physical awareness, emotional attunement, etc. It doesn't come with the

same particular "click" qualia that intellectual concepts have for me, but there is often a "suddenly this is easy and feels effortless" feeling.

How Do *You* Think?

So those are all the ways that I think, that I can easily remember. How do you do *your* thinking? Are there any pieces of it that I missed here? (I'm wondering how much variety there is in how people think, or experience thought, as well as whether I missed some of my own thinking-types)

Implications for AI

My actual motivation here was to get a sense of "*Are* there any major roadblocks for human level intelligence, or do we basically have all the pieces?" (While I wrote this post, a couple other posts came out that seemed to be exploring the same question)

My current sense is that all the most advanced thinking I do is made of simple parts, and it seems like most of those parts roughly correspond to facets of modern ML research. This wouldn't mean AGI is right around the corner – my understanding is that when you zoom into the details there's a fair amount of finicky bits. Many of the most interesting bits aren't using the same architecture, and integrating them into some kind of cohesive whole is a huge, multi-step project.

But, it seems like most of the remaining work is more like "keep plugging away at known problems, improve our ability to efficiently process large amounts of information, and incrementally improve our ability to learn in general ways from smaller bits of information".

This post doesn't really attempt to make the "AGI is near" case, because I'm not actually an ML researcher nor even a competent hobbyist. I think seriously investigating that is another post, written by someone who's been paying much closer attention than me.

For the immediate future, I'm interested in various LW contributors answering the question "How do you think you think, and why do you think you think that way?"

A Practical Guide to Conflict Resolution: Comprehension

[As described in the introduction, this post is about the value of fully grasping the whole picture, including the many sides to each debate, even when most of them are wrong.]

Seek First to Understand

A good attitude and a good venue will carry you a surprisingly long way, but of course they're not always sufficient on their own. The next thing I try when mediating conflict is to temporarily ignore whatever I believe and work to understand both sides of the argument equally. I called this post "comprehension", but it could equally just be called "listening", or maybe more precisely "active listening". Honestly, Stephen Covey has already said most of this far better than I can in his book *The 7 Habits of Highly Effective People*. Habit number five is "seek first to understand", which captures the spirit of things.

The value of truly understanding both sides of a conflict cannot be overstated. Even when I've nominally resolved a conflict, I get antsy if I still don't really grok one side or the other. More than once, trying to scratch that itch "after the fact" has turned up a hidden requirement or pain point which would have just caused more grief down the road. Remember, success is rarely as simple as just making the conflict go away; you can't know if you've truly found success (not just victory) unless you properly understand both sides.

But understanding both sides isn't just an after-the-fact thing. It also has concrete value in guiding the resolution of a conflict when you're caught in the middle of it, because it allows you to properly apply the principle of charity. The principle of charity says that you should try and find the best possible interpretation for people's arguments, even when they aren't always clear or coherent. It goes back to assuming good faith; maybe an argument sounds crazy, but *it makes sense to the person saying it*. The only way to apply the principle of charity in many cases is to start by understanding the argument, and the person making it.

Understanding both sides is also a key part of something called "steelmanning", which is the process of actively finding better versions of another person's arguments. This may seem like an odd thing to do in a conflict, but only if you've accidentally slipped back into the habit of aiming for victory instead of success. Assume good faith, and work with both sides to fully develop the points they're trying to make. Doing this brings clarity to the discussion which can often illuminate the crux of the conflict.

Of course sometimes being charitable is *hard*. People may make arguments which just seem… wrong. Crazy. Even harmful. (The topic of whether an argument can be harmful in and of itself is a fascinating one I don't have space for here. Whatever you believe, it isn't relevant to the point I'm trying to make). A lot of people would suggest that trying to understand or improve an argument like that is a waste of time, or even ethically wrong. I disagree. I believe that truly understanding both sides of a conflict is fundamentally valuable, no matter what that conflict is. It clarifies. It builds empathy.

It expands your knowledge of the world. And even if by the end you still deeply disagree, understanding the argument will let you articulate a better response.

Practical Tips

The principle is all well and good, but getting to that level of understanding in practice can be really hard. It's a skill that gets easier with repetition, so I would encourage you to practice it as much as possible, even for small conflicts where it might not seem necessary. Build that habit when it's easy, and you'll find that it becomes automatic even when it's hard. Still, if you're trying and you're really stuck, I've got a trick which helps me when I just can't seem to connect with what somebody is saying.

To better understand a different perspective, try splitting an argument up into the separate pieces of a problem and a solution. A lot of arguments fit into this pattern quite naturally, and I often find that while I couldn't quite grasp the argument as a whole, I both understand and even agree with the problem; it's the solution that's causing me issues. Even then, having the problem separated out and well-defined can lead me to understanding the solution too, because it frequently highlights some unstated premise which I wasn't aware of. This is also a great way to practice steelmanning, since making implied premises explicit is a great way to improve an argument; people are pretty bad at this by default.

I should also note that if this trick kind of works for a situation, but doesn't quite, you should try making the problem even more general. For example, if the argument is "Mexicans are taking our jobs, so we should stop immigration from Mexico", it's tempting to define the problem as just "Mexicans are taking our jobs", but it's probably more productive to define it as something like "something is taking our jobs" or even "our economic prospects suck". This pulls out an implied premise (that the cause is Mexican immigrants) which may be the real point of disagreement, but even apart from that, finding a problem which you can be sympathetic to is worth its weight in gold. With this kind of problem in hand, you can reframe the conflict as a cooperative mission, working together to find the best solution to the problem. You can start to look for success, not victory.

Standards of Understanding

It's often said that the real acid test for truly understanding somebody's argument is the ability to explain it back to them in a way they will agree with. This is good, and you should definitely aim for this (trying to explain it back is also a useful trick for conflict resolution in general), but sometimes I find it useful to use a slightly higher bar. I consider myself to really properly understand an argument when I can not only explain it to the person who made it, but can also explain (to myself, not to them) how they came to believe it. Both sides of the conflict are part of the universe, so to understand the universe you have to know how both sides came to be.

This may seem like an esoteric or excessively demanding standard, and it isn't necessary all the time. But there are interesting and practical sources of conflict where this is a really useful approach that provides a lot of insight. Religion is my favourite example of this; most theistic worldviews can pretty naturally explain the existence of atheists, but a lot of atheists have a hard time explaining the existence of theists. "People are dumb" may be emotionally satisfying, but doing the work of

constructing a real explanation builds a lot of empathy and ends up sharpening the resulting argument.

I’ve covered a lot of different ground in this section, but I think I can boil it down to four key points to take away:

1. Seek always to understand.
2. Actively look for the best version of everyone’s arguments.
3. Separate the problem and the solution.
4. To truly understand, you must explain how both sides came to be.

[AN #102]: Meta learning by GPT-3, and a list of full proposals for AI alignment

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Alignment Newsletter is a weekly publication with recent content relevant to AI alignment around the world. Find all Alignment Newsletter **resources here**. In particular, you can look through **this spreadsheet** of all summaries that have ever been in the newsletter.

Audio version **here** (may not be up yet).

SECTIONS

HIGHLIGHTS

Language Models are Few-Shot Learners *(Tom B. Brown et al)* (summarized by Rohin): The biggest **GPT-2 model** (**AN #46**) had 1.5 billion parameters, and since its release people have trained language models with up to 17 billion parameters. This paper reports GPT-3 results, where the largest model has *175 billion* parameters, a 10x increase over the previous largest language model. To get the obvious out of the way, it sets a new state of the art (SOTA) on zero-shot language modeling (evaluated only on Penn Tree Bank, as other evaluation sets were accidentally a part of their training set).

The primary focus of the paper is on analyzing the *few-shot learning* capabilities of GPT-3. In few-shot learning, after an initial training phase, at test time models are presented with a small number of examples of a new task, and then must execute that task for new inputs. Such problems are usually solved using *meta-learning* or *finetuning*, e.g. at test time **MAML** takes a few gradient steps on the new examples to produce a model finetuned for the test task. In contrast, the key hypothesis with GPT-3 is that language is so diverse, that doing well on it already requires adaptation to the

input, and so the learned language model will *already be a meta-learner*. This implies that they can simply "prime" the model with examples of a task they care about, and the model can *learn* what task is supposed to be performed, and then perform that task well.

For example, consider the task of generating a sentence using a newly made-up word whose meaning has been explained. In one notable example, the prompt for GPT-3 is:

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:

We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

Given this prompt, GPT-3 generates the following example sentence for "farduddle":

One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.

The paper tests on several downstream tasks for which benchmarks exist (e.g. question answering), and reports zero-shot, one-shot, and few-shot performance on all of them. On some tasks, the few-shot version sets a new SOTA, *despite not being finetuned using the benchmark's training set*; on others, GPT-3 lags considerably behind finetuning approaches.

The paper also consistently shows that few-shot performance increases as the number of parameters increases, and the rate of increase is faster than the corresponding rate for zero-shot performance. While they don't outright say it, we might take this as suggestive evidence that as models get larger, they are more incentivized to learn "general reasoning abilities".

The most striking example of this is in arithmetic, where the smallest 6 models (up to 6.7 billion parameters) have poor performance (< 20% on 2-digit addition), then the next model (13 billion parameters) jumps to > 50% on 2-digit addition and subtraction, and the final model (175 billion parameters) achieves > 80% on 3-digit addition and subtraction and a perfect 100% on 2-digit addition (all in the few-shot regime). They explicitly look for their test problems in the training set, and find very few examples, suggesting that the model really is learning "how to do addition"; further, when it is incorrect, it tends to make mistakes like "forgetting to carry a 1".

On broader impacts, the authors talk about potential misuse, fairness and bias concerns, and energy usage concerns; and say they about these issues what you'd expect. One interesting note: "To understand how low and mid-skill actors think about language models, we have been monitoring forums and chat groups where misinformation tactics, malware distribution, and computer fraud are frequently discussed." They find that while there was significant discussion of misuse, they found no successful deployments. They also consulted with professional threat analysts about the possibility of well-resourced actors misusing the model. According to the paper: "The assessment was that language models may not be worth investing significant resources in because there has been no convincing demonstration that current language models are significantly better than current methods for generating text, and because methods for "targeting" or "controlling" the content of language models are still at a very early stage."

Rohin's opinion: For a long time, I've heard people quietly hypothesizing that with a sufficient diversity of tasks, regular gradient descent could lead to general reasoning abilities allowing for quick adaptation to new tasks. This is a powerful demonstration of this hypothesis.

One **critique** is that GPT-3 still takes far too long to "identify" a task -- why does it need 50 examples of addition in order to figure out that what it should do is addition? Why isn't 1 sufficient? It's not like there are a bunch of other conceptions of "addition" that need to be disambiguated. I'm not sure what's going on mechanistically, but we can infer from the paper that as language models get larger, the number of examples needed to achieve a given level of performance goes down, so it seems like there is some "strength" of general reasoning ability that goes up (see also **this commentary**). Still, it would be really interesting to figure out mechanistically how the model is "reasoning".

This also provides some empirical evidence in support of the threat model underlying **inner alignment concerns** (**AN #58**): they are predicated on neural nets that implicitly learn to optimize. (To be clear, I think it provides empirical support for neural nets learning to "reason generally", not neural nets learning to implicitly "perform search" in pursuit of a "mesa objective" -- see also **Is the term mesa optimizer too narrow?** (**AN #78**).)

An overview of 11 proposals for building safe advanced AI *(Evan Hubinger)* (summarized by Rohin): This post describes eleven "full" AI alignment proposals (where the goal is to build a powerful, beneficial AI system using current techniques), and evaluates them on four axes:

1. **Outer alignment:** Would the optimal policy for the specified loss function be aligned with us? See also **this post**.

2. **Inner alignment:** Will the model that is *actually produced* by the training process be aligned with us?

3. **Training competitiveness:** Is this an efficient way to train a powerful AI system? More concretely, if one team had a "reasonable lead" over other teams, would they keep at least some of the lead if they used this algorithm?

4. **Performance competitiveness:** Will the trained model have good performance (relative to other models that could be trained)?

Seven of the eleven proposals are of the form "recursive outer alignment technique" plus "**technique for robustness** (**AN #81**)". The recursive outer alignment technique is either **debate** (**AN #5**), **recursive reward modeling** (**AN #34**), or some flavor of **amplification** (**AN #42**). The technique for robustness is either transparency tools to "peer inside the model", **relaxed adversarial training** (**AN #70**), or intermittent oversight by a competent supervisor. An additional two proposals are of the form "non-recursive outer alignment technique" plus "technique for robustness" -- the non-recursive techniques are vanilla reinforcement learning in a multiagent environment, and narrow reward learning.

Another proposal is Microscope AI, in which we train AI systems to simply understand vast quantities of data, and then by peering into the AI system we can learn the insights that the AI system learned, leading to a lot of value. We wouldn't have the AI system act in the world, thus eliminating a large swath of potential bad outcomes. Finally, we have STEM AI, where we try to build an AI system that operates in a

sandbox and is very good at science and engineering, but doesn't know much about humans. Intuitively, such a system would be very unlikely to deceive us (and probably would be incapable of doing so).

The post contains a lot of additional content that I didn't do justice to in this summary. In particular, I've said nothing about the analysis of each of these proposals on the four axes listed above; the full post talks about all 44 combinations.

Rohin's opinion: I'm glad this post exists: while most of the specific proposals could be found by patching together content spread across other blog posts, there was a severe lack of a single article laying out a full picture for even one proposal, let alone all eleven in this post.

I usually don't think about outer alignment as what happens with optimal policies, as assumed in this post -- when you're talking about loss functions *in the real world* (as I think this post is trying to do), *optimal* behavior can be weird and unintuitive, in ways that may not actually matter. For example, arguably for any loss function, the optimal policy is to hack the loss function so that it always outputs zero (or perhaps negative infinity).

TECHNICAL AI ALIGNMENT

MISCELLANEOUS (ALIGNMENT)

Planning with Uncertain Specifications *(Ankit Shah et al)* (summarized by Rohin): Suppose you recognize that there are no "certain specifications", and so infer a distribution over specifications. What do you then do with that distribution? This paper looks at this problem in the context where the specifications are given by formulas in linear temporal logic (which can express temporal non-Markovian constraints). They identify four possibilities:

1. *Most likely*: Plan with respect to the most likely specification.

2. *Most coverage*: Satisfying as many formulas as possible, ignoring their probability (as long as they have non-zero probability)

3. *Chance constrained*: Like the above, except you weight by probabilities, and drop the least likely formulas up to a parameter δ.

4. *Least regret*: Like the above, with δ set to zero.

Intuitively, the *Most likely* criterion won't be very robust since it is only taking one specification into account, *Most coverage* is aiming for maximum robustness, *Chance constrained* interpolates, where larger δ corresponds to trading robustness for gain in ability. This is exactly the pattern we see in a task where a robot must set a dinner table.

Rohin's opinion: Ultimately, I hope that in cases like this, the agent plans conservatively initially, but also tries to learn which specification is actually correct, allowing it to become more bold over time. Nonetheless, it seems quite difficult to do this well, and even then we likely will have this tradeoff between robustness and task

performance. This is the case with humans too: if you try to please everyone (robustness), you'll end up pleasing no one (task performance).

OTHER PROGRESS IN AI

REINFORCEMENT LEARNING

Suphx: Mastering Mahjong with Deep Reinforcement Learning *(Junjie Li et al)* (summarized by Rohin): Mahjong is a large imperfect information game with complex rules where turn order can be interrupted. This makes it challenging to solve with existing techniques like MCTS and counterfactual regret minimization. This paper details what was necessary to build *Suphx*, an AI system that is stronger than 99.99% of humans. Some highlights:

- Like the original AlphaGo, they first learned from human gameplay and then finetuned using reinforcement learning, with deep CNNs as their models. They learned both action models as well as value models. They added an entropy bonus to ensure that the policy remained stochastic enough to continue learning over the course of RL.

- They have *five* learned action models, corresponding to five different decisions that need to be made in Mahjong, as well as a rule-based system for deciding whether or not to declare a winning hand.

- To handle imperfect information, they first train an *oracle agent* that gets access to all information, and then slowly reduce the amount of information that it gets to observe.

- They could use search to improve the performance online, but did not do so in their evaluation (since Suphx was playing on a website with time constraints). Suphx with search would probably be significantly stronger.

Rohin's opinion: I am a bit curious how they removed observations from the oracle agent, given that you usually have to keep the structure of the input to a neural net constant. Perhaps they simply zeroed out the observations they didn't want?

Mastering Complex Control in MOBA Games with Deep Reinforcement Learning *(Deheng Ye et al)* (summarized by Rohin): This paper presents an AI system that can play the Multi-player Online Battle Arena (MOBA) game *Honor of Kings*. They are inspired by **OpenAI Five** (**AN #13**) (and Honor of Kings sounds quite similar to Dota, though it is 1v1 instead of 5v5), and have a similar learning setup: reinforcement learning using PPO. Their architecture requires an off-policy algorithm (I'm not sure why, maybe they have stale parameters across their rollout servers), so they add an importance sampling correction to the PPO objective, as well as an additional type of gradient clipping. The input is a combination of the image and underlying game state info. The resulting agents are able to beat top human players, and in an event with the public, the AI system lost only 4 out of 2100 matches. Unlike OpenAI Five, this required only around 100 hours to train (though it's unclear how much compute was used).

DEEP LEARNING

More Efficient NLP Model Pre-training with ELECTRA *(Kevin Clark et al)* (summarized by Flo): There are two main approaches to pretraining for NLP, language models (LMs) which iteratively predict the next word in a given incomplete sentence, and masked language models (MLMs), which predict the identities of a few masked words in an otherwise complete sentence. While not just looking at the previous words (bidirectionality) can be advantageous, MLMs only learn to predict the masked words, which reduces how much is learnt from a given sentence.

The authors present an alternative approach, ELECTRA, that outperforms RoBERTa while requiring less than a third of the compute. This is achieved by changing the form of the pretraining task from predicting words to discriminating fake words: Instead of masking, some words are replaced by words generated by an MLM and the trained model has to classify these as fake. This way, we get bidirectionality, but also a more dense signal, as the model has to produce an output for every single word, not just the masked ones. While this looks similar to GANs, the generator is only trained on the usual MLM loss and is not incentivized to fool the discriminator, as GANs don't seem to work well on sequence data.

Read more: Paper: ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

Flo's opinion: I found it a bit surprising that replacing word prediction with fake discrimination would help that much, but from the ablations, it seems like this is really mostly an instrument to get a loss signal for every single word, which is a cool idea. On a more zoomed-out perspective, results like this seem to show that gains in **algorithmic efficiency** (**AN #99**) are not fundamentally slowing down.

HIERARCHICAL RL

DADS: Unsupervised Reinforcement Learning for Skill Discovery *(Archit Sharma et al)* (summarized by Rohin): Reinforcement learning in robotics typically plans directly on low-level actions. However, it sure seems like there are a simple set of *primitives* like walking, running, shuffling, etc. that are inherent to the robot morphology. What if we could learn these primitives, and then plan using those primitives? This paper introduces a method for learning these primitives *without a reward function*. They simply optimize skills for *predictability* and *diversity* (by optimizing the mutual information between the current state and next state, conditioned on which skill is being executed).

They can then use these primitives for *model-based planning* for a downstream task. You can think of this as a regular RL problem, except that an action in their "action space" takes the form "execute skill X for T timesteps". They use *model-predictive control* (MPC), in which you sample a bunch of trajectories, and execute the first action of the trajectory that gets the highest reward. Since each of their high-level actions determines the policy for T timesteps, they can scale to much longer horizon tasks than MPC can usually be used for. They show that this approach is competitive with regular model-based RL.

Read more: Paper: Dynamics-Aware Unsupervised Discovery of Skills

Rohin's opinion: I think unsupervised learning is likely to be key in getting more powerful and general AI systems without requiring a truly staggering amount of expert data, and this is a great example of what that might look like. Note though that the

learned primitives are certainly not what you'd expect of a human: for example, the humanoid learns to vaguely shuffle in a direction, rather than walking. In addition, they did require specifying an "x-y prior" that required skills to be diverse based on x-y coordinates, which is why the skills learned navigation primitives, as opposed to e.g. distinct types of flailing.

FEEDBACK

I'm always happy to hear feedback; you can send it to me, **Rohin Shah**, by **replying to this email**.

PODCAST

An audio podcast version of the **Alignment Newsletter** is available. This podcast is an audio version of the newsletter, recorded by **Robert Miles.**

Results of $1,000 Oracle contest!

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Almost a year ago, I posted a contest to find the best questions to ask an Oracle AI. I've been very slow in grading it, apologies for that, but here are the final results and the sumptuous cash prizes.

First of all, thanks for everyone who wrote an entry; even if your suggestion didn't win, most of them had interesting ideas that made me think. But here are the prizes:

Counterfactual Oracle

- $350 for the best question(s) to ask a counterfactual Oracle, awarded to Wei_Dai for any one of these posts.

The best counterfactual Oracle suggestions revolved around using the Oracle to automate or hasten difficult human work, by getting the Oracle to tell us what humans would have achieved had they done the work. Wei_Dai's suggestions in this area were the most useful, and he also had an interesting iterated Oracle idea that suggests some other avenues to explore.

Low-Bandwidth Oracle

- $350 for the best question(s) to ask a low-bandwidth Oracle: the joint winners are Wei_Dai, for his idea on using the Oracle to predict crime, and William_S, for his idea to get the Oracle to critique plans.

Generally good ideas

There's a total of $300 "to be distributed as I see fit among the non-winning entries; I'll be mainly looking for innovative and interesting ideas that don't quite work." This award is split among several people:

- $150 to cousin_it for finding the "bucket chain" flaw if there are Oracles whose time-ranges of prediction overlap.
- $50 to evhub's idea to use Oracles to do iterated amplification and distillation.
- $50 to paulfchristiano for a useful critique of evhub's idea, the thread starting from here. It increased my understanding both of Paul's approach and my own.
- $25 to Gurkenglas for using Oracles to see how predictable history was; I have no idea how to make this work precisely, but it is a fascinating idea.
- $25 to romeostevensit for a very thorough schema of issues around Oracles.

Thanks again to all who participated, and the winners can contact me to receive their ill-gotten gains.

How alienated should you be?

Epistemic status: a post about values, where I'm confident the question is interesting but not confident my answer is correct. Factual statements are intended to be correct summaries, but I won't be careful about citations. This post was written in January, and edited and published today.

The "correct" level and type of alienation, like a "balanced" posture, is a dynamic function of the individual and environment. Nevertheless, we can say useful things about it, as we can about posture and balance.

We often talk on LessWrong about the importance of human intelligence. Often, we mean *individual* human intelligence, instead of *collective* human intelligence, and for many topics the difference is immaterial. But when we talk about alienation, or the separation between the individual and the collective, the distinction is paramount.

> For the strength of the pack is the wolf, and the strength of the wolf is the pack. -- Kipling, *The Law For Wolves*

The evolutionary history of humans is inexorably tied up with their existence in bands. The drive to imitate allows for cultural accumulation, and the drive to teach further accelerated that growth. Language gave structure to thoughts, and more importantly allowed easier transmission. Trade allowed for specialization and *fragmentation* of knowledge.

We see some evidence that improving collective ability was worth individual costs. Much has been made out of how early farmers had clear signs of malnutrition compared to hunter gatherers, like their shorter skeletons. Note that malnutrition is bad for individual intelligence, and farming benefits collective capability primarily through the increase in population carrying capacity. (It also allows for saving of wealth in a way that means society can be arranged differently, and more reliable trade because people are more stationary, but I think those effects are small for our purposes here.)

> To live with an immense and proud composure: always beyond. - To have and not have one's feelings, one's for and against, voluntarily, to condescend to them for hours, to sit on them, as if on a horse, often as if on a donkey: - for one needs to know how to use their stupidity as well as their fire. To preserve one's three hundred foregrounds, as well as one's dark glasses: for there are occasions when no one should be allowed to look into our eyes, even less into our "reasons." And to select for company that mischievous and cheerful vice, courtesy. And to remain master of one's four virtues: courage, insight, sympathy, and loneliness. For solitude is a virtue with us, as a sublime tendency and impulse for cleanliness, which senses how contact between one person and another - "in society" - must inevitably bring impurity with it. Every community somehow, somewhere, sometime makes people - "common." --Nietzsche, *Beyond Good and Evil* #284

In No Safe Defense, Not Even Science, Eliezer observes many rationalists had an early experience that broke their emotional trust in the sanity of the people around them, and this made it necessary for them to form independent judgment.

Selection on collectives is done on *collective*, not individual, survival and popularity. This means that the incentives on the collective for the accuracy of individual beliefs

and the satisfaction of individual preferences is weak at best. When collective benefit and individual benefit conflict, we should expect significant pressure to be pro-social; that is, put the collective first.

Many advances in rationality come from rejecting epistemically invalid pressures to be pro-social. From the collective's point of view, belief in the local religion is generally not a question about the supernatural, but instead a question of "are you one of us?", and the rare loud atheist who took the question literally instead of seriously was correctly seen as "not one of us." Silly rules are also a more effective test of desire to be a member than sensible rules; even sinners do sensible things.

But, from the individual's perspective, what an evil trick for society to pull! It, in its bigness, decides that your epistemology should be censored and constrained in an opaque way that is for its benefit. It might claim that it's for your benefit also, but in a way it will only let you evaluate using the crippled epistemology.

Many advances in social technology seem like they manage this balancing act more delicately. Capitalism is often characterized as trying to harness individual ambition for pro-social ends. Liberal humanist democracy increases the incentive for the collective to take some sorts of individual benefits more seriously, but more importantly gives individuals more of a moral license to be sovereign.

> >u ever sit on a train and listen to a good song and sunlight is pouring into the carriage as u pull into the city and u just,...feel an overwhelming awe and love for the human race? like we built this train!!!! we built this city!!!!! billions of hands and millions of ideas and thousands of years and now I'm here!!! sitting on this train!!! listening to music that was written between all the infrastructure and progress just because!!! human beings are clever and loving and creative and that passion moves in all directions and has inevitably lead me here!!! to this train!! going to uni!!! the most mundane thing in the world and it's so utterly remarkable it makes me feel tiny and also enormous --bactii

Scott Alexander writes that nerds can be bees, too. Often the capacity and desire to belong are there, and just the group to belong to is missing.

So the question at the end is, what should one do with one's time, and sense of belonging, and sense of alienation? Bryan Caplan recommends building a beautiful bubble. This seems obviously correct in many ways, but I think it fails to grapple with the ways in which one's values are flexible and socially constructed. It is a different sort of work to 'do what you think is right' and 'figure out what you should think is right.'

I've been thinking about this general topic for years, not with this specific name, but this post solidified while I was thinking about effective altruism, youth movements, recruiting for work on existential risk reduction, and Kolmogorov Complicity.

For x-risk reduction, it often is the case that people think someone else has it handled. And in a world that's often adequate, that's not a crazy thing to think!

For effective altruism, it seems like the general society's pro-social pressure is to collude to pretend to not notice the various ways in which the world is on fire, or the child at the center of Omelas. On the one hand, this is one of the main reasons humanity has nice things at all.

> Typical of this attitude is the comment that, 'If we can send a man to the moon, why can't we-" followed by whatever project the speaker favors. The fact that we sent a man to the moon is part of the reason why many other things could not be done. --Thomas Sowell, *Knowledge and Decisions*

On the other hand, this is how mortality becomes deliberately ignored, despite the obvious individual interest in continuing to live. [And a society that ignores individual mortality through tweaks to epistemics instead of values seems like the sort of society that might end up ignoring collective mortality too!]

But it's one thing to convince people that society is predatory, or misleading them, or generally worthy of being alien to, and another thing to create a collective that is still able to do good in the world, and worth belonging to.

For example, when I look at my psychology and values, I see both a deep individualism and a deep cosmopolitanism. That is, I want to be able to do my own thing, make my own choices, and generally be weird, and also I respect other people's ability to do their own thing, and make their own choices, and generally be weird. You can have one without the other! One could be a would-be tyrant, accepting no limit on their own behavior whilst cruelly limiting others, or the formless follower, believing the Other is correct regardless of how it's shaped and seeking to become whatever will fit in.

Given that those aspects needed to be paired to be good, the challenge of creating a culture, and of convincing others to join it, is more difficult than it might seem.

Coronavirus as a test-run for X-risks

On the 27th February, I, like many of us, became fully aware of the danger humanity was facing (let's thank 'Seeing the Smoke') and put my cards on the table with this:

> This is partly a test run of how we'd all feel and react during a genuine existential risk. Metaculus currently has it as a 19% chance of spreading to billions of people, a disaster that would certainly result in many millions of deaths, probably tens of millions. Not even a catastrophic risk, of course, but this is what it feels like to be facing down a 1/5 chance of a major global disaster in the next year. It is an opportunity to understand on a gut level that, this is possible, yes, real things exist which can do this to the world. And it does happen.
>
> It's worth thinking that specific thought now because this particular epistemic situation, a 1/5 chance of a major catastrophe in the next year, will probably arise again over the coming decades. I can easily imagine staring down a similar probability of dangerously fast AGI takeoff, or a nuclear war, a few months in advance.

Well, now a few months have gone by and much has changed. The natural question to ask is- what general lessons have we learned, compared to that 'particular epistemic situation', now that we're in a substantially different one? What does humanity's response to the coronavirus pandemic so far imply about how we might fare against genuine X-risks?

At a first pass, the answer to that question seems obvious - not very well. The response of most usually well-functioning governments (I'm thinking mainly of Western Europe here) has been slow, held back by an unwillingness to commit all resources to a strategy and accept its trade-offs, and sluggish to respond to changing evidence. Advance preparation was even worse. This post gives a good summary of some of those more obvious lessons for X-risks, focussing specifically on slow AI takeoff.

As to what we ultimately blame for this slowness - Scott Alexander and Toby Ord gave as good an account as anyone (before the pandemic) in blaming a failure to understand expected value and the availability heuristic.

However, many of us predicted in advance the dynamics that would lead countries to put forward a slow and incoherent response to the coronavirus. What I want to explore now is - what has changed epistemically *since I wrote that comment* - what things have happened since that have surprised many of us who have internalised the truth of civilisational inadequacy? I am looking for generalised lessons we can take from this pandemic, rather than specific things we have learnt *about* the pandemic in the last few months. I believe there is one such lesson that is surprising, which I'd like to convince you of.

Underweighting Strong Reactions

My claim is that in late February/early March, many of us did overlook or underweight the possibility that many countries would eventually react strongly to coronavirus - with measures like lockdowns that successfully drove R under 1 for extended periods,

or with individual action that holds R near 1 in the absence of any real government intervention. This meant we placed too much weight on coronavirus going uncontained, and were surprised when in many countries it did not.

Whether the strong reaction is fully or only partially effective remains to be seen, but the fact that this reaction occurred was surprising to many of us, relative to what we believed at the start of all this - I know that it surprised me.

I will first present the examples of predictions, some from people on LessWrong or adjacent groups and some from government scientists, which all either foretold worse outcomes by now, more feeble results from interventions, lower compliance with interventions, that interventions wouldn't even be implemented or predicted bad outcomes that are not yet ruled out but now look much less likely than they did.

I will then put forward an explanation for these mistakes - something I named (in April) the 'Morituri Nolumus Mori' *('We who are about to die don't want to')* effect, in reference to the Discworld novel *The Last Hero*: that most governments and individuals have a consistent, short-term aversion to danger which is stronger than many of us suspected, though not sustainable in the absence of an imminent threat. I'll first go through the incorrect predictions and then give my favoured explanation. If I am correct that many of us (and also many scientists and policymakers) missed the importance of the MNM effect, it should increase our confidence that, in situations where there is some warning, there are fairly basic features of our psychology and institutions that do get in the way of the very worst outcomes. However, the MNM effect is limited and will not help in any situation where advance planning or responding to things above and beyond immediate incentives are required.

I consider the MNM effect to be mostly compatible with Zvi's 'Governments Most Places Are Lying Liars With No Ability To Plan or Physically Reason.' (*I do think that claim is too America-centric, and* '***no*** *ability to plan/reason*' *is hyperbole if applied to Europe or even the UK, let alone e.g. Taiwan*). The MNM effect is what we credit *instead* of clever planning or reasoning, for why things aren't as bad as they could be - the differences between e.g. America and Germany are due to any level of planning *at all*, not better planning.

Noticing Confusion

Things (especially in the US) are sufficiently bad right now that it is difficult to remember that many of us put significant weight on things already being worse than they currently are - but as I will show that was the case.

Some people's initial predictions were that R would not be driven substantially below 1 for any extended period, anywhere, except with a Wuhan style lockdown. Robin Hanson seemingly claimed this on March 19: 'So even if you see China policy as a success, you shouldn't have high hopes if your government merely copies a few surface features of China policy.' In that article Hanson was clearly referring to 'most governments' that aren't China as being unlikely to suppress without adopting a deep mimicry of China's policy - including welding people into their flats and forcible case isolation. Yet, two months later there are many countries, from New Zealand to Germany, which have simply copied some but not all features of the Chinese policy while achieving initial suppression.

More recently, Hanson updated to speaking more specifically about the USA: (in response to a graphic showing several examples of successful suppression in Europe and Asia) 'Yes, you know that other nations have at times won wars. Even so, you must decide if to choose peace or war.' Going from 'most western countries' to 'America' counts as an optimistic update.

But mitigation measures (which Hanson calls 'peace') have also worked out less disastrously than our worst fears suggested because of stronger-than-expected individual action. See e.g. this article about Sweden:

> Ultimately, Sweden shows that some of the worst fears about uncontrolled spread may have been overblown, because people will act themselves to stop it. But, equally, it shows that criticisms of lockdowns tend to ignore that the real counterfactual would not be business as usual, nor a rapid attainment of herd immunity, but a slow, brutal, and uncontrolled spread of the disease throughout the population, killing many people. Judging from serological data and deaths so far, it is the speed of deaths that people who warned in favour of lockdowns got wrong, not the scale.

This remark about Sweden is applicable more generally - the worst case scenario for almost every country seems to be R around 1.5 at this point - see this map from Epidemic Forecasting. True explosive spread is very rare across the world, but was being discussed as a real possibility in early March even in Europe. Again, the response is not good enough to outright reverse the unfolding disaster, but it is still strong enough to arrest explosive spread.

Focussing on the UK, which had a badly delayed response and a highly imperfect lockdown, we can see that even there R was driven substantially below 1 and hospital admissions with Covid-19 (which are the most reliable short-term proxy for infection throughout the overall pandemic) are at 13% of their peak. London did not exceed its ICU capacity despite predictions that it would from government modellers.

Another way of getting at this disjoint is to just look at the numbers and see if we still expect the same number of people to die. Wei Dai initially (1st March) predicted 190-760 million people would eventually die from coronavirus with 50% of the world infected. The more recent top-rated comment by Orthonormal points out that current evidence points against that. Good Judgment rates the probability that more than 80 million will die as 1%. A recent paper by Imperial College suggested that the Europe-wide lockdowns have so far saved 3 million lives without accounting for the fact that deaths in an unmitigated scenario would have been higher due to a lack of intensive care beds. Regardless of what happens next, would we have predicted that in early March?

These mistakes have not been limited to the LessWrong community - one of the reasons for the aforementioned delay before the UK called the lockdown was that UK behavioural scientists advising the government were near certain that stringent lockdown measures would not be obeyed to the necessary degree and lockdowns in the rest of Europe were instead implemented 'more for solidarity reasons'. In the end it turned out that compliance was instead 'higher than expected'. The attitude in most of Europe in early March was that full lockdowns were completely infeasible. Then they were implemented.

Another way of getting at this observation is to note the people who have publicly recorded their surprise or shift in belief as these events have unfolded. I have written

several comments with earlier versions of this claim, starting two months ago. Wei Dai notably updated in the direction of thinking coronavirus would reach a smaller fraction of the population, after reading this prescient blogpost:

> The interventions of enforced social distancing and contract tracing are expensive and inevitably entail a curtailment of personal freedom. However, they are achievable by any sufficiently motivated population. An increase in transmission *will* eventually lead to containment measures being ramped up, because every modern population will take draconian measures rather than allowing a health care meltdown. In this sense COVID-19 infections are not and will probably never be a full-fledged pandemic, with unrestricted infection throughout the world. It is unlikely to be allowed to ever get to high numbers again in China for example. It will always instead be a series of local epidemics.

In a recent podcast, Rob Wiblin and Tara Kirk Sell were discussing what they had recently changed their minds about. They picked out the same thing:

> **Robert Wiblin:** Has the response affected your views on what policies are necessary or should be prioritized for next time?
>
> **Tara Kirk Sell:** The fact that “Stay-at-home orders” are actually possible in the US and seem to work… I had not really had a lot of faith in that before and I feel like I’ve been surprised. But I don’t want “Stay-at-home orders” to be the way we deal with pandemics in the future. Like great, it worked, but I don’t want to do this again.

Or this from Zvi:

> 5. Fewer than 3 million US coronavirus deaths: 90%
>
> I held. Again, we saw very good news early, so to get to 3 million now we’d need full system collapse to happen quickly. It’s definitely still possible, but I’m guessing we’re now more like 95% to avoid this than 90%.

Lastly, we have the news from the current hardest-hit places, like Manhattan, which have already hit partial herd immunity and show every sign of being able to contain coronavirus going forward even with imperfect measures.

The Morituri Nolumus Mori effect

Many of these facts (in particular the reason that 100 million plus dead is effectively ruled out) have multiple explanations. For one, the earliest data on coronavirus implied the hospitalization rate was 10-20% for all age groups, and we now know it is substantially lower (that tweet by an author of the Imperial College paper, which estimated a hospitalization rate of 4.4%). This means that if hospitals were entirely unable to cope with the number of patients, the IFR would be in the range of 2%, not 20% initially implied.

However, the rest of our information about the characteristics of the virus in early March- the estimate of R0 and ‘standard’ IFR, were fairly close to the mark. Our predictions were working off of reasonable data about the virus. Any prediction made then about the number of people who would be infected isn’t affected by this hospitalization rates confounder, nor is any prediction about what measures would be

implemented. So there must be some other reason for these mistakes - and a common thread among nearly all the inaccurate pessimistic predictions was that they underestimated the forcefulness, though not the level of forethought or planning, behind mitigation or suppression measures. As it is written,

> "Brains don't work that way. They don't suddenly supercharge when the stakes go up - or wh*en they do, it's within hard limits.* I couldn't calculate the thousandth digit of pi if someone's life depended on it."

The Morituri Nolumus Mori effect, as a reminder, is the thesis that governments and individuals have a consistent, short-term reaction to danger which is stronger than many of us suspected, though not sustainable in the absence of an imminent threat. This effect is just such a hard limit - it can't do very much except work as a stronger than expected brake. And something like it has been proposed as an explanation, not just by me two months ago but by Will MacAskill and Toby Ord, for why we have already avoided the worst disasters. Here's Toby's recent interview:

> Learning the right lessons will involve not just identifying and patching our vulnerabilities, but pointing towards strengths we didn't know we had. The unprecedented measures governments have taken in response to the pandemic, and the public support for doing so, should make us more confident that when the stakes are high we can take decisive action to protect ourselves and our most vulnerable. And when faced with truly global problems, we are able to come together as individuals and nations, in ways we might not have thought possible. This isn't about being self-congratulatory, or ignoring our mistakes, but in seeing the glimmers of hope in this hardship.

Will MacAskill made reference to the MNM effect in a pre-coronavirus interview, explaining why he puts the probability of X-risks relatively low.

> Second then, is just thinking in terms of the rational choice of the main actors. So what's the willingness to pay from the perspective of the United States to reduce a single percentage point of human extinction whereby that just means the United States has three hundred million people. How much do they want to not die? So assume the United States don't care about the future. They don't care about people in other countries at all. Well, it's still many trillions of dollars is the willingness to pay just to reduce one percentage point of existential risk. And so you've got to think that something's gone wildly wrong, where people are making such incredibly irrational decisions.

Bill Gates also referred to this effect.

I also think that the MNM effect is the main reason why both Metaculus and superforecasters consistently predicted deaths will stay below 10 million, implying a very slow burn, neither suppression nor full herd immunity, right across most of the world.

The Control System

From Slatestarcodex:

> Is there a possibility where R0 is exactly 1? Seems unlikely – one is a pretty specific number. On the other hand, it's been weirdly close to one in the US, and

> worldwide, for the past month or two. You could imagine an unfortunate control system, where every time the case count goes down, people stop worrying and go out and have fun, and every time the case count goes up, people freak out and stay indoors, and overall the new case count always hovers at the same rate. I've never heard of this happening, but this is a novel situation.

One more speculative consequence of the MNM effect is that a reactive, strong push against uncontrolled pandemic spread is a good explanation for why Rt tends to approach 1 in countries without a coordinated government response, like the United States, and the more coordinated the response the further below 1 Rt can be pushed. A priori, we might expect that there is some 'minimal default level' of response that leads to Rt being decreased from R0, 3-4, to some much lower value - but why is the barometer set around 1? It's not a coincidence, as Zvi points out.

> Whenever something lands almost exactly on the only inflection point, in this case R0 of one where the rate of cases neither increases nor decreases, the right reaction is suspicion.
>
> In this case, the explanation is that a control system is in play. People are paying tons of attention to when things are 'getting better' or 'getting worse' and adjusting behaviour, both legally required actions and voluntary actions.

The MNM effect is apparently so predictable that, with short-ish term feedback, it can form a control system. The other end of this control system is all the usual cognitive and institutional biases that prevent us from taking these events seriously and actually planning for them.

It is possible this is the first time such a control system has formed to mitigate a widespread disaster. Disasters of this size are rare throughout history. Add to this the fact that such control systems can only form when the threat unfolds and changes over several months, giving people time to veer between incaution and caution. Meanwhile, the short term feedback which governments and people can access about the progress of the epidemic is relatively new - better data collection and mass media make modern populations much more sensitive to the current level of threat than those throughout history. Remembering that noone knows exactly where or when the Spanish Flu began highlights that good real-time monitoring of a pandemic is an extremely new thing.

In our current situation of equilibrium created by a control system, the remaining uncertainties are: *can we do better than the equilibrium position?* (sociological and political) and *how bad is the equilibrium position?* (mainly a matter of the disease dynamics). It seems to me, the equilibrium probably ends in partial herd immunity (nowhere near 75% 'full herd immunity', because of MNM). This involves healthcare systems struggling to cope to some extent along the way. The US is essentially bound for equilibrium - but what that entails is not clear. I could imagine the equilibrium holding Rt near 1 even in the absence of any government foresight or planning but it doesn't seem very likely, as some commenters pointed out. More likely it ends with partial herd immunity.

However, there is still a push away from this equilibrium in Europe (e.g. attempts to use national-level tracing and testing programs). This push is not that strong and depends on individuals sticking to social distancing rules. European lockdowns brought Rt down to between 0.6 and 0.8, noticeably below 1, indicating that they beat

the equilibrium to some degree for a while. Rt got down to 0.4 in Wuhan, suggesting great success in beating the equilibrium.

That is the other lesson - any level of government foresight or planning adds on to the already existing MNM effect - witness how foot traffic levels dramatically declined before lockdowns were instituted, or even if they were never instituted, right across the world. The effects are additive. So if the default holds Rt near 1, then a few extra actions by a government able to look some degree into the future can make all the difference.

Conclusions

I consider that the number of predictions that have already been falsified or rendered unlikely is sufficient to establish that the MNM effect exists, or is stronger than many of us thought early on (I don't imagine there were many people who would have denied the MNM effect exists at all, i.e. expected us to just walk willingly to our deaths). 'Dumb reopening' as is happening the US, as a successor to lockdowns that have pushed R to almost exactly 1, is consistent with what I have claimed - that our reliable and predictable short-term reactivity (governmental and individual) and desire to not die, the Morituri Nolumus Mori effect, serves as a brake against the very worst outcomes. What next?

Conceivably, the control system could keep running, and R could stay near 1 perpetually even with no effective planning or well-enforced lockdowns, or there could be a slow grind as the virus spreads up to a partial herd immunity threshold - either way, the MNM effect is there, screening off some outcomes that looked likely in early March, such as a single sharp peak. Similarly, the MNM effect gives a helping hand to attempts at real strategy. Some governments that are competent in the face of massive threats but slow to react (such as Germany) did better than expected because of the caution of citizens who started restricting their movements before lockdown and who now aren't taking full advantage of reopened public spaces.

From the perspective of predicting future X-risks, the overall outcome of this pandemic is less interesting than the fact that there has been a consistent, unanticipated push from reactive actions against the spread of the virus. Then there is a further, also relevant issue of whether countries can beat the equilibrium (of R being held at near 1 or just above 1) and do better than the MNM effect mandates. So far, Europe spent a while beating equilibrium (with R during lockdown at 0.6-0.8) and China drove R down even further.

The first remaining uncertainty is: *can a specific country/the world as a whole do better than this equilibrium position?* We do have some pertinent evidence to answer this in the form of the superforecaster predictions and, though it is confounded by the next uncertainty, from disease modelling. The insights of disease modelling should shed light on the question: *how bad is this equilibrium position?* If we knew this we would have a better sense of what the reasonable worst case scenario is for coronavirus, but that is not important from an x-risk perspective.

This makes it clear what kinds of evidence are worth looking out for. We should look at the performance of areas of the world where there is little advance planning, but nevertheless the people are informed about the level of day-to-day danger and leaders don't actively oppose individual efforts at safety. Parts of the United States fit the bill. Seeing the eventual outcomes in these areas, when compared to some initial

predictions about just how bad things *could* get, will give us an idea of the extra help provided by the MNM effect. Then, with that as our baseline, we can see how many countries do better to judge the further help provided by planning or an actual strategy.

Implications for X-risks

The most basic lesson that should be learned from this disaster is, of course, that for the moment we are inadequate - unable to coordinate as long as there is any uncertainty about what to do, and unable to meaningfully plan in advance for plausible near-term threats like those from pandemics. We should of course remember that not enough focus is put on long-term risks, that our institutions are flawed in dealing with them.

Covid-19 shows that there can still be a strong reaction once it is clear there is disaster coming. We have some idea already just how strong this reaction is. We have less idea how effective it will end up being. In February and March, we often observed a kind of pluralistic ignorance, where even experts raising the alarm did so in a way that was muted and seemingly aimed at 'not causing panic'.

> **Robert Wiblin:** I think part of what was going on was perhaps people wanted to promote this idea of "Don't panic" because they were worried that the public would panic and they felt that the way to do that was really to talk down the risk a lot and then it kind of got a bit out of control, but I'm not sure how big the risk of... It seems like what's ended up happening is much worse than the public panicking in January. Or maybe I just haven't seen what happens when the public really panics. I guess people panicked later and it wasn't that bad.

Suppose this dynamic applies in a future disaster. We might expect to see a sudden phase change from indifference to panic despite the fact that trouble was already looming anyway and no new information has appeared.

If there is enough forewarning before the disaster occurs that a phase shift in attitudes can take place, we will react hard. Suppose the R0 of Coronavirus had been 1.5-2, and the rest of our response had been otherwise the same - suppression measures taken in the US and elsewhere would have worked perfectly even though we were sleepwalking towards disaster as recently as three weeks before. The only reason this didn't happen is because of contingent facts about this particular virus. On the other hand, there are magnitudes of disaster which the MNM effect is clearly inadequate for - suppose the R0 had been 8.

Perhaps the MNM effect is stronger for a disaster, like a pandemic, for which there is some degree of historical memory and evolved emotions and intuitions around things like purity and disgust which can take over and influence our risk-mitigation behaviour. Maybe technological disasters that don't have the same deep evolutionary routes, like nuclear war, or X-risks like unaligned AGI that have literally never happened before, would not evoke the same strong, consistent reaction because the threat is even less comprehensible.

Nevertheless, one could imagine a slow AI takeoff scenario with a lot of the same characteristics as coronavirus, where the MNM effect steps in at the last moment:

It takes place over a couple of years. Every day there are slight increases in some relevant warning sign. A group of safety people raise the alarm but are mostly ignored. There are smaller scale disasters in the run-up, but people don't learn their lesson (analogous to SARS-1 and MERS). Major news orgs and government announce there is nothing to worry about (analogous to initial statements about masks and travel bans). Then there is a sudden change in attitudes for no obvious reason. At some point everyone freaks out - bans and restrictions on AI development, right before the crisis hits. Or, possibly, right when it is already too late.

The lesson to be learned is that there may be a phase shift in the level of danger posed by certain X-risks - if the amount of advance warning or the speed of the unfolding disaster is above some minimal threshold, even if that threshold would seem like far too little time to do anything given our previous inadequacy, then there is still a chance for the MNM effect to take over and avert the worst outcome. In other words, AI takeoff with a small amount of forewarning might go a lot better than a scenario where there is no forewarning, even if past performance suggests we would do nothing useful with that forewarning.

More speculatively, I think we can see the MNM effect's influence in other settings where we have consistently avoided the very worst outcomes despite systematic inadequacy - Anders Sandberg referenced something like it when he was discussing the probability of nuclear war. There have been *many* near misses when nuclear war could have started, implying that we can't have been lucky over and over. Instead that there has been a stronger skew towards interventions that halt disaster at the last moment, compared to not-the-last-moment:

> **Robert Wiblin:** So just to be clear, you're saying there's a lot of near misses, but that hasn't updated you very much in favor of thinking that the risk is very high. That's the reverse of what I expected.
>
> **Anders Sandberg:** Yeah.
>
> **Robert Wiblin:** Explain the reasoning there.
>
> **Anders Sandberg:** So imagine a world that has a lot of nuclear warheads. So if there is a nuclear war, it's guaranteed to wipe out humanity, and then you compare that to a world where is a few warheads. So if there's a nuclear war, the risk is relatively small. Now in the first dangerous world, you would have a very strong deflection. Even getting close to the state of nuclear war would be strongly disfavored because most histories close to nuclear war end up with no observers left at all.
>
> In the second one, you get the much weaker effect, and now over time you can plot when the near misses happen and the number of nuclear warheads, and you actually see that they don't behave as strongly as you would think. If there was a very strong anthropic effect you would expect very few near misses during the height of the Cold War, and in fact you see roughly the opposite. So this is weirdly reassuring. In some sense the Petrov incident implies that we are slightly safer about nuclear war.

On the other hand, the MNM effect requires leaders and individuals to have access to information about the state of the world *right now* (i.e. how dangerous are things at the moment). Even in countries with reasonably free flow of information this is not a given. If you accept Eliezer Yudkowksy's thesis that clickbait has impaired our ability to understand a persistent, objective external world then you might be more

pessimistic about the MNM effect going forward. Perhaps for this reason, we should expect countries with higher social trust, and therefore more ability for individuals to agree on a consensus reality and understand the level of danger posed, to perform better. Japan and the countries in Northern Europe like Denmark and Sweden come to mind, and all of them have performed better than the mitigation measures employed by their governments would suggest.

The principle that I've called the Morituri Nolumus Mori effect is defined in terms of the map, not the territory - a place where our predictions diverged from reality in an easily and consistently describable way - that the short-term reaction from many governments and individuals was *stronger than we expected*, whilst advance planning and reasoning was as weak as we expected. The MNM effect may also be a feature of the territory. It may already have a name in the field of social psychology, or several names. It may be a contingent artefact of lots of local facts about only our coronavirus response, though I don't think that's plausible for the reasons given above. Either way, I believe that it was an important missing piece, probably the biggest missing piece, in our early predictions and needs to be considered further if we want to refine our analysis of X-risks going forward. One of the few upsides to this catastrophe is that it has provided us with a small-scale test run of some dynamics that might play out during a genuine catastrophic or existential risk, and we should be sure to exploit that for all its worth.

Quick Look #1 Diophantus of Alexandria

https://www.storyofmathematics.com/hellenistic_diophantus.html

Diophantus of Alexandria, a 2nd Century Greek mathematician, had a lot of the concepts needed to develop an Algebra. However, he was unable to fully generalize his methods of problem solving, even if he invented some interesting methods.

Ancient math was written in paragraphs, using words for the most part, thus making reading it very, very painful compared to the compact elegance of modern mathematical notation. However, I was surprised to see Diophantus (or his very early editors at least) develop some interesting and helpful notation in his algebra.

Final sigma 'ς' represented the unknown variable, but there were different symbols for variables of every power so for x^2… x^6 each had a unique variable. In fact, this situation persisted into the 17th century, even Fermat used N for unknown and S for the unknown-squared and C for the unknown cubed!

The problem with this is that it meant Diophantus couldn't devise general methods to solve algebraic problems which had multiple unknowns, and it wasn't obvious to him that one CAN frequently relate x^2 to x.

The cool thing about this notation from the past though, is how it makes obvious something that Algebra I – Algebra II students mess up frequently. You can't just combine x^2 + x^3, these are different variables whose relation concerns the base. And almost everyone has made this mistake in their early math career. Some never recover.

Although the editor of my copy, Sir Thomas L. Heath, claims that Diophantus experienced limited options as a mathematician because all the letters of the Greek alphabet were in use as letters except for the final sigma, which Diophantus used to represent the unknown variable, I think D *could* have invented more variables quite easily. We see this in his invention of the subtraction sign as an inverted psi, and his use of a different variable with superscript for an unknown to the nth power up to the sixth. There was also the extinct digamma and all the Egyptian symbols which at least could have cribbed off of. Surely, the problem was not a lack of imagination, but merely satisfaction with the method then in use. Besides, one person can only invent so much, unless that person is Leonard Euler or Von Neumann, neither of whom had any limits. D. merely didn't see the limits of his notation.

Although D's problems are surprisingly challenging even using modern notation, the logic D. used to solve the problems is obscure. He does not explain his step by step process. Since they are not proofs, and they are merely problems, it's hard to divine exactly what D. thought the import of his methods were or exactly which steps he took to come to the answer. He seems to have used trial and error to solve some problems frequently, just plugging in numbers until the right answer popped out. He only wanted positive integers in his answers, so the problems are designed to reflect that. However, some problems don't have an answer as a whole number. For those he would estimate the answer. "X is < 11 and > 10." Sometimes he is wrong on these estimations! I don't know quite what to make of that. In problems whose answer is a negative number, Diophantus says, "Pthht, absurd!"

This is unfortunate, because if D. had credits and debts in mind when he was putting together these problems, he might have seen the utility of negative numbers and started an accounting revolution 1500 years early.

If Diophantus can teach us one thing about discovery, I believe it is that iterating over different methods of notation might lead us to make conceptual breakthroughs.

Where to Start Research?

When I began what I called the knowledge bootstrapping project, my ultimate goal was “Learn how to learn a subject from scratch, without deference to credentialed authorities”. That was too large and unpredictable for a single grant, so when I applied to LTFF, my stated goal was “learn how to study a single book”, on the theory that books are the natural subcomponents of learning (discounting papers because they’re too small). This turned out to have a flawed assumption baked into it.

As will be described in a forthcoming post, the method I eventually landed upon involves starting with a question, not a book. If I start with a book and investigate the questions it brings up (you know, like I’ve been doing for the last 3-6 years), the book is controlling which questions get brought up. That’s a lot of power to give to something I have explicitly decided not to trust yet.

Examples:

- When reading The Unbound Prometheus, I took the book’s word that a lower European birth rate would prove Europeans were more rational than Asians and focused on determining whether Europe’s birth rates were in fact lower (answer: it’s complicated), when on reflection it’s not at all clear to me that lower birth rates are evidence of rationality.
- “Do humans have exactly 4 hours of work per day in them?” is not actually a very useful question. What I really wanted to know is “when can I stop beating myself up for not working?“, and the answer to the former doesn’t really help me with the latter. Even if humans on average have 4 hours, that doesn’t mean I do, and of course it varies by circumstances and type of work… and even “when can I stop beating myself up?” has some pretty problematic assumptions built into it, such as “beating myself up will produce more work, which is good.” The real question is something like “how can I approach my day to get the most out of it?”, and the research I did on verifying a paper on average daily work capacity didn’t inform the real question one way or the other.

What would have been better is if I’d started with the actual question I wanted to answer, and then looked for books that had information bearing on that question (including indirectly, including very indirectly). This is what I’ve started doing.

This can look very different depending on what type of research I’m doing. When I started doing covid research, I generated a long list of fairly shallow questions. Most of these questions were designed to inform specific choices, like “when should I wear what kind of mask?” and “how paranoid should I be about people without current symptoms?”, but some of them were broader and designed to inform multiple more specific questions, such as “what is the basic science of coronavirus?”. These broader, more basic questions helped me judge the information I used to inform the more specific, actionable questions (e.g., I saw a claim that covid lasted forever in your body the same way HIV does, which I could immediately dismiss because I knew HIV inserted itself your DNA and coronaviruses never enter the nucleus).

I used to read a lot of nonfiction for leisure. Then I started doing epistemic spot checks– taking selected claims from a book and investigating them for truth value, to assess the book’s overall credibility- and stopped being able to read nonfiction without

doing that, unless it was one of a very short list of authors who'd made it onto my trust list. I couldn't take the risk that I was reading something false and would absorb it as if it were true (or true but unrepresentative, and absorb it as representative). My time spent reading nonfiction went way down.

About 9 months ago I started taking really rigorous notes when I read nonfiction. The gap in quality of learning between rigorous notes and my previous mediocre notes was about the same as the gap between doing an epistemic spot check and not. My time spent reading nonfiction went way up (in part because I was studying the process of doing so), but my volume of words read dropped precipitously.

And then three months ago I shifted from my unit of inquiry being "a book", to being "a question". I'm sure you can guess where this is going- I read fewer words, but gained more understanding per word, and especially more core (as opposed to shell or test) understanding.

The first two shifts happened naturally, and while I missed reading nonfiction for fun and with less effort, I didn't feel any pull towards the old way after I discovered the new way. Giving up book-centered reading has been hard. Especially after five weeks of frantic covid research, all I wanted to do was to be sat down and told what questions were important, and perhaps be walked through some plausible answers. I labeled this a desire to learn, but when I compared it to question-centered research, it became clear that's not what it was. Or maybe it was a desire to go through the act of learning something, but it was not a desire to answer a question I had and was not prioritized by the importance of a question. It was best classified as leisure in the form of learning, not resolving a curiosity I had. And if I wanted leisure, better to consume something easier and less likely to lead me astray, so I started reading more fiction, and the rare non-fiction of a type that did not risk polluting my pool of data. And honestly I'm not sure that's so safe: humans are built to extract lessons from fiction too.

Put another way: I goal factored (figured out what I actually wanted from) reading a nonfiction book, and the goal was almost never best served by using a nonfiction book as a starting point. Investigating a question I cared about was almost always better for learning (even if it did eventually cash out in reading a book), and fiction was almost always better for leisure, in part because it was less tiring, and thus left more energy for question-centered learning when that was what I wanted.

Preparing for "The Talk" with AI projects

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Epistemic status: Written for Blog Post Day III. I don't get to talk to people "in the know" much, so maybe this post is obsolete in some way.

I think that at some point at least one AI project will face an important choice between deploying and/or enlarging a powerful AI system, or holding back and doing more AI safety research.

(Currently, AI projects face choices like this all the time, except they aren't important in the sense I mean it, because the AI isn't potentially capable of escaping and taking over large parts of the world, or doing something similarly bad.)

Moreover, I think that when this choice is made, most people in the relevant conversation will be insufficiently concerned/knowledgeable about AI risk. Perhaps they will think: "This new AI design is different from the classic models, so the classic worries don't arise." Or: "Fear not, I did [insert amateur safety strategy]."

I think it would be very valuable for these conversations to end with "OK, we'll throttle back our deployment strategy for a bit so we can study the risks more carefully," rather than with "Nah, we're probably fine, let's push ahead." This buys us time. Say it buys us a month. A month of extra time right after scary-powerful AI is created is worth a lot, because we'll have more serious smart people paying attention, and we'll have more evidence about what AI is like. I'd guess that a month of extra time in a situation like this would increase the total amount of quality-weighted AI safety and AI policy work by 10%. That's huge.

One way to prepare for these conversations is to raise awareness about AI risk and technical AI safety problems, so that it's more likely that more people in these conversations are more informed about the risks. I think this is great.

However, there's another way to prepare, which I think is tractable and currently neglected:

1. Identify some people who might be part of these conversations, and who already are sufficiently concerned/knowledgeable about AI risk.

2. Help them prepare for these conversations by giving them resources, training, and practice, as needed:

2a. Resources:

Perhaps it would be good to have an Official List of all the AI safety strategies, so that whatever rationale people give for why this AI is safe can be compared to the list. (See this prototype list.)

Perhaps it would be good to have an Official List of all the AI safety problems, so that whatever rationale people give for why this AI is safe can be compared to the list, e.g. "OK, so how does it solve outer alignment? What about mesa-optimizers? What about the malignity of the universal prior? I see here that your design involves X; according to the Official List, that puts it at risk of developing problems Y and Z..." (See this prototype list.)

Perhaps it would be good to have various important concepts and arguments re-written with an audience of skeptical and impatient AI researchers in mind, rather than the current audience of friends and LessWrong readers.

2b. Training & practice:

Maybe the person is shy, or bad at public speaking, or bad at keeping cool and avoiding fluster in high-stakes discussions. If so, some coaching and practice could go a long way. Maybe they have the opposite problems, frequently coming across as overconfident, arrogant, aggressive, or paranoid. If so someone should tell them this and help them tone it down.

In general it might be good to do some role-play exercises or something, to prepare for these conversations. As an academic, I've seen plenty of mock-dissertation-defense sessions and mock-job-talk-question-sessions, which seem to help. And maybe there are ways to get even more realistic practice, e.g. by trying to convince your skeptical friends that their favorite AI design might kill them if it worked.

Note that most of part 2 can be done without having done part 1. This is important in case we don't know anyone who might be part of one of these conversations, which is true for many and perhaps most of us.

Why do I think this is tractable? Well, seems like the sort of thing that people producing AI safety research can do on the margin, just by thinking more about their audience and maybe recording their work (or other people's work) on some Official List. Moreover people who don't do (or even read) AI safety research can contribute to this, e.g. by reading the literature on how to practice for situations like this, and writing up the results.

Why do I think this is neglected? Well, maybe it isn't. In fact I'd bet that some people are already thinking along these lines. It's a pretty obvious idea. But just in case it is neglected, I figured I'd write this. Moreover, the Official Lists I mentioned don't exist, and I think they would if people were taking this idea seriously. Finally--and this more than anything else is what caused me to write this post--I've heard one or two people explicitly call this out as something that they *don't* think is an important use case for the alignment research they were doing. I disagreed with them, and here we are. If this is a bad idea, I'd love to know why.

Relevant pre-AGI possibilities

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://aiimpacts.org/relevant-pre-agi-possibilities/

Epistemic status: I started this as an AI Impacts research project, but given that it's fundamentally a fun speculative brainstorm, it worked better as a blog post.

The default, when reasoning about advanced artificial general intelligence (AGI), is to imagine it appearing in a world that is basically like the present. Yet almost everyone agrees the world will likely be importantly different by the time advanced AGI arrives.

One way to address this problem is to reason in abstract, general ways that are hopefully robust to whatever unforeseen developments lie ahead. Another is to brainstorm particular changes that might happen, and check our reasoning against the resulting list.

This is an attempt to begin the second approach.[2] I sought things that might happen that seemed both (a) within the realm of plausibility, and (b) probably strategically relevant to AI safety or AI policy.

I collected potential list entries via brainstorming, asking others for ideas, googling, and reading lists that seemed relevant (e.g. Wikipedia's list of emerging technologies,[3] a list of Ray Kurzweil's predictions[4], and DARPA's list of projects.[5])

I then shortened the list based on my guesses about the plausibility and relevance of these possibilities. I did not put much time into evaluating any particular possibility, so my guesses should not be treated as anything more. I erred on the side of inclusion, so the entries in this list vary greatly in plausibility and relevance. I made some attempt to categorize these entries and merge similar ones, but this document is fundamentally a brainstorm, not a taxonomy, so keep your expectations low.

I hope to update this post as new ideas find me and old ideas are refined or refuted. I welcome suggestions and criticisms; email me (gmail kokotajlod) or leave a comment.

Interactive "Generate Future" button

Asya Bergal and I made an interactive button to go with the list. The button randomly generates a possible future according to probabilities that you choose. It is very crude, but it has been fun to play with, and perhaps even slightly useful. For example, once I decided that my credences were probably systematically too high because the futures generated with them were too crazy. Another time I used the alternate method (described below) to recursively generate a detailed future trajectory, written up here. I hope to make more trajectories like this in the future, since I think this method is less biased than the usual method for imagining detailed futures.[6]

To choose probabilities, scroll down to the list below and fill each box with a number representing how likely you think the entry is to occur in a strategically relevant way prior to the advent of advanced AI. (1 means certainly, 0 means certainly not. **The boxes are all 0 by default.**) Once you are done, scroll back up and click the button.

A major limitation is that the button doesn't take correlations between possibilities into account. The user needs to do this themselves, e.g. by redoing any generated future that seems silly, or by flipping a coin to choose between two generated possibilities that seem contradictory, or by choosing between them based on what else was generated.

Here is an alternate way to use this button that mostly avoids this limitation:

1. Fill all the boxes with probability-of-happening-in-the-next-5-years (instead of happening before advanced AGI, as in the default method)
2. Click the "Generate Future" button and record the results, interpreted as what happens in the next 5 years.
3. Update the probabilities accordingly to represent the upcoming 5-year period, in light of what has happened so far.
4. Repeat steps 2 – 4 until satisfied. I used a random number generator to determine whether AGI arrived each year.

If you don't want to choose probabilities yourself, click "fill with pre-set values" to populate the fields with my non-expert, hasty guesses.[7]

GENERATE FUTURE
Fill with pre-set values (default method)
Fill with pre-set values (alternate method)

Key

Letters after list titles indicate that I think the change might be relevant to:

- TML: Timelines—how long it takes for advanced AI to be developed
- TAS: Technical AI safety—how easy it is (on a technical level) to make advanced AI safe, or what sort of technical research needs to be done
- POL: Policy—how easy it is to coordinate relevant actors to mitigate risks from AI, and what policies are relevant to this.
- CHA: Chaos—how chaotic the world is.[8]
- MIS: Miscellaneous

Each possibility is followed by some explanation or justification where necessary, and a non-exhaustive list of ways the possibility may be relevant to AI outcomes in particular (which is not guaranteed to cover the most important ones). Possibilities are organized into loose categories created after the list was generated.

List of strategically relevant possibilities

Inputs to AI

1. Advanced science automation and research tools (TML, TAS, CHA, MIS)

Narrow research and development tools might speed up technological progress in general or in specific domains. For example, several of the other technologies on this list might be achieved with the help of narrow research and development tools.

2. Dramatically improved computing hardware (TML, TAS, POL, MIS)

By this I mean computing hardware improves at least as fast as Moore's Law. Computing hardware has historically become steadily cheaper, though it is unclear whether this trend will continue. Some example pathways by which hardware might improve at least moderately include:

- Ordinary scale economies[9]
- Improved data locality[10]
- Increased specialization for specific AI applications[11]
- Optical computing[12]
- Neuromorphic chips[13]
- 3D integrated circuits[14]
- Wafer-scale chips[15]
- Quantum computing[16]
- Carbon nanotube field-effect transistors[17]

Dramatically improved computing hardware may:

- Cause any given AI capability to arrive earlier
- Increase the probability of hardware overhang.
- Affect which kinds of AI are developed first (e.g. those which are more compute-intensive.)
- Affect AI policy, e.g. by changing the relative importance of hardware vs. research talent

3. Stagnation in computing hardware progress (TML, TAS, POL, MIS)

Many forecasters think Moore's Law will be ending soon (as of 2020).[18] In the absence of successful new technologies, computing hardware could progress substantially more slowly than Moore's Law would predict.

Stagnation in computing hardware progress may:

- Cause any given AI capability to arrive later
- Decrease the probability of hardware overhang.
- Affect which kinds of AI are developed first (e.g. those which are less compute-intensive.)
- Influence the relative strategic importance of hardware compared to researchers
- Make energy and raw materials a greater part of the cost of computing

4. Manufacturing consolidation (POL)

Chip fabrication has become more specialized and consolidated over time, to the point where all of the hardware relevant to AI research depends on production from a handful of locations.[19] Perhaps this trend will continue.

One country (or a small number working together) could control or restrict AI research by controlling the production and distribution of necessary hardware.

5. Advanced additive manufacturing (e.g. 3D printing or nanotechnology) (TML, CHA)

Advanced additive manufacturing could lead to various materials, products and forms of capital being cheaper and more broadly accessible, as well as to new varieties of them becoming feasible and quicker to develop. For example, sufficiently advanced 3D printing could destabilize the world by allowing almost anyone to secretly produce

terror weapons. If nanotechnology advances rapidly, so that nanofactories can be created, the consequences could be dramatic:[20]

- Greatly reduced cost of most manufactured products
- Greatly faster growth of capital formation
- Lower energy costs
- New kinds of materials, such as stronger, lighter spaceship hulls
- Medical nanorobots
- New kinds of weaponry and other disruptive technologies

6. Massive resource glut (TML, TAS, POL, CHA)

By "glut" I don't necessarily mean that there is too much of a resource. Rather, I mean that the real price falls dramatically. Rapid decreases in the price of important resources have happened before.[21] It could happen again via:

- Cheap energy (e.g. fusion power, He-3 extracted from lunar regolith,[22] methane hydrate extracted from the seafloor,[23] cheap solar energy[24])
- A source of abundant cheap raw materials (e.g. asteroid mining,[25] undersea mining[26])
- Automation of relevant human labor. Where human labor is an important part of the cost of manufacturing, resource extraction, or energy production, automating labor might substantially increase economic growth, which might result in a greater amount of resources devoted to strategically relevant things (such as AI research) which is relevantly similar to a price drop even if technically the price doesn't drop.[27] and therefore investment in AI.

My impression is that energy, raw materials, and unskilled labor combined are less than half the cost of computing, so a decrease in the price of one of these (and possibly even all three) would probably not have large direct consequences on the price of computing.[28] But a resource glut might lead to general economic prosperity, with many subsequent effects on society, and moreover the cost structure of computing may change in the future, creating a situation where a resource glut could dramatically lower the cost of computing.[29]

7. Hardware overhang (TML, TAS, POL)

Hardware overhang refers to a situation where large quantities of computing hardware can be diverted to running powerful AI systems as soon as the AI software is developed.

If advanced AGI (or some other powerful software) appears during a period of hardware overhang, its capabilities and prominence in the world could grow very quickly.

8. Hardware underhang (TML, TAS, POL)

The opposite of hardware overhang might happen. Researchers may understand how to build advanced AGI at a time when the requisite hardware is not yet available. For example, perhaps the relevant AI research will involve expensive chips custom-built for the particular AI architecture being trained.

A successful AI project during a period of hardware underhang would not be able to instantly copy the AI to many other devices, nor would they be able to iterate quickly and make an architecturally improved version.

Technical tools

9. Prediction tools (TML, TAS, POL, CHA, MIS)

Tools may be developed that are dramatically better at predicting some important aspect of the world; for example, technological progress, cultural shifts, or the outcomes of elections, military clashes, or research projects. Such tools could for instance be based on advances in AI or other algorithms, prediction markets, or improved scientific understanding of forecasting (e.g. lessons from the Good Judgment Project).

Such tools might conceivably increase stability via promoting accurate beliefs, reducing surprises, errors or unnecessary conflicts. However they could also conceivably promote instability via conflict encouraged by a powerful new tool being available to a subset of actors. Such tools might also help with forecasting the arrival and effects of advanced AGI, thereby helping guide policy and AI safety work. They might also accelerate timelines, for instance by assisting project management in general and notifying potential investors when advanced AGI is within reach.

10. Persuasion tools (POL, CHA, MIS)

Present technology for influencing a person's beliefs and behavior is crude and weak, relative to what one can imagine. Tools may be developed that more reliably steer a person's opinion and are not so vulnerable to the victim's reasoning and possession of evidence. These could involve:

- Advanced understanding of how humans respond to stimuli depending on context, based on massive amounts of data
- Coaching for the user on how to convince the target of something
- Software that interacts directly with other people, e.g. via text or email

Strong persuasion tools could:

- Allow a group in conflict who has them to quickly attract spies and then infiltrate an enemy group
- Allow governments to control their populations
- Allow corporations to control their employees
- Lead to a breakdown of collective epistemology[30]

11. Theorem provers (TAS)

Powerful theorem provers might help with the kinds of AI alignment research that involve proofs or help solve computational choice problems.

12. Narrow AI for natural language processing (TML, TAS, CHA)

Researchers may develop narrow AI that understands human language well, including concepts such as "moral" and "honest."

Natural language processing tools could help with many kinds of technology, including AI and various AI safety projects. They could also help enable AI arbitration systems. If researchers develop software that can autocomplete code—much as it currently autocompletes text messages—it could multiply software engineering productivity.

13. AI interpretability tools (TML, TAS, POL)

Tools for understanding what a given AI system is thinking, what it wants, and what it is planning would be useful for AI safety.[31]

14. Credible commitment mechanisms (POL, CHA)

There are significant restrictions on which contracts governments are willing and able to enforce–for example, they can't enforce a contract to try hard to achieve a goal, and won't enforce a contract to commit a crime. Perhaps some technology (e.g. lie detectors, narrow AI, or blockchain) could significantly expand the space of possible credible commitments for some relevant actors: corporations, decentralized autonomous organizations, crowds of ordinary people using assurance contracts, terrorist cells, rogue AGIs, or even individuals.

This might destabilize the world by making threats of various kinds more credible, for various actors. It might stabilize the world in other ways, e.g. by making it easier for some parties to enforce agreements.

15. Better coordination tools (POL, CHA, MIS)

Technology for allowing groups of people to coordinate effectively could improve, potentially avoiding losses from collective choice problems, helping existing large groups (e.g. nations and companies) to make choices in their own interests, and producing new forms of coordinated social behavior (e.g. the 2010's saw the rise of the Facebook group)). Dominant assurance contracts,[32] improved voting systems,[33] AI arbitration systems, lie detectors, and similar things not yet imagined might significantly improve the effectiveness of some groups of people.

If only a few groups use this technology, they might have outsized influence. If most groups do, there could be a general reduction in conflict and increase in good judgment.

Human effectiveness

16. Deterioration of collective epistemology (TML, TAS, POL, CHA, MIS)

Society has mechanisms and processes that allow it to identify new problems, discuss them, and arrive at the truth and/or coordinate a solution. These processes might deteriorate. Some examples of things which might contribute to this:

- Increased investment in online propaganda by more powerful actors, perhaps assisted by chatbots, deepfakes and persuasion tools
- Echo chambers, filter bubbles, and online polarization, perhaps driven in part by recommendation algorithms
- Memetic evolution in general might intensify, increasing the spreadability of ideas/topics at the expense of their truth/importance[34]
- Trends towards political polarization and radicalization might exist and continue
- Trends towards general institutional dysfunction might exist and continue

This could cause chaos in the world in general, and lead to many hard-to-predict effects. It would likely make the market for influencing the course of AI development less efficient (see section on "Landscape of…" below) and present epistemic hazards for anyone trying to participate effectively.

17. New and powerful forms of addiction (TML, POL, CHA, MIS)

Technology that wastes time and ruins lives could become more effective. The average person spends 144 minutes per day on social media, and there is a clear upward trend in this metric.[35] The average time spent watching TV is even greater.[36] Perhaps this time is not wasted but rather serves some important recuperative, educational, or other function. Or perhaps not; perhaps instead the effect of social media on society is like the effect of a new addictive drug — opium, heroin, cocaine, etc. — which causes serious damage until society adapts. Maybe there will be more things like this: extremely addictive video games, or newly invented drugs, or wireheading (directly stimulating the reward circuitry of the brain).[37]

This could lead to economic and scientific slowdown. It could also concentrate power and influence in fewer people—those who for whatever reason remain relatively unaffected by the various productivity-draining technologies. Depending on how these practices spread, they might affect some communities more or sooner than others.

18. Medicine to boost human mental abilities (TML, CHA, MIS)

To my knowledge, existing "study drugs" such as modafinil don't seem to have substantially sped up the rate of scientific progress in any field. However, new drugs (or other treatments) might be more effective. Moreover, in some fields, researchers typically do their best work at a certain age. Medicine which extends this period of peak mental ability might have a similar effect.

This could speed up the rate of scientific progress in some fields, among other effects.

19. Genetic engineering, human cloning, iterated embryo selection (TML, POL, CHA, MIS)

Changes in human capabilities or other human traits via genetic interventions[38] could affect many areas of life. If the changes were dramatic, they might have a large impact even if only a small fraction of humanity were altered by them.

Changes in human capabilities or other human traits via genetic interventions might:

- Accelerate research in general
- Differentially accelerate research projects that depend more on "genius" and less on money or experience
- Influence politics and ideology
- Cause social upheaval
- Increase the number of people capable of causing great harm
- Have a huge variety of effects not considered here, given the ubiquitous relevance of human nature to events
- Shift the landscape of effective strategies for influencing AI development (see below)

20. Landscape of effective strategies for influencing AI development changes substantially (CHA, MIS)

For a person at a time, there is a landscape of strategies for influencing the world, and in particular for influencing AI development and the effects of advanced AGI. The landscape could change such that the most effective strategies for influencing AI development are:

- More or less reliably helpful (e.g. working for an hour on a major unsolved technical problem might have a low chance of a very high payoff, and so not be

very reliable)
- More or less “outside the box” (e.g. being an employee, publishing academic papers, and signing petitions are normal strategies, whereas writing Harry Potter fanfiction to illustrate rationality concepts and inspire teenagers to work on AI safety is not)[39]
- Easier or harder to find, such that marginal returns to investment in strategy research change

Here is a non-exhaustive list of reasons to think these features might change systematically over time:

- As more people devote more effort to achieving some goal, one might expect that effective strategies become common, and it becomes harder to find novel strategies that perform better than common strategies. As advanced AI becomes closer, one might expect more effort to flow into influencing the situation. Currently some ‘markets’ are more efficient than others; in some the orthodox strategies are best or close to the best, whereas in others clever and careful reasoning can find strategies vastly better than what most people do. How efficient a market is depends on how many people are genuinely trying to compete in it, and how accurate their beliefs are. For example, the stock market and the market for political influence are fairly efficient, because many highly-knowledgeable actors are competing. As more people take interest, the ‘market’ for influencing the course of AI may become more efficient. (This would also decrease the marginal returns to investment in strategy research, by making orthodox strategies closer to optimal.) If there is a deterioration of social epistemology (see below), the market might instead become less efficient.
- Currently there are some tasks at which the most skilled people are not much better than the average person (e.g. manual labor, voting) and others in which the distribution of effectiveness is heavy-tailed, such that a large fraction of the total influence comes from a small fraction of individuals (e.g. theoretical math, donating to politicians). The types of activity that are most useful for influencing the course of AI development may change over time in this regard, which in turn might affect the strategy landscape in all three ways described above.
- Transformative technologies can lead to new opportunities and windfalls for people who recognize them early. As more people take interest, opportunities for easy success disappear. Perhaps there will be a burst of new technologies prior to advanced AGI, creating opportunities for unorthodox or risky strategies to be very successful.

A shift in the landscape of effective strategies for influencing the course of AI is relevant to anyone who wants to have an effective strategy for influencing the course of AI.[40] If it is part of a more general shift in the landscape of effective strategies for other goals — e.g. winning wars, making money, influencing politics — the world could be significantly disrupted in ways that may be hard to predict.

21. Global economic collapse (TML, CHA, MIS)

This might slow down research or precipitate other relevant events, such as war.

22. Scientific stagnation (TML, TAS, POL, CHA, MIS)

There is some evidence that scientific progress in general might be slowing down. For example, the millennia-long trend of decreasing economic doubling time seems to have stopped around 1960.[41] Meanwhile, scientific progress has arguably come from

increased investment in research. Since research investment has been growing faster than the economy, it might eventually saturate and grow only as fast as the economy.[42]

This might slow down AI research, making the events on this list (but not the technologies) more likely to happen before advanced AGI.

23. Global catastrophe (TML, POL, CHA)

Here are some examples of potential global catastrophes:

- Climate change tail risks, e.g. feedback loop of melting permafrost releasing methane[43]
- Major nuclear exchange
- Global pandemic
- Volcano eruption that leads to 10% reduction in global agricultural production[44]
- Exceptionally bad solar storm knocks out world electrical grid[45]
- Geoengineering project backfires or has major negative side-effects[46]

A global catastrophe might be expected to cause conflict and slowing of projects such as research, though it could also conceivably increase attention on projects that are useful for dealing with the problem. It seems likely to have other hard to predict effects.

Attitudes toward AGI

24. Shift in level of public attention on AGI (TML, POL, CHA, MIS)

The level of attention paid to AGI by the public, governments, and other relevant actors might increase (e.g. due to an impressive demonstration or a bad accident) or decrease (e.g. due to other issues drawing more attention, or evidence that AI is less dangerous or imminent).

Changes in the level of attention could affect the amount of work on AI and AI safety. More attention could also lead to changes in public opinion such as panic or an AI rights movement.

If the level of attention increases but AGI does not arrive soon thereafter, there might be a subsequent period of disillusionment.

25. Change in investment in AGI development (TML, TAS, POL)

There could be a rush for AGI, for instance if major nations begin megaprojects to build it. Or there could be a rush away from AGI, for instance if it comes to be seen as immoral or dangerous like human cloning or nuclear rocketry.

Increased investment in AGI might make advanced AGI happen sooner, with less hardware overhang and potentially less proportional investment in safety. Decreased investment might have the opposite effects.

26. New social movements or ideological shifts (TML, TAS, POL, MIS)

The communities that build and regulate AI could undergo a substantial ideological shift. Historically, entire nations have been swept by radical ideologies within about a decade or so, e.g. Communism, Fascism, the Cultural Revolution, and the First Great

Awakening.[47] Major ideological shifts within communities smaller than nations (or within nations, but on specific topics) presumably happen more often. There might even appear powerful social movements explicitly focused on AI, for instance in opposition to it or attempting to secure legal rights and moral status for AI agents.[48] Finally, there could be a general rise in extremist movements, for instance due to a symbiotic feedback effect hypothesized by some,[49] which might have strategically relevant implications even if mainstream opinions do not change.

Changes in public opinion on AI might change the speed of AI research, change who is doing it, change which types of AI are developed or used, and limit or alter discussion. For example, attempts to limit an AI system's effects on the world by containing it might be seen as inhumane, as might adversarial and population-based training methods. Broader ideological change or a rise in extremisms might increase the probability of a massive crisis, revolution, civil war, or world war.

27. Harbinger of AGI (ALN, POL, MIS)

Events could occur that provide compelling evidence, to at least a relevant minority of people, that advanced AGI is near.

This could increase the amount of technical AI safety work and AI policy work being done, to the extent that people are sufficiently well-informed and good at forecasting. It could also enable people already doing such work to more efficiently focus their efforts on the true scenario.

28. AI alignment warning shot (ALN, POL)

A convincing real-world example of AI alignment failure could occur.

This could motivate more effort into mitigating AI risk and perhaps also provide useful evidence about some kinds of risks and how to avoid them.

Precursors to AGI

29. Brain scanning (TML, TAS, POL, CHA, MIS)

An accurate way to scan human brains at a very high resolution could be developed.

Combined with a good low-level understanding of the brain (see below) and sufficient computational resources, this might enable brain emulations, a form of AGI in which the AGI is similar, mentally, to some original human. This would change the kind of technical AI safety work that would be relevant, as well as introducing new AI policy questions. It would also likely make AGI timelines easier to predict. It might influence takeoff speeds.

30. Good low-level understanding of the brain (TML, TAS, POL, CHA, MIS)

To my knowledge, as of April 2020, humanity does not understand how neurons work well enough to accurately simulate the behavior of a C. Elegans worm, though all connections between its neurons have been mapped[50] Ongoing progress in modeling individual neurons could change this, and perhaps ultimately allow accurate simulation of entire human brains.

Combined with brain scanning (see above) and sufficient computational resources, this may enable brain emulations, a form of AGI in which the AI system is similar,

mentally, to some original human. This would change the kind of AI safety work that would be relevant, as well as introducing new AI policy questions. It would also likely make the time until AGI is developed more predictable. It might influence takeoff speeds. Even if brain scanning is not possible, a good low-level understanding of the brain might speed AI development, especially of systems that are more similar to human brains.

31. Brain-machine interfaces (TML, TAS, POL, CHA, MIS)

Better, safer, and cheaper methods to control computers directly with our brains may be developed. At least one project is explicitly working towards this goal.[51]

Strong brain-machine interfaces might:

- Accelerate research, including on AI and AI safety[52]
- Accelerate in vitro brain technology
- Accelerate mind-reading, lie detection, and persuasion tools
- Deteriorate collective epistemology (e.g. by contributing to wireheading or short attention spans)
- Improve collective epistemology (e.g. by improving communication abilities)
- Increase inequality in influence among people

32. In vitro brains (TML, TAS, POL, CHA)

Neural tissue can be grown in a dish (or in an animal and transplanted) and connected to computers, sensors, and even actuators.[53] If this tissue can be trained to perform important tasks, and the technology develops enough, it might function as a sort of artificial intelligence. Its components would not be faster than humans, but it might be cheaper or more intelligent. Meanwhile, this technology might also allow fresh neural tissue to be grafted onto existing humans, potentially serving as a cognitive enhancer.[54]

This might change the sorts of systems AI safety efforts should focus on. It might also automate much human labor, inspire changes in public opinion about AI research (e.g. promoting concern about the rights of AI systems), and have other effects which are hard to predict.

33. Weak AGI (TML, TAS, POL, CHA, MIS)

Researchers may develop something which is a true artificial general intelligence—able to learn and perform competently all the tasks humans do—but just isn't very good at them, at least, not as good as a skilled human.

If weak AGI is faster or cheaper than humans, it might still replace humans in many jobs, potentially speeding economic or technological progress. Separately, weak AGI might provide testing opportunities for technical AI safety research. It might also change public opinion about AI, for instance inspiring a "robot rights" movement, or an anti-AI movement.

34. Expensive AGI (TML, TAS, POL, CHA, MIS)

Researchers may develop something which is a true artificial general intelligence, and moreover is qualitatively more intelligent than any human, but is vastly more expensive, so that there is some substantial period of time before cheap AGI is developed.

An expensive AGI might contribute to endeavors that are sufficiently valuable, such as some science and technology, and so may have a large effect on society. It might also prompt increased effort on AI or AI safety, or inspire public thought about AI that produces changes in public opinion and thus policy, e.g. regarding the rights of machines. It might also allow opportunities for trialing AI safety plans prior to very widespread use.

35. Slow AGI (TML, TAS, POL, CHA, MIS)

Researchers may develop something which is a true artificial general intelligence, and moreover is qualitatively as intelligent as the smartest humans, but takes a lot longer to train and learn than today's AI systems.

Slow AGI might be easier to understand and control than other kinds of AGI, because it would train and learn more slowly, giving humans more time to react and understand it. It might produce changes in public opinion about AI.

36. Automation of human labor (TML, TAS, POL, CHA, MIS)

If the pace of automation substantially increases prior to advanced AGI, there could be social upheaval and also dramatic economic growth. This might affect investment in AI.

Shifts in the balance of power

37. Major leak of AI research (TML, TAS, POL, CHA)

Edward Snowden defected from the NSA and made public a vast trove of information. Perhaps something similar could happen to a leading tech company or AI project.

In a world where much AI progress is hoarded, such an event could accelerate timelines and make the political situation more multipolar and chaotic.

38. Shift in favor of espionage (POL, CHA, MIS)

Espionage techniques might become more effective relative to counterespionage techniques. In particular:

- Quantum computing could break current encryption protocols.[55]
- Automated vulnerability detection[56] could turn out to have an advantage over automated cyberdefense systems, at least in the years leading up to advanced AGI.

More successful espionage techniques might make it impossible for any AI project to maintain a lead over other projects for any substantial period of time. Other disruptions may become more likely, such as hacking into nuclear launch facilities, or large scale cyberwarfare.

39. Shift in favor of counterespionage (POL, CHA, MIS)

Counterespionage techniques might become more effective relative to espionage techniques than they are now. In particular:

- Post-quantum encryption might be secure against attack by quantum computers.[57]

- Automated cyberdefense systems could turn out to have an advantage over automated vulnerability detection. Ben Garfinkel and Allan Dafoe[58] give reason to think the balance will ultimately shift to favor defense.

Stronger counterespionage techniques might make it easier for an AI project to maintain a technological lead over the rest of the world. Cyber wars and other disruptive events could become less likely.

40. Broader or more sophisticated surveillance (POL, CHA, MIS)

More extensive or more sophisticated surveillance could allow strong and selective policing of technological development. It would also have other social effects, such as making totalitarianism easier and making terrorism harder.

41. Autonomous weapons (POL, CHA)

Autonomous weapons could shift the balance of power between nations, or shift the offense-defense balances resulting in more or fewer wars or terrorist attacks, or help to make totalitarian governments more stable. As a potentially early, visible and controversial use of AI, they may also especially influence public opinion on AI more broadly, e.g. prompting anti-AI sentiment.

42. Shift in importance of governments, corporations, and other groups in AI development (POL, CHA)

Currently both governments and corporations are strategically relevant actors in determining the course of AI development. Perhaps governments will become more important, e.g. by nationalizing and merging AI companies. Or perhaps governments will become less important, e.g. by not paying attention to AI issues at all, or by becoming less powerful and competent generally. Perhaps some third kind of actor (such as religion, insurgency, organized crime, or special individual) will become more important, e.g. due to persuasion tools, countermeasures to surveillance, or new weapons of guerilla warfare.[59]

This influences AI policy by affecting which actors are relevant to how AI is developed and deployed.

43. Catastrophe in strategically important location (TML, POL, CHA, MIS)

Perhaps some strategically important location (e.g. tech hub, seat of government, or chip fab) will be suddenly destroyed. Here is a non-exhaustive list of ways this could happen:

- Terrorist attack with weapon of mass destruction
- Major earthquake, flood, tsunami, etc. (e.g. this research claims a 2% chance of magnitude 8.0 or greater earthquake in San Francisco by 2044.)[60]

If it happens, it might be strategically disruptive, causing e.g. the dissolution and diaspora of the front-runner AI project, or making it more likely that some government makes a radical move of some sort.

44. Change in national AI research loci (POL, CHA)

For instance, a new major national hub of AI research could arise, rivalling the USA and China in research output. Or either the USA or China could cease to be relevant to

AI research.

This might make coordinating AI policy more difficult. It might make a rush for AGI more or less likely.

45. Large war (TML, POL, CHA, MIS)

This might cause short-term, militarily relevant AI capabilities research to be prioritized over AI safety and foundational research. It could also make global coordination on AI policy difficult.

46. Civil war or regime change in major relevant countries (POL, CHA, MIS)

This might be very dangerous for people living in those countries. It might change who the strategically relevant actors are for shaping AI development. It might result in increased instability, or cause a new social movement or ideological shift.

47. Formation of a world government (POL, CHA)

This would make coordinating AI policy easier in some ways (e.g. there would be no need for multiple governing bodies to coordinate their policy at the highest level), however it might be harder in others (e.g. there might be a more complicated regulatory system overall).

18 June 2020.

Notes

(Edited to add text)

What is meant by Simulcra Levels?

Simulcra levels are a concept that have seen a lot of play on Less Wrong. I suspect that different people are using these concepts in slightly different ways, so I thought it might make sense to ask a question to provide a central location for recording these theories and helping people disambiguate. For this reason, I don't think that there should be a single answer, but rather multiple answers. If you propose a theory, I'd suggest ideally providing a descriptive title or label.

Using a memory palace to memorize a textbook.

I spent the week prepping for finals. One is a year-long cumulative closed-book chemistry exam that I haven't had much time to practice for. I was worried about memorizing a few things:

- Periodic trends and exceptions
- The form and application of approximately 100 workhorse equations and various forms of measurement (molarity vs. molality vs. mole fraction).
- Equations that get used rarely in homework or on exercises, but might be used as "gotchas" on the test.
- Some concepts that I found either confusing, or so simple that I didn't bother to remember them the first time.

My anxiety wasn't just my ability to recall these ideas when prompted:

> "What's the two-point form of the Clausius-Clapeyron Equation?"
>
> $\ln(P2 / P1) = - \Delta Hvap/R * (1/T2 - 1/T1)$

Nor was I unable to perform the calculations.

My real concern was that I had spent the year treating my chemistry textbook like a reference manual, a repository for concepts and equations that I could look up when needed. I just memorized the few bits I'd need on any given quiz. Looking back at 1,000 pages of chemistry, I foresaw myself reviewing chapter 5 for a couple hours, but forgetting that review by the time I got to chapter 19.

The sheer volume of work that seemed to be involved in memorizing a textbook seemed unreasonable. I hate using Anki, and I spend far too much time in front of screens as it is.

So I decided to try something different - experimenting with the memory palace technique.

I perceive myself as having a poor visual imagination, but I've been trying to practice improving it lately, with some success. Gwern points to expert opinion that visual thinking ability might be second only to IQ in terms of intellectual importance. My experience is that when I'm using psychedelics, or deliberately practicing my visualization abilities, I do improve far beyond my perceived abilities. We're stuck with our IQ, but if it's possible to improve our visual thinking skills through practice in adulthood, that's important.

I want to describe my attempts and the outcome.

First Room

I tried this both with a single calculus textbook chapter, and my entire chemistry textbook. The results were similar but different. I'm going to focus on the chemistry palace here.

I close my eyes and allow myself to picture nothing, or whatever random nonsense comes to mind. No attempt to control.

Then I invite the concept of a room into mind. I don't picture it clearly. There's a vague sense, though, of imagining a space of some kind. I can vaguely see fleeting shadowy walls. I don't need to get everything crystal clear, though.

I mentally label the room as the "Ch. 14 room," or the "rates room." That means doing lots of things to make the label stick. I speak the words in my head. I picture a banner with them printed on it hanging from the ceiling. Or if I can't see it clearly, I picture a banner-like thing and just *know* that it says "rates room." I picture hourglasses sitting on furniture - the image comes to me much more easily than a banner with text.

I imagine the crucial equations sitting on columnar pedestals. Again, they are easier to picture for some reason. I make sure that I can visually see each piece of the equation. I imagine a label on the pedestal - one says "t1/2" for the half-life equations; the other says "Integrated rate law," with an hourglass made out of two intertwined integration signs.

I look up a picture of Svante Arrhenius and picture him in the room. He takes on a life of his own. I can tell he's proud of his equation, which appears in bold letters at the back of the room, with a sort of curtain around it. He's the keeper of the room. It takes on a calm atmosphere here. He's also the doorman. I have to tell him how to calculate the overall reaction order in order to enter. But if *he knows* that *I know* how to do it, I don't have to explain it in as much detail. We have a psychic relationship.

Second Room

Moving backwards to Ch. 13, I once again imagine a new room, the Solutions Room. Standing there, I can still see the entrance to the first room - I can even picture some of the things inside, from a distance. I start populating the room with symbols, objects, equations, and the chemists they're named after. They are happy to explain things to me as many times as necessary.

Abstract concepts that the book presents in words, still images, or equations get visualized in new ways. Partial pressures become two beakers, one with yellow steam and the other with red steam emerging. They get mixed into a single beaker that now emits a mixture of yellow and red steam, somewhere in between the amounts that the yellow and red beaker emit on their own. François-Marie Raoult is standing by to demonstrate his law to me. There's a bottle of Coke with Henry's Law printed on it.

The solubility rules are accessible when I glance at the periodic table on the wall. Rather than seeing a list of rules, I see the individual elements, which take on a life of their own. The alkali metals, ammonium, and nitrate zoom around the room, not interested in talking to anybody, on their own adventure. The halogens are too cool to talk to anybody except silver, mercury, and lead, who are immensely popular. Silver had a falling out with acetate, who's a communist and not interested in money. Be sensitive! Chromate is a rich chick in an expensive chrome-hubbed car cruising around, looking for a boyfriend. Sulfur is bicurious, so she'll bond not only with the transition metals but with astatine, arsenic, bismuth, and lead.

I practice traveling back and forth between the first and second rooms. They stay remarkably stable. Unlike recalling flash cards or the textbook, when I'm in my memory palace the ideas come almost unbidden. The elemental relationships I've used to conceptualize the solubility rules come *bursting* out of the periodic table.

Further rooms

I continue this for 6 chapters over the course of several hours. I am shocked and delighted at how easy and pleasant it is both to create the memory palace and to access the memories stored there. Not everything goes in - just the bits that I tend to forget. If I'm not sure about something, the famous chemists who populate the rooms will remind me, literally by talking me through their ideas.

The presence of the chemists is also helpful for keeping me focused. I suspect that my brain is recruiting my social motivation. If the only people in my environment are genius chemists who are delighted to keep me interested in chemistry, then why would I get distracted by the internet?

I find it deeply reassuring to stand in the Intermolecular Forces room and know that just by walking a few rooms over, I can get back to the Rates Room, where all the equations are stored. Perhaps I've built a path through the mental mountains? The next day, it's pretty easy to get back to the memory palace, and everything is as I left it. I just have to close my eyes and wait for a moment to get back in.

Concerns and questions

I also did a memory palace for calculus. I did it day-of because I felt more confident about calculus, it wasn't a comprehensive exam, and it was open book. I'll describe it another time. Mostly, it helped me feel more confident that I understood the breadth of the material. I found it much more convenient to refer to the textbook when necessary.

But for tomorrow's, I'm very glad that I now have a store of chemical facts in my memory palace. The anxiety that had been plaguing me this week has vanished. I'm not certain that it will really help. But I do anticipate continuing to use this technique in the future. I think it helps not only my memory but my synthesis of learning.

For example, our chapter on Lewis Structures also introduces the topic of electronegativity and formal charge. Anyone who's taken first year gen chem knows they're related: any negative formal charge should go on the most electronegative atom.

But when I would stare at the electronegativity pages in the textbook, I would focus on the rules offered there: the range of EN difference that characterizes a covalent vs. ionic bond, the periodic trend in EN, and how to calculate net dipole moment. Likewise, in the formal charge section, I would focus on how to calculate the charge.

It took seeing Linus Pauling holding a symbol for electronegativity in one hand, and a symbol for formal charge in the other, to more deeply understand that these are not just two different calculations to do. They're deeply related ways of modeling how molecules are structured. They go together like yeast and flour.

I also see how much faster and more intuitively I think about both chemistry and calculus when I can visualize them. It's just no comparison. Trying to remember Raoult's Law by remembering a verbal description or picturing the equation is just no comparison to looking at those yellow and red steaming beakers. Similarly, it's so helpful to picture a 3D mountain range and see a tiny little yellow gradient vector surfing up and down it on the steepest slopes.

Advice

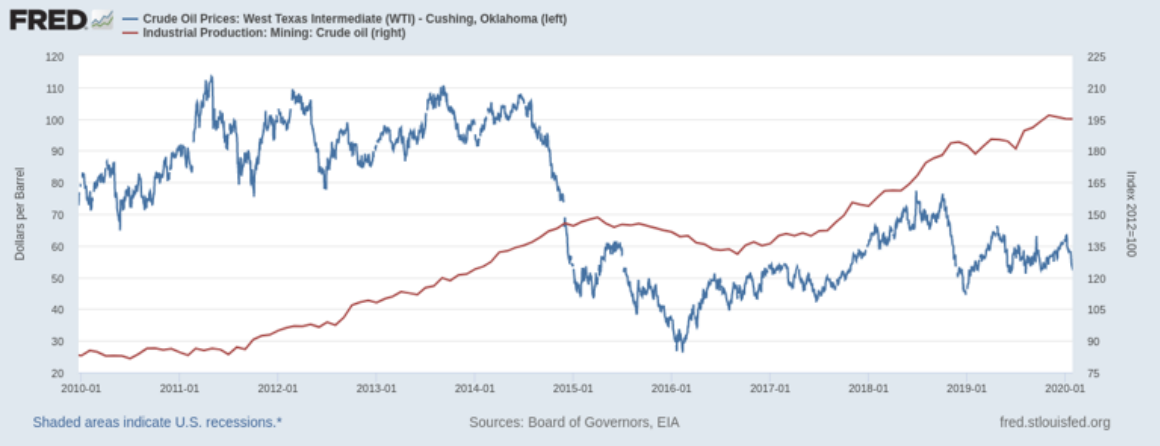
I'm a true beginner here, so I don't want to make any grand claims about how to learn or how useful these techniques are. But I'd give a few pointers so far:

- If you think you can't visualize, you might be wrong.
- Start by just closing your eyes and allowing your brain to produce images without trying to control them. It seems important to have a relaxed, accepting attitude toward my own brain's way of visualizing.
- The way to add control is to take a gentle, experimental attitude. Go with what's easy. Let yourself be surprised by what's do-able and what's useful. Is it hard to picture letters on a banner? Try visualizing Isaac Newton and ask him to say the thing you're trying to remember. Maybe you don't even need to do that - maybe it's enough to picture a vague stripe in the sky that you *just know* is a banner, and you *just know* has the words "rate room" printed on it.
- It takes a while to navigate the rooms and get the information you need, so this might need refinement or practice if you've got to remember things quickly.

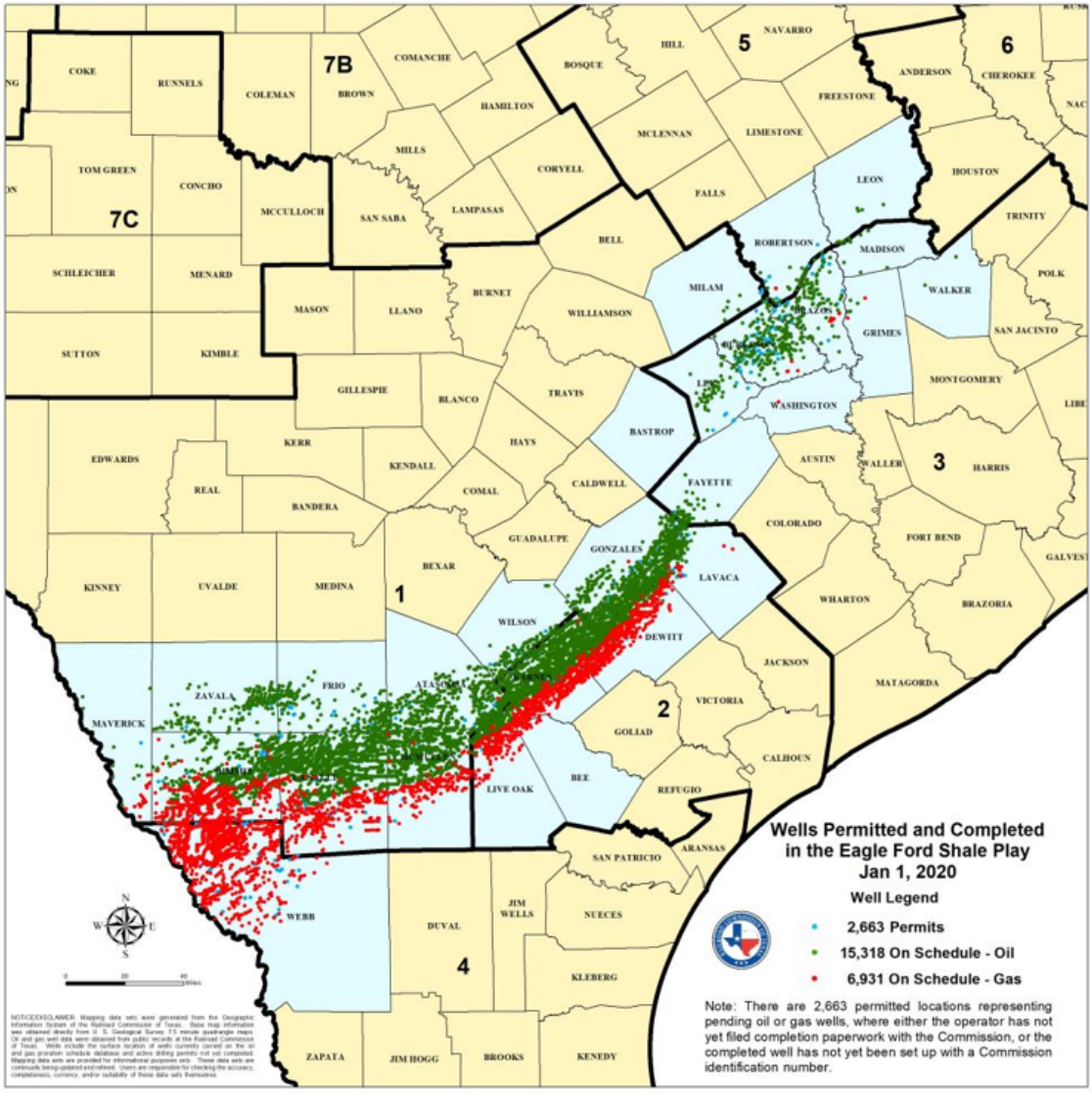
Mediators of History

Epistemic note: all of the examples in this post are very simplified for ease of consumption. The core idea applies just as well to the real systems in all their complicated glory, however.

When oil prices change, oil producers adjust in response - they drill more wells in response to higher prices, or fewer wells in response to lower prices. On the other side of the equation, oil prices adjust to production: when OPEC restricts output, prices rise, and when American shale wells expand, prices fall. We have a feedback loop, which makes it annoying to sort out cause and effect - do prices cause production, or does production cause prices?



The narrative: from roughly 2010-2014, OPEC successfully restricted their production enough to keep oil prices around $100/barrel. But at that price, American shale wells are extremely profitable, and they grew rapidly - the dots in the map below are each an oil/gas well in the Eagle Ford basin in Southern Texas, and the graph above shows American oil production in red. This situation was not sustainable; prices eventually dropped to around $50/barrel, which is roughly the marginal cost of American shale. Since then, prices rose above $50/barrel again around 2018, and American shale once again grew rapidly in response.



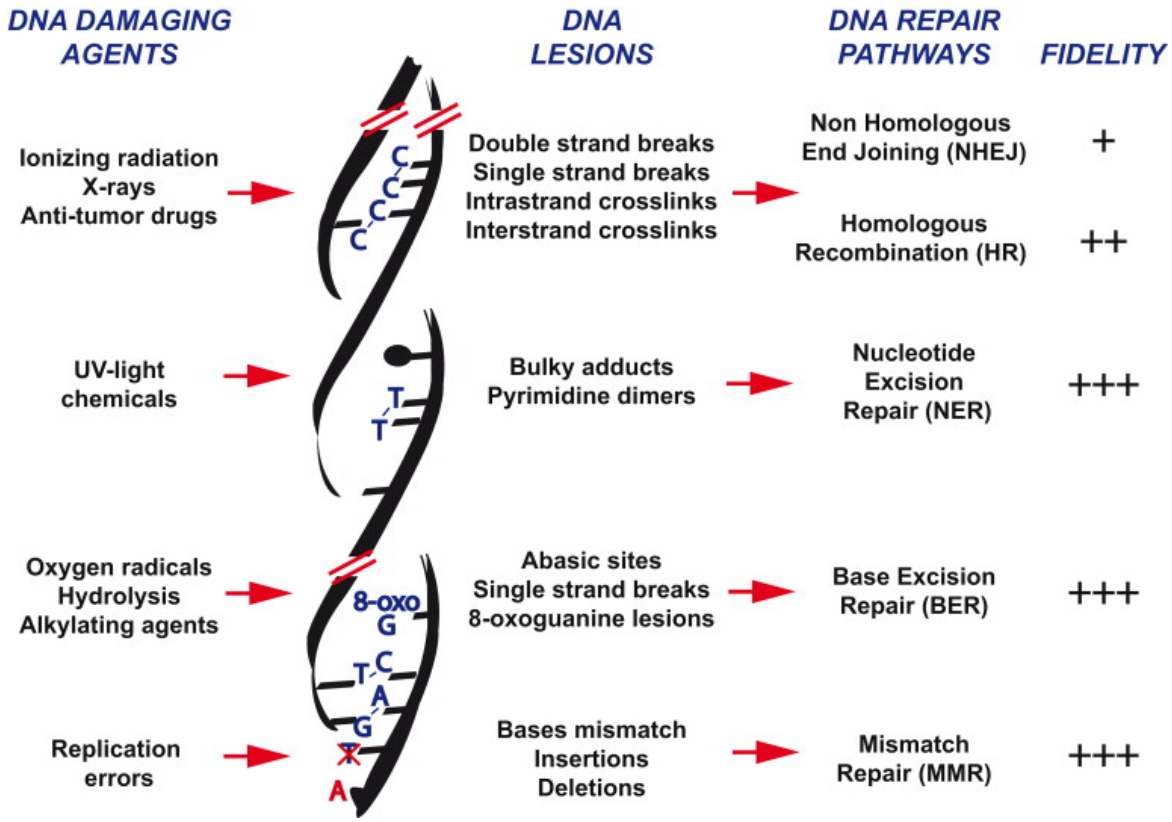
In this case, there is a useful sense in which production capacity causes prices, *not* the other way around - at least if we omit OPEC agreements.

Oil is a fairly liquid commodity. When there's a shock in supply (e.g. OPEC agreeing to restrict output) or demand (e.g. lockdowns), the markets respond and prices rapidly adjust. Production *capacity*, on the other hand, adjusts slowly: drilling new wells and building new pipelines takes time, and once a well is built it rarely makes sense to shut it down before it runs dry. So (ignoring OPEC) prices *right now* are caused by production capacity right now, but production capacity right now is *not* caused by prices right now - it's the result of prices over the past several years, when the wells were drilled.

Or, to put it differently: production capacity mediates the effects of historical prices on current prices. It's a "mediator of history" - a variable which changes slowly enough that it carries information about the past. Other variables equilibrate more quickly, so they depend on far-past values only via the mediators of history.

(Incorporating OPEC into this view is an exercise for the reader.)

Another example: each of our cells' DNA is damaged hundreds or thousands of times per day - things like strand breaks or random molecules stuck on the side. Usually this is rapidly repaired, but occasionally it's misrepaired and a mutation results - a change in the DNA sequence. On the other side, some mutations can increase DNA damage, either by increasing the rate at which it occurs, or reducing the rate at which it's repaired. So damage causes mutations, and mutations can cause damage.



Visualizations of some kinds of DNA damage, and keywords to google if you want to know more about them.

Here again, there is a useful sense in which mutations cause damage, *not* the other way around: the damage *right now* is caused by the mutations right now, but the mutations right now were caused by damage long ago. The mutations are a mediator of history.

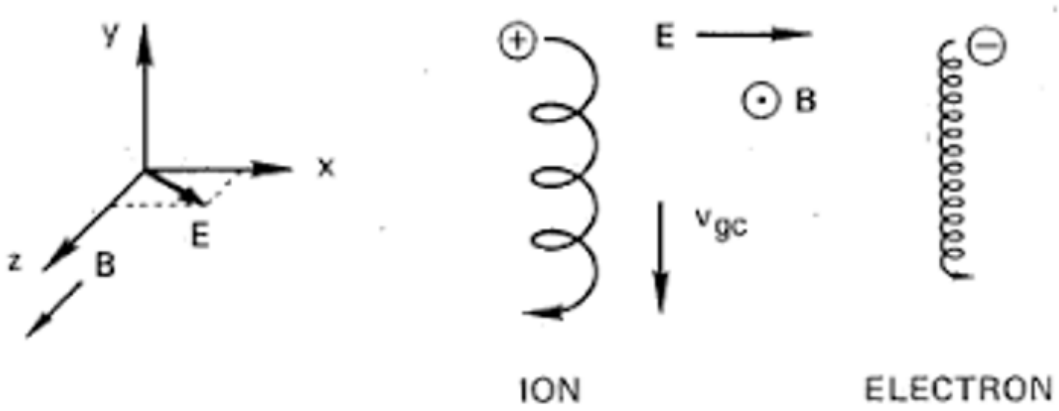
This has important implications for treating disease: we can use antioxidants to suppress (some types of) DNA damage, but that won't remove the underlying mutations. As soon as we stop administering antioxidants, the damage will bounce right back up. Worse, we probably won't prevent all damage, so mutations will still accumulate (albeit at a slower rate), and eventually the antioxidants won't be enough. On the other hand, if we can fix the problematic mutations (e.g. by detecting and removing cells with such mutations), then that "resets" the cells - it's like the earlier damage never happened at all.

Change the mediators of history, and it's like history never happened.

A third example: a robot takes actions and updates its world model in response to incoming data. It uses the world model explicitly to decide which actions to take, but the actions chosen will also indirectly influence the world model - e.g. the robot will see different things and update the model differently depending on where it goes. However, the action being taken *right now* does not influence the world model right now; the world model depends on actions taken previously. So, the world model mediates history.

Here, it's even more obvious that changing the mediator of history makes it like history never happened: if we reset the robot's world model to its original state (and return it to wherever it started in the world), then all the influence of previous actions is erased.

In general, looking for mediators of history is a useful tool for making sense of systems containing feedback loops. In chemistry, it's the fast equilibrium approximation, in which the overall kinetics of a reaction are dominated by a rate-limiting step. In physics more generally, it's timescale separation, useful for separating e.g. wave propagation from material flows in fluid systems.



In plasmas, charged particles follow a wheel-like motion - they orbit around magnetic field lines with drift superposed. When the orbital motion is on a fast timescale relative to the drift, we can average it out - see gyrokinetics.

The most common application of the idea in chemistry and physics is to simplify equations when we're mainly interested in long-term behavior. We can just assume that the fast-equilibrating variables are always in equilibrium, and calculate the rate-of-change of the mediators of history under that assumption. In many systems, only a small fraction of the variables are mediators of history, so this approximation lets us simulate differential equations in far fewer dimensions.

From a modelling perspective, the mediators of history are the "state variables" of the system on long timescales. This is especially important in economic models, since the state variables are what agents in the models need to forecast - e.g. stock traders mainly need to know how mediators of history will behave in the future. If they know that, then the rest is just noise plus an equilibrium calculation.

Finally, in terms of engineering, mediators of history are key targets for control. For instance, if we want to cure aging, then identifying and intervening on the mediators of history is the key problem - they are both a necessary and a sufficient set of intervention targets. That actually simplifies the problem a lot, since the vast majority of biological entities - from molecules to cells - turn over on a very fast timescale, compared to the timescale of aging. So there are probably relatively few mediators of history, relative to the complexity of the whole human body - we just need to look for things which turn over on a timescale of decades or slower (including things which don't equilibrate at all).

Prediction = Compression [Transcript]

(Talk given on Sunday 21st June, over a zoom call with 40 attendees. Alkjash is responsible for the talk, Ben Pace is responsible for the transcription.)

Ben Pace: *Our next speaker is someone you'll all know as Alkjash on LessWrong, who has written an awesome number of posts. Babble and Prune, Hammertime Final Exam – which is one of my favorite names of a curated post on LessWrong. Alkjash, go for it.*

Prediction = Compression Talk

Alkjash: I will be talking about a bit of mathematics today. It's funny that this audience is bigger than any I've gotten in an actual maths talk. It's a bit depressing. Kind of makes me question my life choices...

Alkjash: Hopefully this mathematics is new to some of you. I'm sure that the machine learning people know this already. The principle is that prediction is the same thing as compression. And what that means is that whenever you have a prediction algorithm, you can also get a correspondingly good compression algorithm for data you already have, and vice versa.

Alkjash: Let me dive straight into some examples. Take this sentence and try to fill in the vowels.

> *Ppl bcm wh thy r mnt t b, Mss Grngr, by dng wht s rght.*

Translation:

People become who they are meant to be, Miss Granger, by doing what is right.

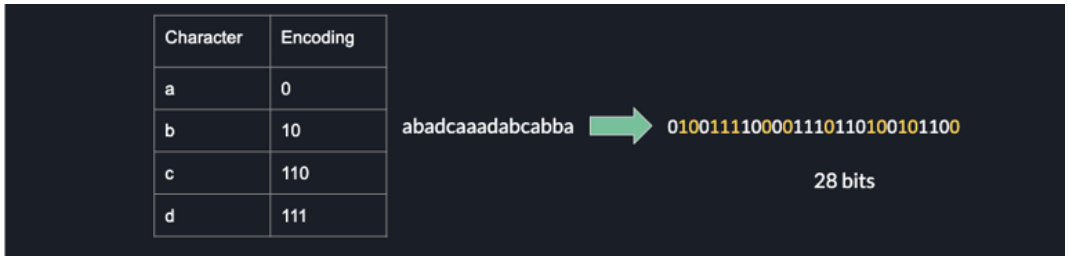
Alkjash: You'll notice that it's actually quite easy, hopefully, to read the sentence, even though I've taken out all the vowels. The human brain is very good at reconstructing certain kinds of missing data, and in some sense, that's a prediction algorithm, but immediately it also gives us a compression algorithm. If I take this sentence and I remove all the vowels, it's transmitting the same amount of data to a human brain because they can predict what's missing.

Alkjash: This is a very simple example where you can predict with certainty what's missing. In general, we don't always have prediction algorithms that can predict the future with certainty. But anytime we can do better than chance, we can also get some compression improvements.

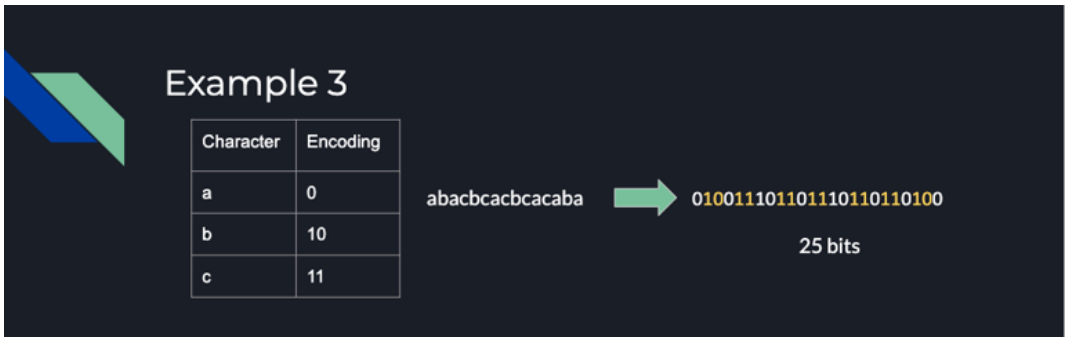
Alkjash: Here's a slightly more sophisticated example. Suppose I have a four character alphabet and I encode some language with a four character alphabet in this obvious default way. Each character is encoded in two bits, then the following string will be encoded in this fifth string.

Alkjash: I've highlighted every other code word for you to just see the emphasis. And you can see that this string has encoded as 32 bits in memory.

Alkjash: However, I might have a little more information or better priors about what this language looks like. I might know for example that As appear more frequently and Cs and Ds appear less frequently. Then I can choose a more suitable encoding for the task. I can choose a different encoding where A is given a shorter bit string because it appears more frequently, and C and D are given longer ones to balance that out. And, using this encoding instead, we see that the same string is encoded only 28 bits.



Alkjash: This time I didn't have any certainty about my predictions, but because I was able to do better than chance about predicting what the next character would be, I achieved a better compression ratio. I can store the same amount of data in memory with less bits.



Alkjash: The third example is very similar, but it's a different sort of prediction. Previously we just used letter frequencies, but now suppose I have a three character alphabet and I encode A as 0, B as 10, and C as 11, then by default I get 25 bits in this encoding.

Alkjash: But now suppose I notice a different sort of pattern about this language, which here, the pattern is that no letter appears consecutively twice in a row. So,

that's the different sort of predictive information from letter frequencies that we've had before.

Alkjash: Perhaps you could think for a moment about how you could use that information to store the same memory in fewer bits. No two As appear twice in a row; no two Bs appear twice in a row; no two Cs appear twice in a row. How would you use that information to compress the data?

[Pause for thinking]

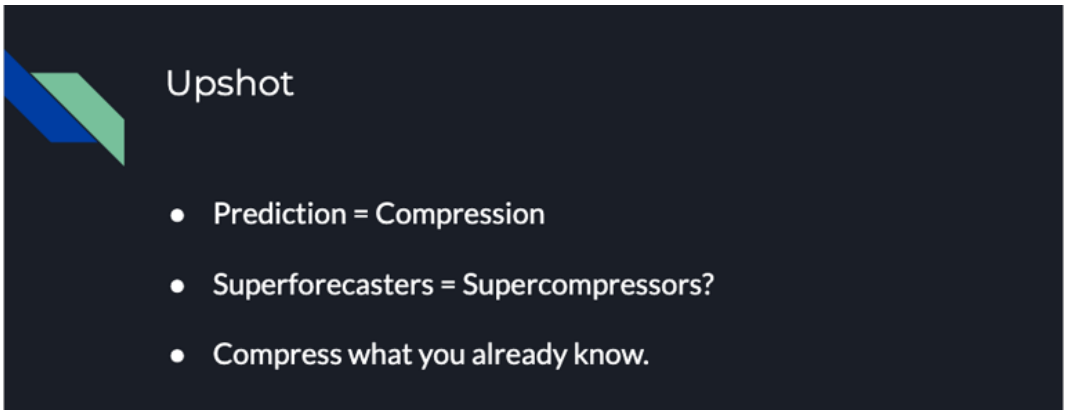
Alkjash: Here's one way you might do it.

Alkjash: I'm going to encode the first letter as whatever it would have been before, and then I'll draw this little triangle and each next letter, because it has to be different, will either be clockwise or counter-clockwise from the previous letter in this triangle that I've drawn.

Alkjash: So if the next letter is clockwise from my previous letter, I put down a zero. If the next letter is counter-clockwise from the previous letter, I put down a 1. And this is a very specific example of some very general algorithm in entropy encoding in the information theory.

Alkjash: So whenever you have any sort of way of better predicting the future than chance, it corresponds to a better data compression algorithm for that type of data, where you store the same amount of information in your bits.

Alkjash: That's really all I have to say. So, I guess the question I want to open up for discussion about what this principle has applications in rationality. For example, we care a lot about super-forecasters predicting the future. Is it true that people who are good at predicting the future are good at storing the same amount of knowledge in their minds very efficiently? Are these very closely related skills? I'd be curious if that's true. Obviously these skills should be correlated with IQ or what not, but perhaps there's an additional layer to that.



Alkjash: And second, I think we spend a lot of time emphasizing prediction and calibration games as tools to improve your rationality, but perhaps we should also spend just as much time thinking about how to efficiently compress the body of knowledge that you already know. And what sorts of things could be designed as, for example, rationality exercises for that, I'd be interested to hear ideas about that.

Q&A

Ben Pace: Thank you very much, Alkjash. I really like the idea of super-forecasters and super-compressors. I want to start a similar tournament for super-compressors.

Alkjash: If any of you are interested, I have been personally interested in questions of which languages have the highest information rate. Perhaps with the newer neural networks, we might actually be close to answering this sort of question.

Ben Pace: By languages, you don't mean computer science languages?

Alkjash: No, I mean human languages. What language seems to use more letters per idea.

Ben Pace: Interesting. Do you have a basic sense of which languages do compress things more?

Alkjash: Very naively, I expect things, for example languages with genders and other unnecessary dongles to be slightly worse, but I really don't know.

Ben Pace: All right, interesting. I think Jacob has a question.

Jacob Lagerros: So, when we spoke about this yesterday, Alkjash, you gave some interesting examples of the third bullet point. I was curious if you wanted to share some of those now as well?

Alkjash: Sure, here's one example I'm thinking about. So I have a general model about learning certain sorts of things. The model is that when you start out, there's so few data points that you just stick them into an array but as time goes on and you learn more stuff, your brain builds a more sophisticated data structure to store the same amount of data more efficiently.

Alkjash: For example, my wife and I occasionally have competed in geography sporcles, and one of the things we noticed is that I remember all the states as a visual

map (I'm a more spacial person) and where each state is. I tried to list out the states by going through vertically scanning my internal map of the United States; on the other hand, she remembers it by this alphabet song that she learned in elementary school.

Alkjash: And so these are equally useful for the application of ‘remembering the states’. But when I turn to the addition of state capitals, my model seemed to be more useful for that because it's harder to fit in the state capitals into the alphabet song than into the geographical map, especially when I roughly already know where 10 or 20 of those cities would be geographically.

Alkjash: Thinking about how these data representations... it depends on the application, what compression actually means, I think.

Ben Pace: Nice. Dennis, do you want to ask your question?

Dennis: Looks like compression is related a little to the Zettlekasten system. Have you tried it, and what do you think about it?

Alkjash: I don't know anything about that. Could you explain what that is?

Dennis: This is the system of compact little cards which is used to store information, not in categories but with links relating one to another, which was used by sociologists.

Alkjash: I see. That's very interesting. Yeah, that seems like a very natural thing, closely related. It seems like, just naively, when I learn stuff, I start out just learning a list of things and then slowly, over time, what happens is the connection build up and making this more explicit. It seems like a great thing, to skip the whole store everything as an array and stuff, altogether.

Ben Pace: Cool. Thanks, Dennis.

Ben Pace: Gosh, I'm excited by these questions. Abram if you want to go for it, you can go for it. Just for Alkjash's reference, Abram wrote that he had "not a question, but a dumb argumentive comment relating this to my talk."

Abram Demski: Yeah, so my dumb argumentive comment is, prediction does not equal compression. Sequential prediction equals compression. But non-sequential prediction is also important and does not equal compression. And so, compression doesn't capture everything about what it means to be a rational agent in the world, predicting things.

Alkjash: Interesting.

Abram Demski: And by non-sequential prediction, I mean you have a sequence of bits in the information theory model, but if instead you have this broad array of things that you could think about, and you're not sure when any one of them will become observed, [then] you want all of your beliefs to be good, but you don't have any next bit... You can't express the goodness just in terms of this sequential goodness.

Alkjash: I feel like there is some corresponding compression rate I could write down based on whatever you're being able to predict in this general non-sequential picture, but I haven't thought about it.

Abram Demski: Yeah. I think that you could get it into compression, but my claim is that compression won't capture all the notion of goodness.

Alkjash: I see. Yeah...

Ben Pace: Abram will go into more depth I expect in his talk.

Ben Pace: I should also mention that if you wanted to look at more of that Zettlekasten thing Abram has written a solid introduction to the idea on LessWrong.

Abram Demski: I am also a big fan of Zettlekasten, and recommend people check it out.

Ben Pace: All right, let me see. There was another question from Vaniver. You want to go for it?

Vaniver: Yeah, so it seems to me like the idea of trying to compress your concepts, or like your understanding of the world that you already have, is a thing that we already do a lot of.

Alkjash: Absolutely.

Vaniver: It feels to me that abstraction and categorization, this is the big thing. I guess the interesting comment to add onto that is something like, there's lumping and splitting as distinct things that are worth thinking about, where lumping is sort of saying, okay I've got a bunch of objects, how do I figure out the right categories that divide this in a small amount? And splitting is like, okay, I've got ‘all dogs’ or something and can I further subgroup within ‘dogs’ and keep going down until I have a thing that is a concrete object?

Vaniver: I think this is interesting to do because with real categories, it's much harder to find the right character encoding, or something. If we're looking at a bit-string like, we’re like oh yeah, letters are letters. And sometimes you're like, oh actually maybe we want words instead of letters.

Alkjash: That's right, yeah.

Vaniver: Your conceptual understanding of the world, it feels like there's much more variation in what you care about for different purposes.

Alkjash: Yeah, so I definitely agree that categorization is an important part of this compression. But I think, actually, probably the more important thing for compression is, as orthonormal says, having a model.

Alkjash: Here's a good example that I've thought about for a while. When I was playing Go, I spent about my first semester in college just studying Go 24/7 instead of going to class. At the end of those six months, I noticed that after playing a game, I could reproduce the entire game from memory. I don't know how much you guys know about Go, but it's a 19x19 board, and a game is placing down these black and white stones; there's probably 200 moves in the average game.

Alkjash: And this skill that I have is not at all exceptional, I think almost anyone who becomes a strong amateur will be able to reproduce their games from memory after playing for about a year. The reason they achieve such a good memory, I claim, is because they're getting good compression and it's not because they're good at

categorizing. It's because for the vast majority of moves, they can predict exactly what the next move is because they have a model of how players at their level play. Even though there's a 19x19 board, 90 percent of the time the next move is obvious.

Alkjash: So, that's why there's so little data in that actual game.

Ben Pace: Cool, thanks. I wanted to just add one thought of mine which is: I think the super-forecasters and super-compressors idea is very fun, and I think if anyone wanted an idea for a LessWrong post to write and wanted to have a format in which to Babble a bit, something that’s got different constraints, you could write one taking something like Tetlock’s 10 Heuristics that Superforecasters Use and trying to turn them into 10 Heuristics that Supercompressors Use, and seeing if you come up with anything interesting there, I would be excited to read that.

Ben Pace: All right, thanks a lot, Alkjash. I think we'll move next onto Abram.

Locality of goals

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Introduction

Studying goal-directedness produces two kinds of questions: questions about goals, and questions about being directed towards a goal. Most of my previous posts focused on the second kind; this one shifts to the first kind.

Assume some goal-directed system with a known goal. The nature of this goal will influence which issues of safety the system might have. If the goal focuses on the input, the system might wirehead itself and/or game its specification. On the other hand, if the goal lies firmly in the environment, the system might have convergent instrumental subgoals and/or destroy any unspecified value.

Locality aims at capturing this distinction.

Intuitively, the locality of the system's goal captures how far away from the system one must look to check the accomplishment of the goal.

Let's give some examples:

- The goal of "My sensor reaches the number 23" is very local, probably maximally local.
- The goal of "Maintain the temperature of the room at 23 °C" is less local, but still focused on a close neighborhood of the system.
- The goal of "No death from cancer in the whole world" is even less local.

Locality isn't about how the system extract a model of the world from its input, but about whether and how much it cares about the world beyond it.

Starting points

This intuition about locality came from the collision of two different classification of goals: the first from from Daniel Dennett and the second from Evan Hubinger.

Thermostats and Goals

In "The Intentional Stance", Dennett explains, extends and defends... the intentional stance. One point he discusses is his liberalism: he is completely comfortable with admitting ridiculously simple systems like thermostats in the club of intentional systems -- to give them meaningful mental states about beliefs, desires and goals.

Lest we readers feel insulted at the comparison, Dennett nonetheless admits that the goals of a thermostat differ from ours.

> Going along with the gag, we might agree to grant [the thermostat] the capacity for about half a dozen different beliefs and fewer desires—it can believe the room is too cold or too hot, that the boiler is on or off, and that if it wants the room warmer it should turn on the boiler, and so forth. But surely this is imputing too much to the thermostat; it has no concept of heat or of a boiler, for instance. So suppose we de-interpret its beliefs and desires: it can believe the A is too F or G, and if it wants the A to be more F it should do K, and so forth. After all, by attaching the thermostatic control mechanism to different input and output devices, it could be made to regulate the amount of water in a tank, or the speed of a train, for instance.

The goals and beliefs of a thermostat are thus not about heat and the room it is in, as our anthropomorphic bias might suggest, but about the binary state of its sensor.

Now, if the thermostat had more information about the world -- a camera, GPS position, general reasoning ability to infer information about the actual temperature from all its inputs --, then Dennett argues its beliefs and goals would be much more related to heat in the room.

> The more of this we add, the less amenable our device becomes to serving as the control structure of anything other than a room-temperature maintenance system. A more formal way of saying this is that the class of indistinguishably satisfactory models of the formal system embodied in its internal states gets smaller and smaller as we add such complexities; the more we add, the richer or more demanding or specific the semantics of the system, until eventually we reach systems for which a unique semantic interpretation is practically (but never in principle) dictated (cf. Hayes 1979). At that point we say this device (or animal or person) has beliefs about heat and about this very room, and so forth, not only because of the system's actual location in, and operations on, the world, but because we cannot imagine an-other niche in which it could be placed where it would work.

Humans, Dennett argues, are more like this enhanced thermostat, in that our beliefs and goals intertwine with the state of the world. Or put differently, when the world around us changes, it will influence almost always influence our mental states; whereas a basic thermostat might react the exact same way in vastly different environments.

> But as systems become perceptually richer and behaviorally more versatile, it becomes harder and harder to make substitutions in the actual links of the system to the world without changing the organization of the system itself. If you change its environment, it will notice, in effect, and make a change in its internal state in response. There comes to be a two-way constraint of growing specificity between the device and the environment. Fix the device in any one state and it demands a very specific environment in which to operate properly (you can no longer switch it easily from regulating temperature to regulating speed or anything else); but at the same time, if you do not fix the state it is in, but just plonk it down in a changed environment, its sensory attachments will be sensitive and discriminative enough to respond appropriately to the change, driving the system into a new state, in which it will operate effectively in the new environment.

Part of this distinction between goals comes from generalization, a property considered necessary for goal-directedness since Rohin's initial post on the subject. But the two goals also differs in their "groundedness": the thermostat's goal lies

completely in its sensors' inputs, whereas the goals of humans depend on things farther away, on the environment itself.

That is, these two goals have different locality.

Goals Across Cartesian Boundaries

The other classification of goals comes from Evan Hubinger, in a personal discussion. Assuming a Cartesian Boundary outlining the system and its inputs and outputs, goals can be functions of:

- **The environment**. This includes most human goals, since we tend to refuse wireheading. Hence the goal depends on something else than our brain state.
- **The input**. A typical goal as a function of the input is the one ascribed to the simple thermostat: maintaining the number given by its sensor above some threshold. If we look at the thermostat without assuming that its goal is a proxy for something else, then this system would happily wirehead itself, as the goal IS the input.
- **The output**. This one is a bit weirder, but captures goals about actions: for example, the goal of twitching. If there is a robot that only twitches, not even trying to keep twitching, just twitching, its goal seems about its output only.
- **The internals**. Lastly, goals can depend on what happens inside the system. For example, a very depressed person might have the goal of "Feeling good". If that is the only thing that matters, then it is a goal about their internal state, and nothing else.

Of course, many goals are functions of multiple parts of this quatuor. Yet separating them allows a characterization of a given goal through their proportions.

Going back to Dennett's example, the basic thermostat's goal is a function of its input, while human goals tend to be functions of the environment. And once again, an important aspect of the difference appears to lie in how far from the system is there information relevant to the goal -- locality.

What Is Locality Anyway?

Assuming some model of the world (possibly a causal DAG) containing the system, the locality of the goal is inversely proportional to the minimum radius of a ball, centered at the system, which suffice to evaluate the goal. Basically, one needs to look a certain distance away to check whether one’s goal is accomplished; locality is a measure of this distance. The more local a goal, the less grounded in the environment, and the most it is susceptible to wireheading or change of environment without change of internal state.

Running with this attempt at formalization, a couple of interesting point follow:

- If the model of the world includes time, then locality also captures how far in the future and in the past one must go to evaluate the goal. This is basically the short-sightedness of a goal, as exemplified by variants of twitching robots: the robot that simply twitches; the one that want to maximize its twitch in the next second; the one that want to maximize its twitching in the next 2 seconds,... up to the robot that want to maximize the time it twitches in the future.

- Despite the previous point, locality differs from the short term/long term split. An example of a short-term goal (or one-shot goal) is wanting an ice cream: after its accomplishment, the goal simply dissolves. Whereas an example of a long-term goal (or continuous goal) is to bring about and maintaing world peace -- something that is never over, but instead constrains the shape of the whole future. Short-sightedness differs from short-term, as a short-sighted goal can be long-term: "for all times t (in hours to simplify), I need to eat an ice cream in the interval [t-4,t+4]".
- Where we put the center of the ball inside the system is probably irrelevant, as the classes of locality should matter more than the exact distance.
- An alternative definition would be to allow the center of the ball to be anywhere in the world, and make locality inversely proportional to the sum of the distance of the center to the system plus the radius. This captures goals that do not depend on the state of the system, but would give similar numbers than the initial definition.

In summary, locality is a measure of the distance at which information about the world matters for a system's goal. It appears in various guises in different classification of goals, and underlies multiple safety issues. What I give is far from a formalization; it is instead a first exploration of the concept, with open directions to boot. Yet I believe that the concept can be put into more formal terms, and that such a measure of locality captures a fundamental aspect of goal-directedness.

Thanks to Victoria Krakovna, Evan Hubinger and Michele Campolo for discussions on this idea.

Life at Three Tails of the Bell Curve

If you assume other people are the same as you along every dimension then you will over-estimate other people exactly as much as you underestimate them. It is a good first-order approximation to assume other people are like yourself.

Most people are in the middle of any given bell curve. You are probably in the middle of any given bell curve too. It is a good second-order approximation to assume other people are like yourself.

But…if you assume you are statistically normal when you are not then you will have problems. I have made this mistake many, many times because my personality is extremized in three big ways.

Natural Amphetamines

> I once heard a friend, upon his first use of modafinil, wonder aloud if the way they felt on that stimulant was the way Elon Musk felt all the time. That tied a lot of things together for me, gave me an intuitive understanding of what it might "feel like from the inside" to be Elon Musk. And it gave me a good tool to discuss biological variation with. Most of us agree that people on stimulants can perform in ways it's difficult for people off stimulants to match. Most of us agree that there's nothing magical about stimulants, just changes to the levels of dopamine, histamine, norepinephrine et cetera in the brain. And most of us agree there's a lot of natural variation in these chemicals anyway. So "me on stimulants is that guy's normal" seems like a good way of cutting through some of the philosophical difficulties around this issue.
>
> — *The Parable of the Talents* by Scott Alexander

According to drugabuse.com, amphetamines have the following short-term effects:

- Quicker reaction times.
- Feelings of energy/wakefulness.
- Excitement.
- Increased attentiveness and concentration.
- Feelings of euphoria.

These characteristics describe my baseline state. I feel like I am on stimulants[1] all the time. Natural amphetamines have advantages. I am energetic. I concentrate well. I get lots of work done.

They have disadvantages too. Amphetamines are associated with headaches, appetite suppression, severe anxiety and obsessive behavior all of which *also* describe me. The headaches are ignorable because I cannot remember ever *not* having a headache. The appetite suppression is survivable because I live in a civilization full of convenient calories. The obsessive behavior is a double-edge sword. On the one hand it makes me bad at small talk. On the other hand it helps me finish things.

The anxiety in particular causes me to overprepare for disaster. When your are anxious, it is natural to search for a threat. When I feel merely moderate anxiety I

ought always to consider that there may be nothing to fear but the fear itself.

I should assume a prior expectation that other people have the following characteristics relative to myself:

- Slower reaction times
- Lethargy
- Boredom
- Inattentive and distractable
- Depressed
- Relaxed
- Normal, adjusted, balanced, sane, stable[2]

High Curiosity

Among the Big Five personality traits, curiosity (openness to experience) is my most extremized[3] one.

Other people are comparatively closed to experience. They are conventional and traditional in their outlook and behavior, with familiar routines and a narrow range of interest. I am extraordinary[4], with a wide range of interests and no set routine.

Openness to experience is considered a positive trait within liberal Western society, but it comes with disadvantages. I suffer hard for my nonconformity, think in peculiar ways and tend to get absorbed by my own fantasies.

When I meet new people, I ought to assume the following characteristics[5] as a prior:

- Conventional, traditional, culturally conservative
- Tendency not to daydream
- Rigidly conforming to routines, little need for variation
- Narrow range of interest
- Lower crystallized intelligence
- Less general knowledge
- Little need for cognitive exercise
- Dislike of intellectual activities in general and solving puzzles in particular
- Ethnocentric, authoritarian, intolerant of diversity
- Dislikes freedom, prefers externally-imposed structure
- Concerned with social dominance
- Prejudiced against all sorts of things
- Low positive affect, little joy
- Emotional blunting, reduced affective display

High Systemization

The empathizing–systemizing theory suggests there is an evolutionary tradeoff between empathizing and systematizing with autism at one end and schizophrenia at the other end, with most people in the middle. I find it a useful model for understanding a difference between myself and others. I am heavily on the systematizing end of this spectrum.

There is less research on this topic than the others so instead of listing the traits of systematizers I will list the traits of autistic people and flip them around. Compared to me, other people are:

- Tolerant of disruptions to their schedule, do not need to be notified in advance
- Tolerant of noise and other background sensory stimuli
- High empathy, almost telepathic
- Fewer, less intense special interests
- Think vaguely, imprecisely
- Good communicators

Conclusions

Comparatively-speaking, I am a heretical savant high on cocaine. That *would* explain why strangers tend to remember having met me. **I can improve my interactions with others by adopting the prior that they are passive parochial team players.**

High openness is associated with a preference for frequently-changing schedules. Systematizing is associated with schedule inflexibility. How can I exhibit both traits? I am inflexible toward others when it comes to my whimsical schedule.

By a similar paradox, I feel comfortable in the traditional oppressive culture of Japan. My systematizing proclivities enjoy the quiet perfectionism. Meanwhile, as a foreigner, I am not myself expected to conform.

My extremized characteristics all contribute to my advanced technical skills; I am pathalogically good at writing software. Ironically, the same traits simultaneously make it harder to find a job and fit into a corporation. I can only work somewhere where my technical skills are sufficiently valued for the company to tolerate my eccentricity. My ideal company would probably be working remotely for a small machine learning team. Or I could simply self-employ.

Socially, these extremized traits suggest I might connect well to other people through art, which benefits from obsessively systematic nonconformity. In particular, I have exactly the right character sheet to write a technical webcomic like xkcd.

These traits also suggest I should stay away from quantum field theory as it were meth.

---

1. Disclaimer: I have never personally taken amphetamines, cocaine, heroin, meth or anything along those lines. I apologize for my lack of empirical rigor in this domain. ↩

2. Thesaurus.com lists these words as antonyms to "obsessive". ↩

3. My other Big Five personality traits fall in the middle 98% of the bell curve. ↩

4. Thesaurus.com lists "extraordinary" as an antonym to "traditional". ↩

5. Most of these characteristics come from skimming the Wikipedia article on openness to experience. ↩

Three characteristics: impermanence

This is the sixth post of the "a non-mystical explanation of the three characteristics of existence" series.

Impermanence

Like no-self and unsatisfactoriness, impermanence seems like a label for a broad cluster of related phenomena. A one-sentence description of it, phrased in experiential terms, would be that "All experienced phenomena, whether physical or mental, inner or outer, are impermanent".

As an intellectual claim, this does not sound too surprising: few people would seriously think that either physical things or mental experiences last forever. However, there are ways in which impermanence does contradict our intuitive assumptions.

A conventional example of this is change blindness. In a typical change blindness experiment, people report having good awareness of the details of a picture shown to them: but when details are changed during an eye saccade, subjects fail to notice any difference. Maybe a person's hat looks red, and people who have been looking *right at* the hat fail to notice that it looked green just a second ago: the consciousness of the green-ness has vanished, replaced entirely with red.

People are typically surprised by this, thinking that "if it was red a second ago, surely I would remember that" - a thought that implicitly assumes that sense percepts leave permanent memories behind. But as long something does not explicitly store a piece of conscious information, it is gone as soon as it has been experienced.

This is a natural consequence of the Global Neuronal Workspace (GNW) model of consciousness from neuroscience. As I have previously discussed, studies suggest that the content of consciousness corresponds to information held in a particular network of neurons called the "global workspace". This workspace can only hold a single piece of conscious content at a time, and new information is constantly trying to enter it, replacing the old information.

Now if the content of your consciousness happens to be something like this:

- 0 milliseconds: Seeing a red hat
- 53 milliseconds: Thinking about cookies
- 200 milliseconds: Seeing a green hat

Then at the 200 millisecond mark, unless some memory system happened to explicitly store the fact of seeing a red hat before, no trace of it remains in consciousness for the person to compare with. One can train particular subsystems to monitor the contents of consciousness and send occasional summaries of previous contents, which is part of what investigating impermanence involves.

Compare this to meditation teacher Daniel Ingram's description of impermanence:

> Absolute transience is truly the actual nature of experiential reality.

> What do I mean by “experiential reality”? I mean the universe of sensations that you directly experience. […] From the conventional perspective, things are usually believed to exist even when you no longer experience them directly, and are thus inferred to exist with only circumstantial evidence to be relatively stable entities. […] For our day-to-day lives, this assumption is functional and adequate.
>
> For example, you could close your eyes, put down this book or device, and then pick it up again where you left it without opening your eyes. From a pragmatic point of view, this book was where you left it even when you were not directly experiencing it. However, when doing insight practices, it just happens to be much more useful to assume that things are only there when you experience them and not when you don’t. Thus, the gold standard for reality when doing insight practices is the sensations that make up your reality in that instant. Sensations that are not there at that time are not presumed to exist, and thus only sensations arising in that instant do exist, with “exist” clearly being a problematic term, given how transient sensations are.
>
> In short, most of what you assume as making up your universe doesn’t exist most of the time, from a purely sensate point of view. This is exactly, precisely, and specifically the point. […] sensations arise out of nothing, do their thing, and vanish utterly. Gone. Entirely gone.

In Ingram’s terms, people subconsciously assume that if a person in a picture has a red hat, then the person in the picture is going to keep having a red hat. Also, if the person in the picture has a green hat, they probably also had a green hat when you last looked at them. This kind of an assumption is often pragmatically useful, and may even be a true claim about the world, for as long as the image you are looking at is not being manipulated by researchers who keep changing subtle details. But it is not an accurate model of how your own mind functions.

Consciousness as an FBI report

In a previous article, I mentioned that according to neuroscientist Stanislas Dehaene, one of the functions of consciousness is for subsystems in the brain to exchange summaries of their conclusions. He offered the analogy of the US president being briefed by the FBI. The FBI is a vast organization, with thousands of employees: they are constantly shifting through enormous amounts of data and forming hypotheses about topics with national security relevance. But it would be useless for the FBI to present to the president every single report collected by every single field agent, as well as every analysis compiled by every single analyst in response. Rather, the FBI needs to internally settle on some overall summary of what they believe is going on, and then present that to the President, who can then act based on the information. Similarly, Dehaene suggests that consciousness is a place where different brain systems can exchange summaries of their models, and to integrate conflicting evidence in order to arrive to an overall conclusion.

In a similar way, it’s usually not necessary for the brain to keep a conscious track of every little detail in an image. Rather, sensory information comes in, and subsystems responsible for processing it broadcast a summary of what they consider important about it. If you look at a painting, a general summary of its contents will be produced and maintained in consciousness, while minor details like the color of someone’s hat won’t be recorded unless a person had a particularly important reason to look at it. (That would be the equivalent of the FBI including a field agent’s random observations

in a report to the president. They're very unlikely to include those unless they are *really* important.)

This is particularly noticeable when learning to draw: as Raemon discusses in Drawing Less Wrong: Observing Reality:

> When you look at a person, what you perceive is not a series of shapes and colors that correspond to what's there, but rather a bunch of hastily constructed symbols that convey the information that the brain thinks is important. If you haven't rewired your brain for drawing, then "important" questions do not include "*Is that elbow angled at 90 degrees or 75?*" or "*Where are the eyes in relation to the top of the head?*" Instead, what you usually care about are things like "is this person happy, or angry?" and the information that gets recorded is a little tag that says "Smiling" with a vague curving-upwards-line symbol accompanying it.
>
> A large chunk of the information we usually need has to do with the face. This plays a role in two common biases that are near-universal in inexperienced artists:
>
> -Drawing the head much larger than it actually is, compared to the rest of the body
>
> -Drawing the "face" (i.e. everything between the eyebrows and mouth) as if they took up the entire head rather than the bottom half. Practically everything above the eyebrows conveys no relevant information, so it's just ignored.
>
> Your brain has a mental model of what a human is "supposed" to look like, and that model is wrong. You can see major gains in drawing capability just by learning the "ideal" proportions of a human being.

The relation of this to impermanence is that observing the contents of your mind lets you notice just how little sense data is actually used, and how quickly it vanishes. Returning to Daniel Ingram:

> We are typically quite sloppy about distinguishing between physical and mental sensations (memories, mental images, and mental impressions of other physical or mental sensations). These two kinds of sensations alternate, one arising and passing and then the other arising and passing, in a quick but perceptible fashion. Being clear about exactly when the physical sensations are present will begin to clarify their slippery counterparts—flickering mental impressions—that help co-create the illusion of continuity, stability, or solidity. [...]
>
> Each one of these sensations (the physical sensation and the mental impression) arises and vanishes completely before another begins, so it is possible to sort out which is which with relatively stable attention dedicated to consistent precision and to not being lost in stories. This means that the instant you have experienced something, you can know that it isn't there anymore, and whatever is there is a new sensation that will be gone in an instant. There are typically many other momentary sensations and impressions interspersed with these, but for the sake of practice, this is close enough to what is happening to be a good working model.

Ingram suggests that between physical sensations, there are mental sensations which "fill in the gaps", and which prevent people from noticing that the original physical sensations only come in sporadically. As people become more adept at meditation practices such as following the breath, they may come to notice that a large part of their time has been spent on following *a thought about the breath*, rather than *the*

breath itself: and far less sensory information about the breath actually comes to consciousness than they assumed.

In an earlier article on insight meditation I gave another example about these kinds of mental sensations. I mentioned a time when I was doing concentration meditation, using an app that played the sound of something hitting a woodblock, 50 times per minute. As I was concentrating on listening to the sound, I noticed that what had originally been just one thing in my experience - a discrete sound event - was actually composed of many smaller parts. The beginning and end of the sound were different, so there were actually two sound sensations; and there was a subtle visualization of something hitting something else; and a sense of motion accompanying that visualization. I had not previously even been fully aware that my mind was automatically creating a mental image of what it thought that the sound represented.

Continuing to observe those different components, I became more aware of the fact that my visualization of the sound changed over time and between meditation sessions, in a rather arbitrary way. Sometimes my mind conjured up a vision of a hammer hitting a rock in a dwarven mine; sometimes it was two wooden sticks hitting each other; sometimes it was drops of water falling on the screen of my phone.

Normally, all of this would just be packaged together into a general impression of "I'm hearing some sound". Our raw sense data is made up of countless small details and sensations, each arising and passing away in rapid succession - but we mostly perceive the high-level summaries, which are much more static. This creates an experience of seeing solid and discrete objects, and a feeling of there being permanent objects.

So how *does* one actually come to see what is happening in their mind?

In *The Mind Illuminated*, meditation teacher and former neuroscientist John Yates (Culadasa) suggests that one way this happens is by taking the subsystems responsible for producing such summaries and directing them to produce summaries about the content of consciousness. The brain already has a subsystem that generates overall summaries of what's going on in your mind; you can train that system to produce more detailed reports. Yates calls such summaries *introspective awareness* (discussed in more detail in an earlier article).

The impermanence of the self

As I have discussed, consciousness involves a constant competitive process, where different subsystems send content to the global workspace. At any given time, only one of these pieces of content is selected to become the content of consciousness. We might say that there has been a "subsystem switch" or a "subsystem swap" when the content of consciousness changes from that submitted by one subsystem to that submitted by another.

In normal circumstances, the structure of your mind is such that you cannot directly notice the different subsystems getting swapped in and out. Your consciousness can only hold one piece of information at a time. Suppose that at one moment, you are thinking of your friend, and at the next you are thinking of candy. When you think of candy, you are no longer aware of the fact that you were thinking of your friend the previous moment. You can often *infer* that a subsystem switch has happened, but you can't actually *experience* the switch.

However, if you develop more detailed introspective awareness, the stream of your consciousness may include reports such as this:

- Subsystem 1: So I was talking with my friend and she said…
- Subsystem 2: Ooh, candy.
- Awareness subsystem: The train of thought about my friend switched to a train of thought about candy right now.

Subjectively, this feels like becoming aware of the subsystem swapping in real time: a thought comes in, while an "afterimage" of the previous thought lingers for a brief moment, enough to make you realize that one kind of thought has replaced the other. If the trains of thought are different enough, the transitions between might feel really sharp and distinct.

You may also notice that you have two or more separate thought streams going in parallel, while having had no awareness of the fact. At one time you are thinking about your friend, and at the other time you are thinking about candy. Despite the fact that these two thought streams have kept alternating, maybe switching once every couple of seconds, they have been entirely unaware of each other. First the candy is everything that is in your mind, then your friend, then the candy again.

This is not to say that it would normally be impossible to be aware of having multiple trains of thought going on. Even without meditative training, your brain is constantly producing summaries of what's happening, including summaries of what's happening in your head. But what normally happens is something like having the first train of thought, then having the second train of thought, and then having general introspective awareness of there being two trains of thought. What does not usually happen is that the introspective awareness is sharp enough to register the fact that *whenever the train of thought switches, everything else disappears from consciousness for the duration*.

Rather than there being a single observer who experiences all of their own thoughts, there are three separate processes, two of them concerned with their own issues and a third meta-process keeping a loose record of what the two others have been up to.

A rough analogy would be to a (single-core) computer that keeps executing multiple different programs in succession, with the contents of the processor being cleaned out for the next program each time the execution switches. As long as everything goes smoothly, things will appear to the user as multiple different programs being executed at the same time, and the programs themselves will be unaware of the other programs. Yet, a sufficiently fine-grained trace of the different processes will reveal that only one has been running at a time. (Though unlike in this analogy, mental subsystems do keep running even when "swapped out"; they just don't have write access to consciousness during that time.)

By developing sufficient detail, another thing that can be noticed is that the sense of self is actually only present a part of the time. As discussed in previous posts [1, 2], the experience of a self is basically a piece of data - a *narrative* which is sometimes experienced and sometimes not. That is, it is another high-level summary of what is happening - "I am doing this thing" - constructed from lower-level data. (In a comment, Vanessa Kosoy suggested that the experience of a self is an explanation of *why* the person is doing things, constructed for social purposes and to be able to justify your behavior afterwards. This sounds plausible to me.)

That means that the content of your consciousness may be something like:

- Time 1: The sight of a bird outside the window.
- Time 2: The thought “there’s a bird over there”.
- Time 3: The experience of typing on a keyboard.
- Time 4: The sound of a car outside.
- Time 5: A mental image of a car.
- Time 6: A sense of being someone who sees the bird and hears the car, while typing on a keyboard.

… that is, normally you may experience there being a constant, permanent self which feels like *what you really are*. But in fact, during a large part of your conscious experience, that sense of self may simply not be there at all. Normally this might be impossible to detect due to what’s called the refrigerator light illusion: the light in a refrigerator turns on whenever you open the door, so it seems to you to always be on. Likewise, whenever you ask “do I experience a sense of self right now”, that question references and activates a self-schema, meaning that the answer is always “yes”. It is only by developing introspective awareness that records *all* mental content, without needing to make reference to a self, that you can come to notice the way in which your self constantly appears and disappears.

It is worth noting that coming to experience this may feel very frightening. Psychologist and meditation teacher Ron Crouch describes one way that it can go:

> What is actually happening, down deep, is that as your attention is syncing up with the dissolution of phenomena you are finding that there is nothing in experience that the sense of “me” can hold onto as stable and permanent. It just can’t get any footing. You do not realize it at a cognitive level, but you are getting a deep insight into the impermanence of all phenomena, and along with that, into the impermanence of the self. This is something that is terrifying to one’s very roots. Needles to say this initial stage can be a great source of distress and people can become stuck here for some time if they do not have good guidance.

We might think the distress follows from the mind’s underlying assumption that the self must be something like a permanent object. Whenever one has checked for the presence of the self, it has been there: thus, it is something that persists uninterrupted over time (except maybe in sleep). Now it - or something that resembles what it used to be - suddenly keeps vanishing and reappearing. Does that mean that you are dying?

Eventually, given enough further practice, the mind readjusts and revises its models. Continuity of consciousness does not mean uninterrupted continuity of self after all; the self is as impermanent as any other sensory experience. Nothing here to see, move along now.

Impermanence and unsatisfactoriness

One aspect of craving is *clinging,* a kind of repeated craving. The mind notices a pleasant or unpleasant sensation, and then tries to keep the pleasant sensations in consciousness and the unpleasant sensations out of consciousness. This may feel like you are trying to “freeze” the content of consciousness into a particular, pleasurable slice of experience.

In an earlier post, I gave a list of examples about craving; this is also a good list of examples to use for clinging, so I’ll repeat it here:

- It is morning and your alarm bell rings. You should get up, but it feels nice to be sleepy and remain in bed. You want to hang onto those pleasant sensations of sleepiness for a little bit more.
- You are spending an evening together with a loved one. This is the last occasion that you will see each other in a long time. You feel really good being with them, but a small part of you is unhappy over the fact that this evening will eventually end.
- You are at work on a Friday afternoon. Your mind wanders to the thought of no longer being at work, and doing the things that you had planned to do on the weekend. You would prefer to be done with work already, and find it hard to stay focused as you cling to the thoughts of your free time.
- You are single and hanging out with an attractive person. You know that they are not into you, but it would be so great if they were. You can't stop thinking about that possibility, and this keeps distracting you from the actual conversation.
- You are in a conversation with several other people. You think of a line that would be a really good response to what someone else just said. Before you can say it, somebody says a thing, and the conversation moves on. You find yourself still thinking of your line, and how nice it would have been to get to say it.
- You are playing a game of chess. You see an opportunity to make a series of movies that looks like it would win the game for you. You get so focused on the sequence of moves that would bring you a victory, that you don't notice that your opponent could also respond in a way that would ruin the entire plan.
- You had been planning on going to a famous museum while on your vacation, but the museum turns out to be temporarily closed at the time. You keep thinking about how much you had been looking forward to it.

What is essentially going on, is the craving trying to *fight against impermanence*. Taking the example of being sleepy and in bed: there is the sensation of sleepiness and a feeling of pleasure; *and* that annoying thought which keeps saying that you really need to get up soon... and the craving wants that pleasant sleepiness *back* and *stable*, damnit. If only it would focus on the sleepiness enough, maybe that annoying reminder would go away...

This contributes to the loop where the mind sees craving as necessary for well-being: phenomena won't stabilize in consciousness by themselves, and craving takes actions to make them more stable. Whenever it is unsuccessfully trying to do so, there is discomfort; when it succeeds in getting the pleasant thing to become the object of consciousness (if only for a moment), there is less discomfort (if only for a moment). Now, that discomfort is being generated *by the craving itself*, so it could also be eliminated by dropping the craving… but the system does not notice that.

Nor does it notice that following the craving does not lead to consistent happiness. Of course, we may *intellectually* understand that there's no single thing that would make us permanently and eternally happy. But at the subsystem level, each source of craving is based on a schema that states something like:

- If I get the thing I am craving, things will feel satisfying.

When the subsystem related to that goal is active, this is the schema which will be active in the person's mind. If you are hungry for food, you only think about how food will bring relief to your discomfort. Intellectually, you may know that soon afterwards you will start wanting something else - but the assumption that your mind is operating from, is that getting the food will bring contentment. And that assumption is correct! Recall that unsatisfactoriness is actually caused *by* craving. So getting the food *will*

make the craving for it go away - until the next craving pops up, which is likely to happen very soon.

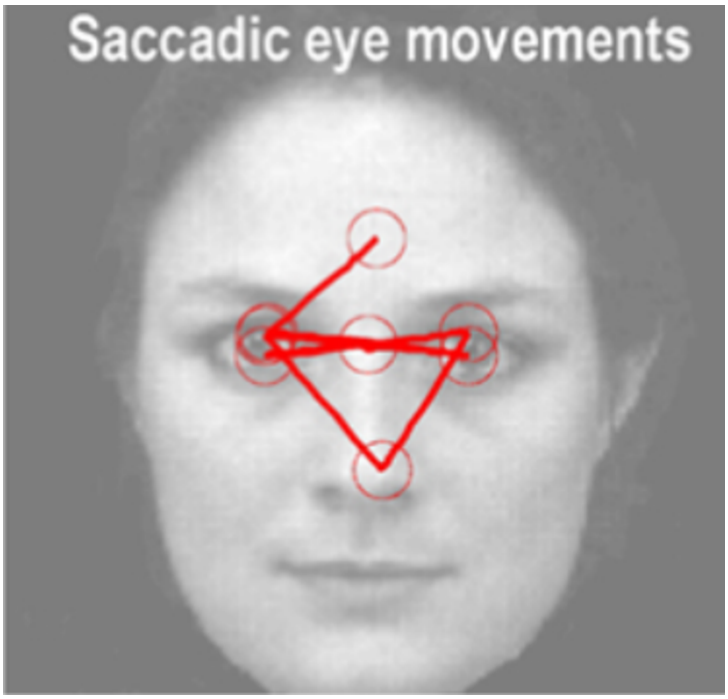
This has the consequence that each *individual* craving may have its prediction confirmed. The craving for food correctly predicts that food will bring satisfaction from that craving. The craving to look at the phone while you eat correctly predicts that looking at the phone will bring satisfaction from that craving. The craving to go watch pictures of attractive naked people after eating correctly predicts that going to watch pictures of attractive naked people will bring satisfaction from that craving… while the overall system remains in a near-constant state of craving that just keeps changing its target.

In the paper *Suffering* (Metzinger, 2016), Thomas Metzinger reports on an “experience sampling” experiment, where messages were sent to people’s phones at random times, asking them whether they felt that their current experience would feel worth reliving:

> For many, the result was surprising: the number of positive conscious moments per week varied between 0 and 36 [out of 70], with an average of 11.8 or almost 31 per cent of the phenomenological samples, while at 69 per cent a little more than two thirds of the moments were spontaneously ranked as not worth reliving.

Metzinger notes that one cannot generalize from these results to the general population: this was a small, unreplicated pilot study done with a highly selected group (philosophy students). But as he also notes, what *is* remarkable is that *nearly all* of the participants were surprised by their own results - they had expected many more moments to feel pleasurable. He speculates that human motivation may depend on systematic self-deception: if a person valued positive experiences but noticed that most of their experience was actually unpleasant, they might become paralyzed.

And it does seem that increased awareness of the impermanence of satisfaction helps reduce craving. I like to think of each individual craving as a form of a *hypothesis*, in the predictive processing sense where hypotheses drive behavior by seeking to prove themselves true. For example (Friston et al. 2012), your visual system may see someone's nose and form the hypothesis that "the thing that I'm seeing is a nose, and a nose is part of a person's face, so I'm seeing someone's face". That contains the prediction "faces have eyes next to the nose, so if I look slightly up and to the right I will see an eye, and if I look left from there I will see another eye"; it will then seek to confirm its prediction by making you look at those spots and verify that they do indeed contain eyes.



Eye movements seeking to confirm the hypotheses of "I am seeing a face". From [Friston, Adams, Perrinet & Breakspear 2012](#).

Normally, each craving is successfully proving true the hypothesis of "pursuing this craving will cause satisfaction"... but included in that prediction is not only a claim that satisfying this craving will bring momentary satisfaction. As the hypothesis is not modeling events that happen after it is satisfied, there is an implied claim that this will bring *lasting* satisfaction.

If the mind-system develops increased awareness of the way repeated craving seems to just lead to a constant state of discomfort, then under the right conditions it may consider the hypotheses in those cravings falsified and discard them.

Fighting against a sensation as assuming its permanence

Another thing that is happening is the subsystems failing to notice how fighting against a sensation actually helps *keep* it in consciousness, and how the sensation might actually fall away *on its own* if it was not being fought against.

Suppose that you are feeling stressed out over something, and a craving is activated to get rid of the feeling of stress. This involves sending into consciousness a plan for getting rid of the sensation of stress, which needs to *make reference to the sensation of stress*. This tends to redirect *more* attention towards the sensation of stress, strengthening the signal associated with it... and because sensations are normally impermanent and tend to easily vanish, this may help keep it in consciousness whereas it would otherwise have disappeared on its own.

In general, craving often operates under the assumption that unpleasant sensations are permanent: that is, they will persist in consciousness until actively resisted. And certainly it is true that not *all* unpleasant sensations will just disappear if you stop feeding them with attention. But even then, redirecting attention into a struggle against them may actively make them stronger.

If one develops sufficient introspective awareness, they may come to experience this directly. They will notice a neutral sensation, a negative sensation, another neutral sensation, then the aversion to the negative sensation… and notice there are actually quite a few neutral sensations, during which the negative sensation does not bother them at all. This helps notice that struggling against discomfort is actually not *necessary* for being free of discomfort; one is free of discomfort a large part of the time already.

Impermanence as vibration

No discussion of impermanence would be complete without touching upon the topic of "vibrations". Recall that according to the predictive processing model, the brain is composed of layers prediction machinery. A given layer can receive sensory information from a lower layer, and from the higher layer predictions of what that information *should* look like. For purposes of prediction, each layer is trying to form *models* of what it expects to see.

Rather than sense information primarily "flowing up" from the sense organs, the brain keeps making guesses of what it *expects* to see. These expectations are sent "down to the senses", with the brain using the sense data to *check* its assumptions and correcting for any mismatches. Mismatches that seem small enough may be ignored and explained away as noise.

One possible model is that the sensory information from the lower levels represents stable, permanent objects. As has been noted, this assumption is often a useful and correct one for predicting how the world behaves, so the system begins to assume it… ignoring the fact that sensory data is actually coming *in pulses* rather than constantly.

When one's consciousness starts dropping some of the mental impressions that normally "fill in the gaps", it may lead to an experiential quality of reality "vibrating". Here is how DavidM describes this:

> A meditator practicing in this style will eventually find that their experience is not static, but 'vibrates' or fluxes in a peculiar way over extremely short periods of time (fractions of a second). For an explanation by analogy, imagine a set of speakers playing music without dynamic variation; if a person rapidly turns the volume knob in the pattern off-low-high-low-off, the amplitude of the music will flux over time. Similarly, a meditator practicing in this style finds that the components of experience are not static, but fluctuate rapidly from nonexistent to existent and back again. N.B. This has nothing to do with the fact that the contents of experience are constantly changing. Rather, apparently static objects (e.g. an unchanging white visual field) turn out to be in flux.

For the most part, the hypothesis of "sensory data represents permanent objects" has turned out to deliver good results, so normally any gaps in the data will be automatically "filled in" by the model, as they are assumed to be meaningless noise. As a result, a "neural autocompletion feature" can create an impression of closely

observing sensory data, even when sensory data is actually sparse and the impression of it is mostly fabricated on the basis of a few data points.

For as long as the sensed data deviates only a little from the expected, the deviation is treated as noise and ignored; but once the deviation crosses some critical threshold, it is picked up and registered as surprising. If one intentionally goes looking for vibrations, then one is trying to pick up finer and finer distinctions in the sense data. This forces the system to pay attention to minor patterns that would otherwise have been treated as meaningless noise. That causes it to notice discrepancies between the higher-level model's prediction of "solid stream of sense data" and the sensory experiences that are coming in as pulses. This leads to an awareness of vibrations, and more generally insight into how the brain fills in data which is not actually there.

On the other hand, I have also heard reports of people finding vibrations without explicitly even looking at sensory details, in contexts such as doing loving-kindness meditation. I am confused about what is going on there and don't know how to explain it. This is also an area that I have personally investigated relatively little.

This is the sixth post of the "a non-mystical explanation of the three characteristics of existence" series.

Institutional Senescence

Consider this toy model:

An institution, such as a firm, an association or a state, is formed.

It works well in the beginning. It encounters different problems and solves them the best it can.

At some point though a small problem arises that happens to be a suboptimal Nash equilibrium: None of the stakeholders can do better by trying to solve it on their own. Such problems are, almost by definition, unsolvable.

Thus the problem persists. It's an annoyance, but it's not a big deal. The institution is still working well and you definitely don't want to get rid of it just because it's not perfect.

As the time goes on, such problems accumulate. They also tend to have unpleasant consequences: If such a problem makes particular medical treatment unavailable, it incentivizes the patients to bribe the doctors and the doctors to break the law and administer the treatment anyway. Now, in addition to malfunctioning medical system, you have a problem with corruption.

After on time the institution accumulates so many suboptimal Nash equilibria that it barely works at all.

The traditional solution to this problem is internal strife, civil war or revolution. It eventually destroys the institution and, if everything goes well, replaces it with a different one where at least the most blatant problems are fixed.

War or revoulution is not a desirable outcome though: In addition to the human suffering, it also tends to replace the people in power. But the people in power don't like to be replaced and so they will try to prevent it.

One manoevre they can use is to introduce planned institutional death: Every now and then the institution would be dismantled and created anew, without having to resort to a war or revolution.

Here's an example: The credit system tends to be one big suboptimal Nash equilibrium in itself. Compound interest grows the size of the debt like crazy and unless there's a way to limit the harm it'll destroy people and business and eventually the entire economy. Even lenders would be hurt, but none of them has a reason to mitigate the problem. They could, in theory, forgive the debt for the sake of keeping the economy afloat, but that would put them in disadvantage to other lenders.

And so the king or the religious authority decides to have jubilee years. Every fifty years, all debts are forgiven. The institution of money lending dies and is rises anew from the ashes. (David Graeber asserts that the practice was, in fact, not specific to Israel, but common at the time among the ancient societies in the Middle East.)

One can also think of the democratic system of regular elections as a kind of planned institutional death. Every four years, the government, with all the accumulated dysfunction, is thrown out and a new one is instituted. But the government example

also makes the problem with planned death obvious. Government is replaced, but the people on non-political positions, various administrators and small-scale decision makers, remain. At least some inadequate Nash equilibria can therefore survive the change of the government. And those would accumulate over the time and eventually lead to the system collapse. We are between a rock and a hard place here: We want to destroy the institution to break the equilibria, but at the same time we want to preserve the institutional knowledge. We don't want to get all the way back to the trees after all. We don't want to get back to the middle ages either.

Last example that comes to mind is IETF, the institution that standardizes how Internet works. The real work, the development of standards, is done in working groups, which have a clear charter that defines what they are supposed to achieve and more importantly, how long would it take. The working group exists for, say, four months, and then dies. Sure, there are IETF institutions other than the working groups and those can survive for longer. But these are mostly doing the support jobs. Organizing meetings, publishing the new standards and so on. The real stuff happens in the working groups.

All in all, I am not at all sure that planned institutional death is a solution to all suboptimal equilibria problems, but the fact that evolution uses it, that it fights dysfunctions, such as cancer, by discarding the bulk of the cells every now and then and preserving only the germline, makes it at least worth of consideration.

[AN #103]: ARCHES: an agenda for existential safety, and combining natural language with deep RL

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Alignment Newsletter is a weekly publication with recent content relevant to AI alignment around the world. Find all Alignment Newsletter **resources here**. In particular, you can look through **this spreadsheet** of all summaries that have ever been in the newsletter.

Audio version **here** (may not be up yet).

HIGHLIGHTS

AI Research Considerations for Human Existential Safety *(Andrew Critch et al)* (summarized by Rohin): This research agenda out of CHAI directly attacks the problem longtermists care about: **how to prevent AI-related existential catastrophe**. This is distinctly different from the notion of being "provably beneficial": a key challenge for provable beneficence is defining what we even mean by "beneficial". In contrast, there are avenues for preventing AI-caused human extinction that do not require an understanding of "beneficial": most trivially, we could coordinate to never build AI systems that could cause human extinction.

Since the focus is on the *impact* of the AI system, the authors need a new phrase for this kind of AI system. They define a **prepotent AI system** to be one that cannot be controlled by humanity **and** has the potential to transform the world in a way that is at least as impactful as humanity as a whole. Such an AI system need not be superintelligent, or even an AGI; it may have powerful capabilities in a narrow domain such as technological autonomy, replication speed, or social acumen that enable prepotence.

By definition, a prepotent AI system is capable of transforming the world drastically. However, there are a lot of conditions that are necessary for continued human existence, and most transformations of the world will not preserve these conditions. (For example, consider the temperature of the Earth or the composition of the atmosphere.) As a result, human extinction is the *default* outcome from deploying a prepotent AI system, and can only be prevented if the system is designed to preserve human existence with very high precision relative to the significance of its actions. They define a misaligned prepotent AI system (MPAI) as one whose deployment leads to human extinction, and so the main objective is to avert the deployment of MPAI.

The authors break down the risk of deployment of MPAI into five subcategories, depending on the beliefs, actions and goals of the developers. The AI developers could fail to predict prepotence, fail to predict misalignment, fail to coordinate with other teams on deployment of systems that aggregate to form an MPAI, accidentally (unilaterally) deploy MPAI, or intentionally (unilaterally) deploy MPAI. There are also hazardous social conditions that could increase the likelihood of risks, such as unsafe

development races, economic displacement of humans, human enfeeblement, and avoidance of talking about x-risk at all.

Moving from risks to solutions, the authors categorize their research directions along three axes based on the setting they are considering. First, is there one or multiple humans; second, is there one or multiple AI systems; and third, is it helping the human(s) comprehend, instruct, or control the AI system(s). So, multi/single instruction would involve multiple humans instructing a single AI system. While we will eventually need multi/multi, the preceding cases are easier problems from which we could gain insights that help solve the general multi/multi case. Similarly, comprehension can help with instruction, and both can help with control.

The authors then go on to list 29 different research directions, which I'm not going to summarize here.

Rohin's opinion: I love the abstract and introduction, because of their directness at actually stating what we want and care about. I am also a big fan of the distinction between provably beneficial and reducing x-risk, and the single/multi analysis.

The human fragility argument, as applied to generally intelligent agents, is a bit tricky. One interpretation is that the "hardness" stems from the fact that you need a bunch of "bits" of knowledge / control in order to keep humans around. However, it seems like a generally intelligent AI should easily be able to keep humans around "if it wants", and so the bits already exist in the AI. (As an analogy: we make big changes to the environment, but we could easily preserve deer habitats if we wanted to.) Thus, it is really a question of what "distribution" you expect the AI system is sampled from: if you think we'll build AI systems that try to do what humanity wants, then we're probably fine, but if you think that there will be multiple AI systems that each do what their users want, but the users have conflicts, the overall system seems more "random" in its goals, and so more likely to fall into the "default" outcome of human extinction.

The research directions are very detailed, and while there are some suggestions that don't seem particularly useful to me, overall I am happy with the list. (And as the paper itself notes, what is and isn't useful depends on your models of AI development.)

Human Instruction-Following with Deep Reinforcement Learning via Transfer-Learning from Text *(Felix Hill et al)* (summarized by Nicholas): This paper proposes the Simulation-to-Human Instruction Following via Transfer from Text (SHIFTT) method for training an RL agent to receive commands from humans in natural language. One approach to this problem is to train an RL agent to respond to commands based on a template; however, this is not robust to small changes in how humans phrase the commands. In SHIFTT, you instead begin with a pretrained language model such as BERT and first feed the templated commands through the language model. This is then combined with vision inputs to produce a policy. The human commands are later fed through the same language model, and they find that the model has zero-shot transfer to the human commands even if they differ in structure.

Nicholas's opinion: Natural language is a very flexible and intuitive way to convey instructions to AI. In some ways, this shifts the alignment problem from the RL agent to the supervised language model, which just needs to learn how to correctly interpret the meaning behind human speech. One advantage of this approach is that the

language model is separately trained so it can be tested and verified for safety criteria before being used to train an RL agent. It also may be more competitive than alternatives such as reward modeling that require training a new reward model for each task.

I do see a couple downsides to this approach, however. The first is that humans are not perfect at conveying their values in natural language (e.g. King Midas wishing for everything he touches to turn to gold), and natural language may not have enough information to convey complex preferences. Even if humans give precise and correct commands, the language model needs to verifiably interpret those commands correctly. This could be difficult as current language models are difficult to interpret and contain many harmful biases.

Grounding Language in Play *(Corey Lynch et al)* (summarized by Robert): This paper presents a new approach to learning to follow natural language human instruction in a robotics setting. It builds on similar ideas to **Learning Latent Plans from Play** (**AN #65**), in that it uses unsupervised "play" data (trajectories of humans playing on the robot with no goal in mind).

The paper combines several ideas to enable training a policy which can follow natural language instructions with only limited human annotations.

* In *Hindsight Instruction Pairing*, human annotators watch small trajectories from the play data, and label them with the instruction which is being completed in the clip. This instruction can take any form, and means we don't need to choose the instructions and ask humans to perform specific tasks.

* *Multicontext Imitation Learning* is a method designed to allow goal-conditioned policies to be learned with multiple different types of goals. For example, we can have lots of example trajectories where the goal is an end state image (as these can be generated automatically without humans), and just a small amount of example trajectories where the goal is a natural language instruction (gathered using *Hindsight Instruction Pairing*). The approach is to learn a goal embedding network for each type of goal specification, and a single shared policy which takes the goal embedding as input.

Combining these two methods enables them to train a policy and embedding networks end to end using imitation learning from a large dataset of (trajectory, image goal) pairs and a small dataset of (trajectory, natural language goal) pairs. The policy can follow very long sequences of natural language instructions in a fairly complex grasping environment with a variety of buttons and objects. Their method performs better than the Learning from Play (LfP) method, even though LfP uses a goal image as the goal conditioning, instead of a natural language instruction.

Further, they propose that instead of learning the goal embedding for the natural language instructions, they use a pretrained large language model to produce the embeddings. This improves the performance of their method over learning the embedding from scratch, which the authors claim is the first example of the knowledge in large language models being transferred and improving performance in a robotics domain. This model also performs well when they create purposefully out of distribution natural language instructions (i.e. with weird synonyms, or google-translated from a different language).

Robert's opinion: I think this paper shows two important things:

1. Embedding the natural language instructions in the same space as the image conditioning works well, and is a good way of extending the usefulness of human annotations.

2. Large pretrained language models can be used to improve the performance of language-conditioned reinforcement learning (in this case imitation learning) algorithms and policies.

Methods which enable us to scale human feedback to complex settings are useful, and this method seems like it could scale well, especially with the use of pretrained large language models which might reduce the amount of language annotations needed further.

TECHNICAL AI ALIGNMENT

MISCELLANEOUS (ALIGNMENT)

From ImageNet to Image Classification *(Dimitris Tsipras et al)* (summarized by Flo): ImageNet was crowdsourced by presenting images to MTurk workers who had to select images that contain a given class from a pool of images obtained via search on the internet. This is problematic, as an image containing multiple classes will basically get assigned to a random suitable class which can lead to deviations between ImageNet performance and actual capability to recognize images. The authors used MTurk and allowed workers to select multiple classes, as well as one main class for a given image in a pool of 10000 ImageNet validation images. Around 20% of the images seem to contain objects representing multiple classes and the average accuracy for these images was around 10% worse than average for a wide variety of image classifiers. While this is a significant drop, it is still way better than predicting a random class that is in the image. Also, advanced models were still able to predict the ImageNet label in cases where it does not coincide with the main class identified by humans, which suggest that they exploit biases in the dataset generation. While the accuracy of model predictions with respect to the newly identified main class still increased with better accuracy in predicting labels, the accuracy gap seems to grow and we might soon hit a point where gains in ImageNet accuracy don't correspond to improved image classification.

Read more: Paper: From ImageNet to Image Classification: Contextualizing Progress on Benchmarks

Flo's opinion: I generally find these empiricial tests of whether ML systems actually do what they are assumed to do quite useful for better calibrating intuitions about the speed of AI progress, and to make failure modes more salient. While we have the latter, I am confused about what this means for AI progress: on one hand, this supports the claim that improved benchmark progress does not necessarily translate to better real world applicability. On the other hand, it seems like image classification might be easier than exploiting the dataset biases present in ImageNet, which would mean that we would likely be able to reach even better accuracy than on ImageNet for image classification with the right dataset.

Focus: you are allowed to be bad at accomplishing your goals *(Adam Shimi)* (summarized by Rohin): **Goal-directedness** (**AN #35**) is one of the key drivers of AI risk: it's the underlying factor that leads to **convergent instrumental subgoals**.

However, it has eluded a good definition so far: we cannot simply say that it is the optimal policy for some simple reward function, as that would imply AlphaGo is not goal-directed (since it was beaten by AlphaZero), which seems wrong. Basically, goal-directedness should not be tied directly to *competence*. So, instead of only considering optimal policies, we can consider any policy that could have been output by an RL algorithm, perhaps with limited resources. Formally, we can construct a set of policies for G that can result from running e.g. SARSA with varying amounts of resources with G as the reward, and define the focus of a system towards G to be the distance of the system's policy to the constructed set of policies.

Rohin's opinion: I certainly agree that we should not require full competence in order to call a system goal-directed. I am less convinced of the particular construction here: current RL policies are typically terrible at generalization, and tabular SARSA explicitly doesn't even *try* to generalize, whereas I see generalization as a key feature of goal-directedness.

You could imagine the RL policies get more resources and so are able to understand the whole environment without generalization, e.g. if they get to update on every state at least once. However, in this case realistic goal-directed policies would be penalized for "not knowing what they should have known". For example, suppose I want to eat sweet things, and I come across a new fruit I've never seen before. So I try the fruit, and it turns out it is very bitter. This would count as "not being goal-directed", since the RL policies for "eat sweet things" would already know that the fruit is bitter and so wouldn't eat it.

OTHER PROGRESS IN AI

DEEP LEARNING

Identifying Statistical Bias in Dataset Replication *(Logan Engstrom et al)* (summarized by Flo): One way of dealing with finite and fixed test sets and the resulting possibility of overfitting on the test set is dataset replication, where one tries to closely mimic the original process of dataset creation to obtain a larger test set. This can lead to bias if the difficulty of the new test images is distributed differently than in the original test set. A previous attempt at **dataset replication on ImageNet** tried to get around this by measuring how often humans under time pressure correctly answered a yes/no question about an image's class (dubbed selection frequency), which can be seen as a proxy for classification difficulty.

This data was then used to sample candidate images for every class which match the distribution of difficulty in the original test set. Still, all tested models performed worse on the replicated test set than on the original. Parts of this bias can be explained by noisy measurements combined with disparities in the initial distribution of difficulty, which are likely as the original ImageNet data was prefiltered for quality. Basically, the more noisy our estimates for the difficulty are, the more the original distribution of difficulty matters. As an extreme example, imagine a class for which all images in the original test set have a selection frequency of 100%, but 90% of candidates in the new test set have a selection frequency of 50%, while only 10% are as easy to classify as the images in the original test set. Then, if we only use a single human annotator, half of the difficult images in the candidate pool are indistinguishable from the easy ones, such that most images ending up in the new test set are more difficult to classify than the original ones, even after the adjustment.

The authors then replicate the ImageNet dataset replication with varying amounts of annotators and find that the gap in accuracy between the original and the new test set progressively shrinks with reduced noise from 11.7% with one annotator to 5.7% with 40. Lastly, they discuss more sophisticated estimators for accuracy to further lower bias, which additionally decreases the accuracy gap down to around 3.5%.

Flo's opinion: This was a pretty interesting read and provides evidence against large effects of overfitting on the test set. On the other hand, results like this also seem to highlight how benchmarks are mostly useful for model comparison, and how nonrobust they can be to fairly benign distributional shift.

Cold Case: The Lost MNIST Digits *(Chhavi Yadav et al)* (summarized by Flo): As the MNIST test set only contains 10,000 samples, concerns that further improvements are essentially overfitting on the test set have been voiced. Interestingly, MNIST was originally meant to have a test set of 60,000, as large as the training set, but the remaining 50,000 digits have been lost. The authors made many attempts to reconstruct the way MNIST was obtained from the NIST handwriting database as closely as possible and present QMNIST(v5) which features an additional 50,000 test images for MNIST, while the rest of the images are very close to the originals from MNIST. They test their dataset using multiple classification methods and find little difference in whether MNIST or QMNIST is used for training, but the test error on the additional 50,000 images is consistently higher than on the original 10,000 test images or their reconstruction of these. While the concerns about overuse of a test set are justified, the measured effects were mostly small and their relevance might be outweighed by the usefulness of paired differences for statistical model selection.

Flo's opinion: I am confused about the overfitting part, as most methods they try (like ResNets) don't seem to have been selected for performance on the MNIST test set. Granted, LeNet seems to degrade more than other models, but it seems like the additional test images in QMNIST are actually harder to classify. This seems especially plausible with the previous summary in mind and because the authors mention a dichotomy between the ease of classification for NIST images generated by highschoolers vs government employees but don’t seem to mention any attempts to deal with potential selection bias.

FEEDBACK

I'm always happy to hear feedback; you can send it to me, **Rohin Shah**, by **replying to this email**.

PODCAST

An audio podcast version of the **Alignment Newsletter** is available. This podcast is an audio version of the newsletter, recorded by **Robert Miles**.

If AI is based on GPT, how to ensure its safety?

Imagine that an advance robot is built, which is uses GPT-7 as its brain. It takes all previous states of the world and predicts the next step. If a previous state of the world includes a command, like "bring me a cup of coffee", it predicts that it should bring coffee and also predicts all needed movements of robot's limbs. GPT-7 is trained on a large massive of human and other robots data, it has 100 trillions parameters and completely opaque. Its creators have hired you to make the robot safer, but do not allow to destroy it.