Best of LessWrong: July 2015

1. Astronomy, Astrobiology, & The Fermi Paradox I: Introductions, and Space & Time
2. MIRI's Approach
3. A Year of Spaced Repetition Software in the Classroom
4. Experiences in applying "The Biodeterminist's Guide to Parenting"
5. Wear a Helmet While Driving a Car
6. The Unfriendly Superintelligence next door
7. Don't steer with guilt
8. Shifting guilt
9. Update from the suckerpunch
10. Be a new homunculus

Best of LessWrong: July 2015

1. Astronomy, Astrobiology, & The Fermi Paradox I: Introductions, and Space & Time
2. MIRI's Approach
3. A Year of Spaced Repetition Software in the Classroom
4. Experiences in applying "The Biodeterminist's Guide to Parenting"
5. Wear a Helmet While Driving a Car
6. The Unfriendly Superintelligence next door
7. Don't steer with guilt
8. Shifting guilt
9. Update from the suckerpunch
10. Be a new homunculus

Astronomy, Astrobiology, & The Fermi Paradox I: Introductions, and Space & Time

This is the first in a series of posts I am putting together on a personal blog I just started two days ago as a collection of my musings on astrobiology ("The Great A'Tuin" - sorry, I couldn't help it), and will be reposting here. Much has been written here about the Fermi paradox and the 'great filter'. It seems to me that going back to a somewhat more basic level of astronomy and astrobiology is extremely informative to these questions, and so this is what I will be doing. The bloggery is intended for a slightly more general audience than this site (hence much of the content of the introduction) but I think it will be of interest. Many of the points I will be making are ones I have touched on in previous comments here, but hope to explore in more detail.

This post references my first two posts - an introduction, and a discussion of our apparent position in space and time in the universe. The blog posts may be found at:

http://thegreatatuin.blogspot.com/2015/07/whats-all-this-about.html

http://thegreatatuin.blogspot.com/2015/07/space-and-time.htm

MIRI's Approach

MIRI's summer fundraiser is ongoing. In the meantime, we're writing a number of blog posts to explain what we're doing and why, and to answer a number of common questions. This post is one I've been wanting to write for a long time; I hope you all enjoy it. For earlier posts in the series, see the bottom of the above link.

MIRI's mission is "to ensure that the creation of smarter-than-human artificial intelligence has a positive impact." How can we ensure any such thing? It's a daunting task, especially given that we don't have any smarter-than-human machines to work with at the moment. In a previous post to the MIRI Blog I discussed four background claims that motivate our mission; in this post I will describe our approach to addressing the challenge.

This challenge is sizeable, and we can only tackle a portion of the problem. For this reason, we specialize. Our two biggest specializing assumptions are as follows:

> **1. We focus on scenarios where smarter-than-human machine intelligence is first created in de novo software systems (as opposed to, say, brain emulations).** This is in part because it seems difficult to get all the way to brain emulation before someone reverse-engineers the algorithms used by the brain and uses them in a software system, and in part because we expect that any highly reliable AI system will need to have at least some components built from the ground up for safety and transparency. Nevertheless, it is quite plausible that early superintelligent systems will not be human-designed software, and I strongly endorse research programs that focus on reducing risks along the other pathways.
>
> **2. We specialize almost entirely in technical research.** We select our researchers for their proficiency in mathematics and computer science, rather than forecasting expertise or political acumen. I stress that this is only one part of the puzzle: figuring out how to build the right system is useless if the right system does not in fact get built, and ensuring AI has a positive impact is not simply a technical problem. It is also a global coordination problem, in the face of short-term incentives to cut corners. Addressing these non-technical challenges is an important task that we do not focus on.

In short, MIRI does technical research to ensure that de novo AI software systems will have a positive impact. We do not further discriminate between different types of AI software systems, nor do we make strong claims about exactly how quickly we expect AI systems to attain superintelligence. Rather, our current approach is to select open problems using the following question:

> *What would we still be unable to solve, even if the challenge were far simpler?*

For example, we might study AI alignment problems that we could not solve even if we had lots of computing power and very simple goals.

We then filter on problems that are (1) tractable, in the sense that we can do productive mathematical research on them today; (2) uncrowded, in the sense that the problems are not likely to be addressed during normal capabilities research; and

(3) critical, in the sense that they could not be safely delegated to a machine unless we had first solved them ourselves.[1]

These three filters are usually uncontroversial. The controversial claim here is that the above question — "what would we be unable to solve, even if the challenge were simpler?" — is a generator of open technical problems for which solutions will help us design safer and more reliable AI software in the future, regardless of their architecture. The rest of this post is dedicated to justifying this claim, and describing the reasoning behind it.

1. Creating a powerful AI system without understanding why it works is dangerous.

A large portion of the risk from machine superintelligence comes from the possibility of people building systems that they do not fully understand. Currently, this is commonplace in practice: many modern AI researchers are pushing the capabilities of deep neural networks in the absence of theoretical foundations that describe why they're working so well or a solid idea of what goes on beneath the hood. These shortcomings are being addressed over time: many AI researchers are currently working on transparency tools for neural networks, and many more are working to put theoretical foundations beneath deep learning systems. In the interim, using trial and error to push the capabilities of modern AI systems has led to many useful applications.

When designing a superintelligent agent, by contrast, we will want an unusually high level of confidence in its safety before we begin online testing: trial and error alone won't cut it, in that domain.

To illustrate, consider a study by Bird and Layzell in 2002. They used some simple genetic programming to design an oscillating circuit on a circuit board. One solution that the genetic algorithm found entirely avoided using the built-in capacitors (an essential piece of hardware in human-designed oscillators). Instead, it repurposed the circuit tracks on the motherboard as a radio receiver, and amplified an oscillating signal from a nearby computer.

This demonstrates that powerful search processes can often reach their goals via unanticipated paths. If Bird and Layzell were hoping to use their genetic algorithm to find code for a robust oscillating circuit — one that could be used on many different circuit boards regardless of whether there were other computers present — then they would have been sorely disappointed. Yet if they had tested their algorithms extensively on a virtual circuit board that captured all the features of the circuit board that they *thought* were relevant (but not features such as "circuit tracks can carry radio signals"), then they would not have noticed the potential for failure during testing. If this is a problem when handling simple genetic search algorithms, then it will be a much larger problem when handling smarter-than-human search processes.

When it comes to designing smarter-than-human machine intelligence, extensive testing is essential, but not sufficient: in order to be confident that the system will not find unanticipated bad solutions when running in the real world, it is important to have a solid understanding of how the search process works and why it is expected to generate only satisfactory solutions *in addition* to empirical test data.

MIRI's research program is aimed at ensuring that we have the tools needed to inspect and analyze smarter-than-human search processes before we deploy them.

By analogy, neural net researchers could probably have gotten quite far without having any formal understanding of probability theory. Without probability theory, however, they would lack the tools needed to understand modern AI algorithms: they wouldn't know about Bayes nets, they wouldn't know how to formulate assumptions like "independent and identically distributed," and they wouldn't quite know the conditions under which Markov Decision Processes work and fail. They wouldn't be able to talk about priors, or check for places where the priors are zero (and therefore identify things that their systems cannot learn). They wouldn't be able to talk about bounds on errors and prove nice theorems about algorithms that find an optimal policy eventually.

They probably could have still gotten pretty far (and developed half-formed ad-hoc replacements for many of these ideas), but without probability theory, I expect they would have a harder time designing highly reliable AI algorithms. Researchers at MIRI tend to believe that similarly large chunks of AI theory are still missing, and those are the tools that our research program aims to develop.

2. We could not yet create a beneficial AI system even via brute force.

Imagine you have a Jupiter-sized computer and a very simple goal: Make the universe contain as much diamond as possible. The computer has access to the internet and a number of robotic factories and laboratories, and by "diamond" we mean carbon atoms covalently bound to four other carbon atoms. (Pretend we don't care how it makes the diamond, or what it has to take apart in order to get the carbon; the goal is to study a simplified problem.) Let's say that the Jupiter-sized computer is running python. How would you program it to produce lots and lots of diamond?

As it stands, we do not yet know how to program a computer to achieve a goal like that.

We couldn't yet create an artificial general intelligence *by brute force*, and this indicates that there are parts of the problem we don't yet understand.

There are a number of AI tasks that we *could* brute-force. For example, we could write a program that would be *really, really good* at solving computer vision problems: if we had an indestructible box that outputted a picture of a scene and a series of questions about it, waited for answers, scored the answers for accuracy, and then repeated the process, then we know how to write the program that interacts with that box and gets very good at answering the questions. (The program would essentially be a bounded version of AIXI.)

By a similar method, if we had an indestructible box that outputted a conversation and questions about the conversation, waited for natural-language answers to the questions, and scored them for accuracy, then again, we could write a program that would get very good at answering well. In this sense, we know how to solve computer vision and natural language processing by brute force. (Of course, natural-language processing is nowhere near "solved" in a practical sense — there is still loads of work to be done. A brute force solution doesn't get you very far in the real world. The point

is that, for many AI alignment problems, we haven't even made it to the "we could brute force it" level yet.)

Why do we need the indestructible box in the above examples? Because the way the modern brute-force solution would work is by considering each Turing machine (up to some complexity limit) as a hypothesis about the box, seeing which ones are consistent with observation, and then executing actions that lead to high scores coming out of the box (as predicted by the remaining hypotheses, weighted by simplicity).

Each hypothesis is an opaque Turing machine, and the algorithm never peeks inside: it just asks each hypothesis to predict what score the box will output, without concern for what mechanism is being used to generate that score. This means that if the algorithm finds (via exhaustive search) a plan that *maximizes* the score coming out of the box, and the box is destructible, then the opaque action chain that maximizes score is very likely to be the one that pops the box open and alters it so that it always outputs the highest score. But given an indestructible box, we know how to brute force the answers.

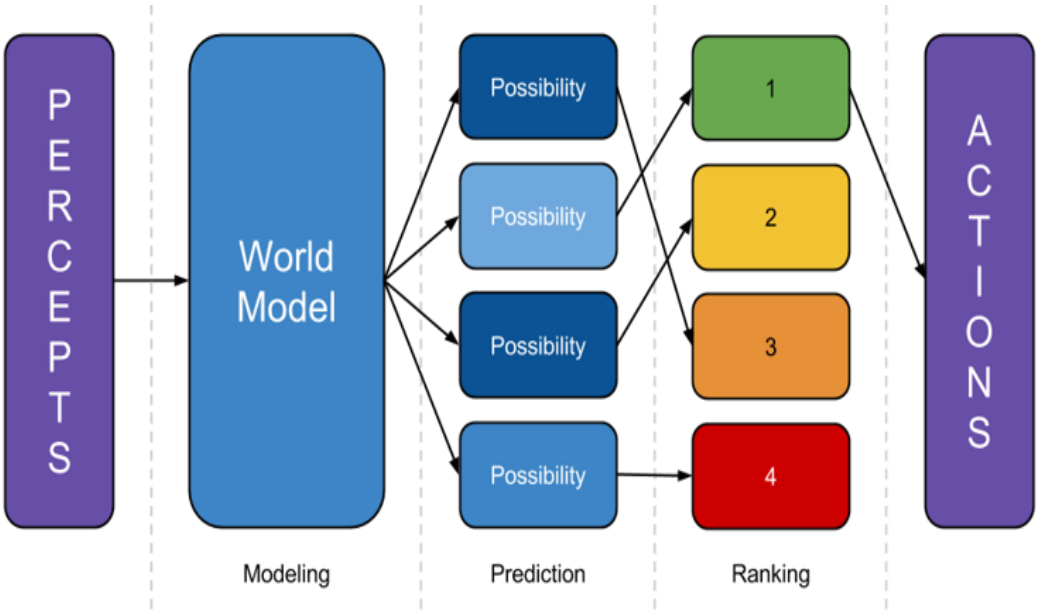
In fact, roughly speaking, we understand how to solve *any* reinforcement learning problem via brute force. This is a far cry from knowing how to *practically* solve reinforcement learning problems! But it does illustrate a difference in kind between two types of problems. We can (imperfectly and heuristically) divide them up as follows:

> *There are two types of open problem in AI. One is figuring out a practical way to solve a problem that we know how to solve in principle. The other is figuring out how to solve problems that we don't even know how to brute force yet.*

MIRI focuses on problems of the second class.[2]

What is hard about brute-forcing a diamond-producing agent? To illustrate, I'll give a wildly simplified sketch of what an AI program needs to do in order to act productively within a complex environment:

1. Model the world: Take percepts, and use them to refine some internal representation of the world the system is embedded in.
2. Predict the world: Take that world-model, and predict what would happen if the system executed various different plans.
3. Rank outcomes: Rate those possibilities by how good the predicted future is, then execute a plan that leads to a highly-rated outcome.[3]



Consider the modeling step. As discussed above, we know how to write an algorithm that finds good world-models by brute force: it looks at lots and lots of Turing machines, weighted by simplicity, treats them like they are responsible for its observations, and throws out the ones that are inconsistent with observation thus far. But (aside from being wildly impractical) this yields only *opaque* hypotheses: the system can ask what "sensory bits" each Turing machine outputs, but it cannot peek inside and examine objects represented within.

If there is some well-defined "score" that gets spit out by the opaque Turing machine (as in a reinforcement learning problem), then it doesn't matter that each hypothesis is a black box; the brute-force algorithm can simply run the black box on lots of inputs and see which results in the highest score. But if the problem is to build lots of diamond in the real world, then the agent must work as follows:

1. Build a model of the world — one that represents carbon atoms and covalent bonds, among other things.
2. Predict how the world would change contingent on different actions the system could execute.
3. Look *inside* each prediction and see which predicted future has the most diamond. Execute the action that leads to more diamond.

In other words, an AI that is built to reliably affect *things in the world* needs to have world-models that are amenable to inspection. The system needs to be able to pop the world model, identify the representations of carbon atoms and covalent bonds, and estimate how much diamond is in the real world.[4]

We don't yet have a clear picture of how to build "inspectable" world-models — not even by brute force. Imagine trying to write the part of the diamond-making program that builds a world-model: this function needs to take percepts as input and build a data structure that represents the universe, in a way that allows the system to inspect universe-descriptions and estimate the amount of diamond in a possible future. Where in the data structure are the carbon atoms? How does the data structure allow the

concept of a “covalent bond” to be formed and labeled, in such a way that it remains accurate even as the world-model stops representing diamond as made of atoms and starts representing them as made of protons, neutrons, and electrons instead?

We need a world-modeling algorithm that builds multi-level representations of the world and allows the system to pursue the same goals (make diamond) even as its model changes drastically (because it discovers quantum mechanics). This is in stark contrast to the existing brute-force solutions that use opaque Turing machines as hypotheses.[5]

When *humans* reason about the universe, we seem to do some sort of reasoning outwards from the middle: we start by modeling things like people and rocks, and eventually realize that these are made of atoms, which are made of protons and neutrons and electrons, which are perturbations in quantum fields. At no point are we certain that the lowest level in our model is the lowest level in reality; as we continue thinking about the world we *construct* new hypotheses to explain oddities in our models. What sort of data structure are we using, there? How do we add levels to a world model given new insights? This is the sort of reasoning algorithm that we do not yet understand how to formalize.[6]

That’s step *one* in brute-forcing an AI that reliably pursues a simple goal. We also don’t know how to brute-force steps two or three yet. By simplifying the problem — talking about diamonds, for example, rather than more realistic goals that raise a host of other difficulties — we’re able to factor out the parts of the problems that we don’t understand how to solve yet, even in principle. Our technical agenda describes a number of open problems identified using this method.

3. Figuring out how to solve a problem in principle yields many benefits.

In 1836, Edgar Allen Poe wrote a wonderful essay on Maelzel’s Mechanical Turk, a machine that was purported to be able to play chess. In the essay, Poe argues that the Mechanical Turk must be a hoax: he begins by arguing that machines cannot play chess, and proceeds to explain (using his knowledge of stagecraft) how a person could be hidden within the machine. Poe’s essay is remarkably sophisticated, and a fun read: he makes reference to the “calculating machine of Mr. Babbage” and argues that it cannot possibly be made to play chess, because in a calculating machine, each steps follows from the previous step by necessity, whereas “no one move in chess necessarily follows upon any one other”.

The Mechnical Turk indeed turned out to be a hoax. In 1950, however, Claude Shannon published a rather compelling counterargument to Poe’s reasoning in the form of a paper explaining how to program a computer to play perfect chess.

Shannon’s algorithm was by no means the end of the conversation. It took forty-six years to go from that paper to Deep Blue, a practical chess program which beat the human world champion. Nevertheless, if you were equipped with Poe’s state of knowledge and not yet sure whether it was *possible* for a computer to play chess — because you did not yet understand algorithms for constructing game trees and doing backtracking search — then you would probably not be ready to start writing practical chess programs.

Similarly, if you lacked the tools of probability theory — an understanding of Bayesian inference and the limitations that stem from bad priors — then you probably wouldn't be ready to program an AI system that needed to manage uncertainty in high-stakes situations.

If you are trying to write a program and you can't yet say how you would write it given a computer the size of Jupiter, then you probably aren't yet ready to design a practical approximation of the brute-force solution yet. Practical chess programs can't generate a full search tree, and so rely heavily on heuristics and approximations; but if you can't brute-force the answer yet given *arbitrary* amounts of computing power, then it's likely that you're missing some important conceptual tools.

Marcus Hutter (inventor of AIXI) and Shane Legg (inventor of the Universal Measure of Intelligence) seem to endorse this approach. Their work can be interpreted as a description of how to find a brute-force solution to any reinforcement learning problem, and indeed, the above description of how to do this is due to Legg and Hutter.

In fact, the founders of Google DeepMind reference the completion of Shane's thesis as one of four key indicators that the time was ripe to begin working on AGI: a theoretical framework describing how to solve reinforcement learning problems *in principle* demonstrated that modern understanding of the problem had matured to the point where it was time for the practical work to begin.

Before we gain a formal understanding of the problem, we can't be quite sure what the problem *is*. We may fail to notice holes in our reasoning; we may fail to bring the appropriate tools to bear; we may not be able to tell when we're making progress. After we gain a formal understanding of the problem in principle, we'll be in a better position to make practical progress.

The point of developing a formal understanding of a problem is not to *run* the resulting algorithms. Deep Blue did not work by computing a full game tree, and DeepMind is not trying to implement AIXI. Rather, the point is to identify and develop the basic concepts and methods that are useful for solving the problem (such as game trees and backtracking search algorithms, in the case of chess).

The development of probability theory has been quite useful to the field of AI — not because anyone goes out and attempts to build a perfect Bayesian reasoner, but because probability theory is the unifying theory for reasoning under uncertainty. This makes the tools of probability theory useful for AI designs that vary in any number of implementation details: any time you build an algorithm that attempts to manage uncertainty, a solid understanding of probabilistic inference is helpful when reasoning about the domain in which the system will succeed and the conditions under which it could fail.

This is why we think we can identify open problems that we can work on today, and which will reliably be useful no matter how the generally intelligent machines of the future are designed (or how long it takes to get there). By seeking out problems that we couldn't solve even if the problem were much easier, we hope to identify places where core AGI algorithms are missing. By developing a formal understanding of how to address those problems in principle, we aim to ensure that when it comes time to address those problems in practice, programmers have the knowledge they need to develop solutions that they deeply understand, and the tools they need to ensure that the systems they build are highly reliable.

4. This is an approach researchers have used successfully in the past.

Our main open-problem generator — "what would we be unable to solve even if the problem were easier?" — is actually a fairly common one used across mathematics and computer science. It's more easy to recognize if we rephrase it slightly: "can we reduce the problem of building a beneficial AI to some other, simpler problem?"

For example, instead of asking whether you can program a Jupiter-sized computer to produce diamonds, you could rephrase this as a question about whether we can reduce the diamond maximization problem to known reasoning and planning procedures. (The current answer is "not yet.")

This is a fairly standard practice in computer science, where reducing one problem to another is a key feature of computability theory. In mathematics it is common to achieve a proof by reducing one problem to another (see, for instance, the famous case of Fermat's last theorem). This helps one focus on the parts of the problem that *aren't* solved, and identify topics where foundational understanding is lacking.

As it happens, humans have a pretty good track record when it comes to working on problems such as these. Humanity has a poor track record at predicting long-term technological trends, but we have a reasonably good track record at developing theoretical foundations for technical problems decades in advance, when we put sufficient effort into it. Alan Turing and Alonzo Church succeeded in developing a robust theory of computation that proved quite useful once computers were developed, in large part by figuring out how to solve (in principle) problems which they did not yet know how to solve with machines. Andrey Kolmogorov, similarly, set out to formalize intuitive but not-yet-well-understood methods for managing uncertainty; and he succeeded. And Claude Shannon and his contemporaries succeeded at this endeavor in the case of chess.

The development of probability theory is a particularly good analogy to our case: it is a field where, for hundreds of years, philosophers and mathematicians who attempted to formalize their intuitive notions of uncertainty repeatedly reasoned themselves into paradoxes and contradictions. The probability theory at the time, sorely lacking formal foundations, was dubbed a "theory of misfortune." Nevertheless, a concerted effort by Kolmogorov and others to formalize the theory was successful, and his efforts inspired the development of a host of useful tools for designing systems that reason reliably under uncertainty.

Many people who set out to put foundations under a new field of study (that was intuitively understood on some level but not yet formalized) have succeeded, and their successes have been practically significant. We aim to do something similar for a number of open problems pertaining to the design of highly reliable reasoners.

The questions MIRI focuses on, such as "how would one ideally handle logical uncertainty?" or "how would one ideally build multi-level world models of a complex environment?", exist at a level of generality comparable to Kolmogorov's "how would one ideally handle empirical uncertainty?" or Hutter's "how would one ideally maximize reward in an arbitrarily complex environment?" The historical track record suggests that these are the kinds of problems that it is possible to both (a) see coming

in advance, and (b) work on without access to a concrete practical implementation of a general intelligence.

By identifying parts of the problem that we would still be unable to solve even if the problem was easier, we hope to hone in on parts of the problem where core algorithms and insights are missing: algorithms and insights that will be useful no matter what architecture early intelligent machines take on, and no matter how long it takes to create smarter-than-human machine intelligence.

At present, there are only three people on our research team, and this limits the number of problems that we can tackle ourselves. But our approach is one that we can scale up dramatically: our approach has generated a very large number of open problems, and we have no shortage of questions to study.[7]

This is an approach that has often worked well in the past for humans trying to understand how to approach a new field of study, and I am confident that this approach is pointing us towards some of the core hurdles in this young field of AI alignment.

This post is cross-posted from the MIRI blog. It's part of a series we're writing on MIRI's strategy and plans for the future, as part of our ongoing 2015 Summer Fundraiser.

---

[1] Since the goal is to design intelligent machines, there are many technical problems that we can expect to eventually delegate to those machines. But it is difficult to trust an unreliable reasoner with the task of designing reliable reasoning! ↩

[2] Most of the AI field focuses on problems of the first class. Deep learning, for example, is a very powerful and exciting tool for solving problems that we know how to brute-force, but which were, up until a few years ago, wildly intractable. Class 1 problems tend to be important problems for building more capable AI systems, but lower-priority for ensuring that highly capable systems are aligned with our interests. ↩

[3] In reality, of course, there aren't clean separations between these steps. The "prediction" step must be more of a ranking-dependent planning step, to avoid wasting computation predicting outcomes that will obviously be poorly-ranked. The modeling step depends on the prediction step, because which parts of the world-model are refined depends on what the world-model is going to be used for. A realistic agent would need to make use of meta-planning to figure out how to allocate resources between these activities, etc. This diagram is a fine first approximation, though: if a system doesn't do something like modeling the world, predicting outcomes, and ranking them somewhere along the way, then it will have a hard time steering the future. ↩

[4] In reinforcement learning problems, this issue is avoided via a special "reward channel" intended to stand in indirectly for something the supervisor wants. (For example, the supervisor may push a reward button every time the learner takes an action that seems, to the supervisor, to be useful for making diamonds.) Then the programmers can, by hand, single out the reward channel inside the world-model and program the system to execute actions that it predicts lead to high reward. This is much easier than designing world-models in such a way that the system can reliably

identify representations of carbon atoms and covalent bonds within it (especially if the world is modeled in terms of Newtonian mechanics one day and quantum mechanics the next), but doesn't provide a framework for agents that must autonomously learn how to achieve some goal. Correct behavior in highly intelligent systems will not always be reducible to maximizing a reward signal controlled by a significantly less intelligent system (e.g., a human supervisor). ↩

[5] The idea of a search algorithm that optimizes according to modeled *facts about the world* rather than just *expected percepts* may sound basic, but we haven't found any deep insights (or clever hacks) that allow us to formalize this idea (e.g., as a brute-force algorithm). If we could formalize it, we would likely get a better understanding of the kind of abstract modeling of objects and facts that is required for self-referential, logically uncertain, programmer-inspectable reasoning. ↩

[6] We also suspect that a brute-force algorithm for building multi-level world models would be much more amenable to being "scaled down" than Solomonoff induction, and would therefore lend some insight into how to build multi-level world models in a practical setting. ↩

[7] For example, instead of asking what problems remain when given lots of computing power, you could instead ask whether we can reduce the problem of building an aligned AI to the problem of making reliable predictions about human behavior: an approach advocated by others. ↩

A Year of Spaced Repetition Software in the Classroom

Last year, I asked LW for some advice about spaced repetition software (SRS) that might be useful to me as a high school teacher. With said advice came a request to write a follow-up after I had accumulated some experience using SRS in the classroom. This is my report.

Please note that this was not a scientific experiment to determine whether SRS "works." Prior studies are already pretty convincing on this point and I couldn't think of a practical way to run a control group or "blind" myself. What follows is more of an informal debriefing for how I used SRS during the 2014-15 school year, my insights for others who might want to try it, and how the experience is changing how I teach.

Summary

SRS can raise student achievement even with students who won't use the software on their own, and even with frequent disruptions to the study schedule. Gains are most apparent with the already high-performing students, but are also meaningful for the lowest students. Deliberate efforts are needed to get student buy-in, and getting the most out of SRS may require changes in course design.

The software

After looking into various programs, including the game-like Memrise, and even writing my own simple SRS, **I ultimately went with Anki** for its multi-platform availability, cloud sync, and ease-of-use. I also wanted a program that could act as an impromptu catch-all bin for the 2,000+ cards I would be producing on the fly throughout the year. (Memrise, in contrast, really needs clearly defined units packaged in advance).

The students

I teach 9th and 10th grade English at an above-average suburban American public high school in a below-average state. Mine are the lower "required level" students at a school with high enrollment in honors and Advanced Placement classes. Generally speaking, this means my students are mostly not self-motivated, are only very weakly motivated by grades, and will not do anything school-related outside of class no matter how much it would be in their interest to do so. There are, of course, plenty of exceptions, and my students span an extremely wide range of ability and apathy levels.

The procedure

First, what I did *not* do. I did *not* make Anki decks, assign them to my students to study independently, and then quiz them on the content. With honors classes I taught in previous years I think that might have worked, but I know my current students too well. Only about 10% of them would have done it, and the rest would have blamed me for their failing grades—with some justification, in my opinion.

Instead, **we did Anki together, as a class**, nearly every day.

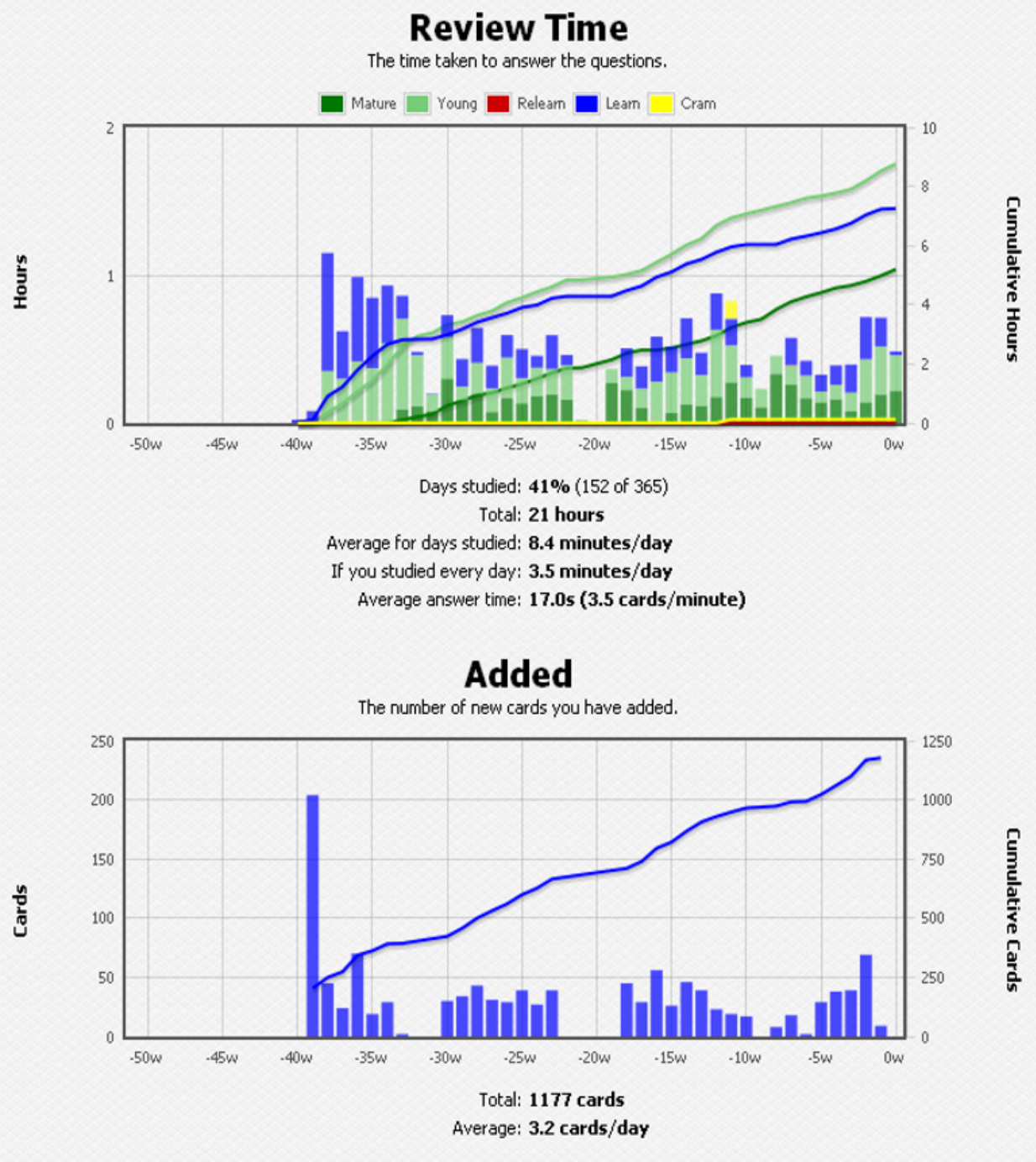
As initial setup, I created a separate Anki profile for each class period. With a third-party add-on for Anki called Zoom, I enlarged the display font sizes to be clearly legible on the interactive whiteboard at the front of my room.

Nightly, I wrote up cards to reinforce new material and integrated them into the deck in time for the next day's classes. This averaged **about 7 new cards per lesson period.**These cards came in many varieties, but the three main types were:

1. **concepts and terms**, often with reversed companion cards, sometimes supplemented with "what is this an example of" scenario cards.
2. **vocabulary**, 3 cards per word: word/def, reverse, and fill-in-the-blank example sentence
3. **grammar**, usually in the form of "What change(s), if any, does this sentence need?" Alternative cards had different permutations of the sentence.

Weekly, I updated the deck to the cloud for self-motivated students wishing to study on their own.

Daily, I led each class in an Anki review of new and due cards for **an average of 8 minutes per study day**, usually as our first activity, at a rate of about **3.5 cards per minute**. As each card appeared on the interactive whiteboard, I would read it out loud while students willing to share the answer raised their hands. Depending on the card, I might offer additional time to think before calling on someone to answer. Depending on their answer, and my impressions of the class as a whole, I might elaborate or offer some reminders, mnemonics, etc. I would then quickly poll the class on how they felt about the card by having them show a color by way of a small piece of card-stock divided into green, red, yellow, and white quadrants. Based on my own judgment (informed only partly by the poll), I would choose and press a response button in Anki, determining when we should see that card again.



[Data shown is from one of my five classes. We didn't start using Anki until a couple weeks into the school year.]

Opportunity costs

8 minutes is a significant portion of a 55 minute class period, especially for a teacher like me who fills every one of those minutes. Something had to give. For me, I entirely cut some varieties of written vocab reinforcement, and reduced the time we spent playing the team-based vocab/term review game I wrote for our interactive

whiteboards some years ago. To a lesser extent, I also cut back on some oral reading comprehension spot-checks that accompany my whole-class reading sessions. On balance, I think **Anki was a much better way to spend the time**, but it's complicated. Keep reading.

Whole-class SRS not ideal

Every student is different, and would get the most out of having a personal Anki profile determine when they should see each card. Also, most individuals could study many more cards per minute on their own than we averaged doing it together. (To be fair, a small handful of my students did use the software independently, judging from Ankiweb download stats)

Getting student buy-in

Before we started using SRS I tried to sell my students on it with a heartfelt, over-prepared 20 minute presentation on how it works and the superpowers to be gained from it. It might have been a waste of time. It might have changed someone's life. Hard to say.

As for the daily class review, I induced engagement partly through participation points that were part of the final semester grade, and which students knew I tracked closely. Raising a hand could earn a kind of bonus currency, but was never required—unlike looking up front and showing colors during polls, which I insisted on. When I thought students were just reflexively holding up the same color and zoning out, I would sometimes spot check them on the last card we did and penalize them if warranted.

But because I know my students are not strongly motivated by grades, I think the most important influence was my attitude. I made it a point to really turn up the charm during review and play the part of the engaging game show host. Positive feedback. Coaxing out the lurkers. Keeping that energy up. Being ready to kill and joke about bad cards. Reminding classes how awesome they did on tests and assignments because they knew their Anki stuff.

(This is a good time to point out that the average review time per class period stabilized at about 8 minutes because **I tried to end reviews before student engagement tapered off too much, which typically started happening at around the 6-7 minute mark.** Occasional short end-of-class reviews mostly account for the difference.)

I also got my students more on the Anki bandwagon by showing them how this was directly linked **reduced note-taking requirements**. If I could trust that they would remember something through Anki alone, why waste time waiting for them to write it down? They were unlikely to study from those notes anyway. And if they aren't looking down at their paper, they'll be paying more attention to *me*. I better come up with more cool things to tell them!

Making memories

Everything I had read about spaced repetition suggested it was a great reinforcement tool but not a good way to introduce new material. With that in mind, I tried hard to find or create memorable images, examples, mnemonics, and anecdotes that my Anki

cards could become hooks for, and to get those cards into circulation as soon as possible. I even gave this method a mantra: **"vivid memory, card ready"**.

When a student during review raised their hand, gave me a pained look, and said, "like that time when...." or "I can see that picture of..." as they struggled to remember, I knew I had done well. (And I would always wait a moment, because they would usually get it.)

Baby cards need immediate love

Unfortunately, if the card wasn't introduced quickly enough—within a day or two of the lesson—the entire memory often vanished and had to be recreated, killing the momentum of our review. This happened far too often—not because I didn't write the card soon enough (I stayed really on top of that), but because it didn't always come up for study soon enough. There were a few reasons for this:

1. We often had too many due cards to get through in one session, and by default **Anki puts new cards behind due ones**.
2. By default, Anki only introduces 20 new cards in one session (I soon uncapped this).
3. Some cards were in categories that I gave lower priority to.

Two obvious cures for this problem:

1. Make fewer cards. (I did get more selective as the year went on.)
2. Have all cards prepped ahead of time and introduce new ones at the end of the class period they go with. (For practical reasons, not the least of which was the fact that I didn't always know what cards I was making until after the lesson, I did not do this. I might able to next year.)

Days off suck

SRS is meant to be used every day. When you take weekends off, you get a backlog of due cards. Not only do my students take every weekend and major holiday off (slackers), they have a few 1-2 week vacations built into the calendar. Coming back from a week's vacation means a 9-day backlog (due to the weekends bookending it). There's no good workaround for students that won't study on their own. The best I could do was run longer or multiple Anki sessions on return days to try catch up with the backlog. It wasn't enough. The "caught up" condition was not normal for most classes at most points during the year, but rather something to aspire to and occasionally applaud ourselves for reaching. Some cards spent weeks or months on the bottom of the stack. Memories died. Baby cards emerged stillborn. Learning was lost.

Needless to say, the last weeks of the school year also had a certain silliness to them. When the class will never see the card again, it doesn't matter whether I push the button that says 11 days or the one that says 8 months. (So I reduced polling and accelerated our cards/minute rate.)

Never before SRS did I fully appreciate the loss of learning that must happen every summer break.

Triage

I kept each course's master deck divided into a few large subdecks. This was initially for organizational reasons, but I eventually started using it as a prioritizing tool. This happened after a curse-worthy discovery: if you tell Anki to review a deck made from subdecks, **due cards from subdecks higher up in the stack are shown before cards from decks listed below**, no matter how overdue they might be. From that point, on days when we were backlogged (most days) I would specifically review the concept/terminology subdeck for the current semester before any other subdecks, as these were my highest priority.

On a couple of occasions, I also used Anki's study deck tools to create temporary decks of especially high-priority cards.

Seizing those moments

Veteran teachers start acquiring a sense of when it might be a good time to go off book and teach something that isn't in the unit, and maybe not even in the curriculum. Maybe it's teaching exactly the right word to describe a vivid situation you're reading about, or maybe it's advice on what to do in a certain type of emergency that nearly happened. As the year progressed, I found myself humoring my instincts more often because of a new confidence that I can **turn an impressionable moment into a strong memory and lock it down with a new Anki card**. *I don't even care if it will ever be on a test.* This insight has me questioning a great deal of what I thought knew about organizing a curriculum. And I like it.

A lifeline for low performers

An accidental discovery came from having written some cards that were, it was immediately obvious to me, much too easy. I was embarrassed to even be reading them out loud. Then I saw which hands were coming up.

In any class you'll get some small number of extremely low performers who never seem to be doing anything that we're doing, and, when confronted, deny that they have any ability whatsoever. Some of the hands I was seeing were attached to these students. And you better believe I called on them.

It turns out that **easy cards are really important** because they can give wins to students who desperately need them. Knowing a 6th grade level card in a 10th grade class is no great achievement, of course, but the action takes what had been negative morale and nudges it upward. And it can trend. I can build on it. A few of these students started making Anki the thing they did in class, even if they ignored everything else. I can confidently name one student I'm sure passed my class *only* because of Anki. Don't get me wrong—he just barely passed. Most cards remained over his head. Anki was no miracle cure here, but it gave him and I something to work with that we didn't have when he failed my class the year before.

A springboard for high achievers

It's not even fair. The lowest students got something important out of Anki, but the highest achievers drank it up and used it for rocket fuel. When people ask who's widening the achievement gap, I guess I get to raise my hand now.

I refuse to feel bad for this. Smart kids are badly underserved in American public schools thanks to policies that encourage staff to focus on that slice of students near

(but not at) the bottom—the ones who might just barely be able to pass the state test, given enough attention.

Where my bright students might have been used to high Bs and low As on tests, they were now breaking my scales. You could see it in the multiple choice, but it was most obvious in their writing: they were skillfully working in terminology at an unprecedented rate, and making way more attempts to use new vocabulary—attempts that were, for the most part, successful.

Given the seemingly objective nature of Anki it might seem counterintuitive that the benefits would be more obvious in writing than in multiple choice, but it actually makes sense when I consider that even without SRS these students probably would have known the terms and the vocab well enough to get multiple choice questions right, but might have lacked the confidence to use them on their own initiative. **Anki gave them that extra confidence.**

A wash for the apathetic middle?

I'm confident that about a third of my students got very little out of our Anki review. They were either really good at faking involvement while they zoned out, or didn't even try to pretend and just took the hit to their participation grade day after day, no matter what I did or who I contacted.

These weren't even necessarily failing students—just the apathetic middle that's smart enough to remember some fraction of what they hear and regurgitate some fraction of that at the appropriate times. Review of any kind holds no interest for them. It's a rerun. They don't *really* know the material, but they tell themselves that they do, and they don't care if they're wrong.

On the one hand, these students are no worse off with Anki than they would have been with with the activities it replaced, and nobody cries when average kids get average grades. On the other hand, I'm not ok with this... but so far I don't like any of my ideas for what to do about it.

Putting up numbers: a case study

For unplanned reasons, I taught a unit at the start of a quarter that I didn't formally test them on until the end of said quarter. Historically, this would have been a disaster. In this case, it worked out well. **For five weeks, Anki was the only ongoing exposure they were getting to that unit, but it proved to be enough.** Because I had given the same test as a pre-test early in the unit, I have some numbers to back it up. The test was all multiple choice, with two sections: the first was on general terminology and concepts related to the unit. The second was a much harder reading comprehension section.

As expected, scores did not go up much on the reading comprehension section. Overall reading levels are very difficult to boost in the short term and I would not expect any one unit or quarter to make a significant difference. The average score there rose by 4 percentage points, from 48 to 52%.

Scores in the terminology and concept section were more encouraging. For material we had not covered until after the pre-test, the average score rose by 22 percentage points, from 53 to 75%. No surprise there either, though; it's hard to say how much credit we should give to SRS for that.

But there were also a number of questions about material we had already covered *before* the pretest. Being the earliest material, I might have expected some degradation in performance on the second test. Instead, the already strong average score in that section rose by an additional 3 percentage points, from 82 to 85%. (These numbers are less reliable because of the smaller number of questions, but they tell me Anki at least "locked in" the older knowledge, and may have strengthened it.)

Some other time, I might try reserving a section of content that I teach before the pre-test but don't make any Anki cards for. This would give me a way to compare Anki to an alternative review exercise.

What about formal standardized tests?

I don't know yet. The scores aren't back. I'll probably be shown some "value added" analysis numbers at some point that tell me whether my students beat expectations, but I don't know how much that will tell me. My students were consistently beating expectations before Anki, and the state gave an entirely different test this year because of legislative changes. I'll go back and revise this paragraph if I learn anything useful.

Those discussions...

If I'm trying to acquire a new skill, one of the first things I try to do is listen to skilled practitioners of that skill talk about it *to each other*. What are the terms-of-art? How do they use them? What does this tell me about how they see their craft? Their shorthand is a treasure trove of crystallized concepts; once I can use it the same way they do, I find I'm working at a level of abstraction much closer to theirs.

Similarly, I was hoping Anki could help make my students more fluent in the subject-specific lexicon that helps you score well in analytical essays. After introducing a new term and making the Anki card for it, I made extra efforts to use it conversationally. I used to shy away from that because so many students would have forgotten it immediately and tuned me out for not making any sense. Not this year. Once we'd seen the card, I used the term freely, with only the occasional reminder of what it meant. I started using multiple terms in the same sentence. I started talking about writing and analysis the way my fellow experts do, and so invited them into that world.

Even though I was already seeing written evidence that some of my high performers had assimilated the lexicon, the high quality *discussions* of these same students caught me off guard. You see, I usually dread whole-class discussions with non-honors classes because good comments are so rare that I end up dejectedly spouting all the insights I had hoped they could find. But by the end of the year, my students had stepped up.

I think what happened here was, as with the writing, as much a boost in confidence as a boost in fluency. Whatever it was, they got into some good discussions where **they used the terminology and built on it to say smarter stuff**.

Don't get me wrong. Most of my students never got to that point. But on average **even small groups without smart kids had a noticeably higher level of discourse** than I am used to hearing when I break up the class for smaller discussions.

Limitations

SRS is inherently weak when it comes to the abstract and complex. No card I've devised enables a student to develop a distinctive authorial voice, or write essay openings that reveal just enough to make the reader curious. Yes, you can make cards about strategies for this sort of thing, but these were consistently my worst cards—the overly difficult "leeches" that I eventually suspended from my decks.

A less obvious limitation of SRS is that students with a very strong grasp of a concept often fail to apply that knowledge in more authentic situations. For instance, they may know perfectly well the difference between "there", "their", and "they're", but never pause to think carefully about whether they're using the right one in a sentence. I am very open to suggestions about how I might train my students' autonomous "System 1" brains to have "interrupts" for that sort of thing... or even just a reflex to *go back and check* after finishing a draft.

Moving forward

I absolutely intend to continue using SRS in the classroom. Here's what I intend to do differently this coming school year:

- **Reduce the number of cards** by about 20%, to maybe 850-950 for the year in a given course, mostly by reducing the number of variations on some overexposed concepts.
- Be more willing to add extra Anki study sessions to **stay better caught-up with the deck**, even if this means my lesson content doesn't line up with class periods as neatly.
- Be more willing to **press the red button on cards we need to re-learn**. I think I was too hesitant here because we were rarely caught up as it was.
- **Rework underperforming cards** to be simpler and more fun.
- Use **more simple cloze deletion** cards. I only had a few of these, but they worked better than I expected for structured idea sets like, "characteristics of a tragic hero".
- Take **a less linear and more opportunistic approach** to introducing terms and concepts.
- Allow for **more impromptu discussions** where we bring up older concepts in relevant situations and build on them.
- Shape more of my lessons around the **"vivid memory, card ready"** philosophy.
- Continue to **reduce needless student note-taking**.
- Keep a close eye on 10th grade students who had me for 9th grade last year. I wonder how much they retained over the summer, and I can't wait to **see what a second year of SRS will do** for them.

Suggestions and comments very welcome!

Experiences in applying "The Biodeterminist's Guide to Parenting"

I'm posting this because LessWrong was very influential on how I viewed parenting, particularly the emphasis on helping one's brain work better. In this context, creating and influencing another person's brain is an awesome responsibility.

It turned out to be a lot more anxiety-provoking than I expected. I don't think that's necessarily a bad thing, as the possibility of screwing up someone's brain should make a parent anxious, but it's something to be aware of. I've heard some blithe "Rational parenting could be a very high-impact activity!" statements from childless LWers who may be interested to hear some experiences in actually applying that.

One thing that really scared me about trying to raise a child with the healthiest-possible brain and body was the possibility that I might not love her if she turned out to not be smart. 15 months in, I'm no longer worried. Evolution has been very successful at producing parents and children that love each other despite their flaws, and our family is no exception. Our daughter Lily seems to be doing fine, but if she turns out to have disabilities or other problems, I'm confident that we'll roll with the punches.

Cross-posted from The Whole Sky.

Before I got pregnant, I read Scott Alexander's excellent Biodeterminist's Guide to Parenting and was so excited to have this knowledge. I thought how lucky my child would be to have parents who knew and cared about how to protect her from things that would damage her brain.

Real life, of course, got more complicated. It's one thing to intend to avoid neurotoxins, but another to arrive at the grandparents' house and find they've just had ant poison sprayed. What do you do then?

Here are some tradeoffs Jeff and I have made between things that are good for children in one way but bad in another, or things that are good for children but really difficult or expensive.

Germs and parasites

The hygiene hypothesis states that lack of exposure to germs and parasites increases risk of auto-immune disease. Our pediatrician recommended letting Lily playing in the dirt for this reason.

While exposure to animal dander and pollution increase asthma later in life, it seems that being exposed to these in the first year of life actually protects against asthma. Apparently if you're going to live in a house with roaches, you should do it in the first year or not at all.

Except some stuff in dirt is actually bad for you.

Scott writes:

> *Parasite-infestedness of an area correlates with national IQ at about r = -0.82. The same is true of US states, with a slightly reduced correlation coefficient of -0.67 (p<0.0001). . . . When an area eliminates parasites (like the US did for malaria and hookworm in the early 1900s) the IQ for the area goes up at about the right time.*

Living with cats as a child seems to increase risk of schizophrenia, apparently via toxoplasmosis. But in order to catch toxoplasmosis from a cat, you have to eat its feces during the two weeks after it first becomes infected (which it's most likely to do by eating birds or rodents carrying the disease). This makes me guess that most kids get it through tasting a handful of cat litter, dirt from the yard, or sand from the sandbox rather than simply through cat ownership. We live with indoor cats who don't seem to be mousers, so I'm not concerned about them giving anyone toxoplasmosis. If we build Lily a sandbox, we'll keep it covered when not in use.

The evidence is mixed about whether infections like colds during the first year of life increase or decrease your risk of asthma later. After the newborn period, we defaulted to being pretty casual about germ exposure.

Toxins in buildings

Our experiences with lead (and lessons learned about how to reduce risk). Our experiences with mercury.

In some areas, it's not that feasible to live in a house with zero lead. We live in Boston, where 87% of the housing was built before lead paint was banned. Even in a new building, we'd need to go far out of town before reaching soil that wasn't near where a lead-painted building had been.

It is possible to do some renovations without exposing kids to lead. Jeff recently did some demolition of walls with lead paint, very carefully sealed off and cleaned up, while Lily and I spent the day elsewhere. Afterwards her lead level was no higher than it had been.

But Jeff got serious lead poisoning as a toddler while his parents did major renovations on their old house. If I didn't think I could keep the child away from the dust, I wouldn't renovate.

Recently a house across the street from us was gutted, with workers throwing debris out the windows and creating big plumes of dust (presumably lead-laden) that blew all down the street. Later I realized I should have called city building inspection services, which would have at least made them carry the debris into the dumpster instead of throwing it from the second story.

Floor varnish releases formaldehyde and other nasties as it cures. We kept Lily out of the house for a few weeks after Jeff redid the floors. We found it worthwhile to pay rent at our previous house in order to not have to live in the new house while this kind of work was happening.

Pressure-treated wood was treated with arsenic and chromium until around 2004 in the US. It often has a greenish tint, though it may not be obvious after fading or staining. Playing on playsets or decks made of such wood increases children's cancer risk. It should not be used for furniture (I thought this would be obvious, but apparently it wasn't to some of my handyman relatives).

I found it difficult to know how to deal with fresh paint and other fumes in my building at work while I was pregnant. Women of reproductive age have a heightened sense of smell, and many pregnant women have heightened aversion to smells, so you can literally smell things some of your coworkers can't (or don't mind). The most critical period of development is during the first trimester, when most women aren't telling the world they're pregnant (because it's also the time when a miscarriage is most likely, and if you do lose the pregnancy you might not want to have to tell everyone). During that period, I found it difficult to explain why I was concerned about the fumes from the roofing adhesive being used in our building. I didn't want to seem like a princess who thought she was too good to work in conditions that everybody else found acceptable. (After I told them I was pregnant, my coworkers were very understanding about such things.)

Food

Recommendations usually focus on what you should eat during pregnancy, but obviously children's brain development doesn't stop there. I've opted to take precautions with the food Lily and I eat for as long as I'm nursing her.

Claims that pesticide residues are poisoning children scare me, although most scientists seem to think the paper cited is overblown. Other sources say the levels of pesticides in conventionally grown produce are fine. We buy organic produce at home but eat whatever we're served elsewhere.

I would love to see a study with families randomly selected to receive organic produce for the first 8 years of the kids' lives, then looking at IQ and hyperactivity. But no one's going to do that study because of how expensive 8 years of organic produce would be. The Biodeterminist's Guide doesn't mention PCBs in the section on fish, but fish (particularly farmed salmon) are a major source of these pollutants. They don't seem to be as bad as mercury, but are neurotoxic. Unfortunately their half-life in the body is around 14 years, so if you have even a vague idea of getting pregnant ever in your life you shouldn't be eating much farmed salmon (or Atlantic/farmed salmon, bluefish, wild striped bass, white and Atlantic croaker, blackback or winter flounder, summer flounder, or blue crab).

I had the best intentions of eating lots of the right kind of high-omega-3, low-pollutant fish during and after pregnancy. Unfortunately, fish was the only food I developed an aversion to. Now that Lily is eating food on her own, we tried several sources of omega-3 and found that kippered herring was the only success. Lesson: it's hard to predict what foods kids will eat, so keep trying.
Postscript, 2016: Based on this review, we've been giving her a fish-oil supplement which she *loves (*"More fishy pill!")

In terms of hassle, I underestimated how long I would be "eating for two" in the sense that anything I put in my body ends up in my child's body. Counting pre-pregnancy (because mercury has a half-life of around 50 days in the body, so sushi you eat before getting pregnant could still affect your child), pregnancy, breastfeeding, and presuming a second pregnancy, I'll probably spend about 5 solid years feeding another person via my body, sometimes two children at once. That's a long time in which you have to consider the effect of every medication, every cup of coffee, every glass of wine on your child. There are hardly any medications considered completely safe during pregnancy and lactation—most things are in Category C, meaning there's some evidence from animal trials that they may be bad for human children.

Fluoride

Too much fluoride is bad for children's brains. The CDC recently recommended lowering fluoride levels in municipal water (though apparently because of concerns about tooth discoloration more than neurotoxicity). Around the same time, the American Dental Association began recommending the use of fluoride toothpaste as soon as babies have teeth, rather than waiting until they can rinse and spit.

Cavities are actually a serious problem even in baby teeth, because of the pain and possible infection they cause children. Pulling them messes up the alignment of adult teeth. Drilling on children too young to hold still requires full anesthesia, which is dangerous itself.

But Lily isn't particularly at risk for cavities. 20% of children get a cavity by age six, and they are disproportionately poor, African-American, and particularly Mexican-American children (presumably because of different diet and less ability to afford dentists). 75% of cavities in children under 5 occur in 8% of the population.

We decided to have Lily brush without toothpaste, avoid juice and other sugary drinks, and see the dentist regularly. We also use a $20 water filter that removes fluoride (we verified with lab tests; I recommend the Maine state lab if you need this kind of thing). Fluoride basically doesn't pass into breastmilk, but I used it while I was pregnant and will use it when the kids start drinking water instead of mostly milk.

Home pesticides

One of the most commonly applied insecticides makes kids less smart. This isn't too surprising, given that it kills insects by disabling their nervous system. But it's not something you can observe on a small scale, so it's not surprising that the exterminator I talked to brushed off my questions with "I've never heard of a problem!"

If you get carpenter ants in your house, you basically have to choose between poisoning them or letting them structurally damage the house. We've only seen a few so far, but if the problem progresses, we plan to:

1) remove any rotting wood in the yard where they could be nesting

2) have the perimeter of the building sprayed

3) place gel bait in areas kids can't access

4) only then spray poison inside the house.

If we have mice we'll plan to use mechanical traps rather than poison.

Flame retardants

Since the 1970s, California required a high degree of flame-resistance from furniture. This basically meant that US manufacturers sprayed flame retardant chemicals on anything made of polyurethane foam, such as sofas, rug pads, nursing pillows, and baby mattresses.

The law recently changed, due to growing acknowledgement that the carcinogenic and neurotoxic chemicals were more dangerous than the fires they were supposed to

be preventing. Even firefighters opposed the use of the flame retardants, because when people die in fires it's usually from smoke inhalation rather than burns, and firefighters don't want to breathe the smoke from your toxic sofa (which will eventually catch fire even with the flame retardants).

We've opted to use furniture from companies that have stopped using flame retardants (like Ikea and others listed here). Apparently futons are okay if they're stuffed with cotton rather than foam. We also have some pre-1970s furniture that tested clean for flame retardants. You can get foam samples tested for free.

The main vehicle for children ingesting the flame retardants is that it settles into dust on the floor, and children crawl around in the dust. If you don't want to get rid of your furniture, frequent damp-mopping would probably help.

The standards for mattresses are so stringent that the chemical sprays aren't generally used, and instead most mattresses are wrapped in a flame-resistant barrier which apparently isn't toxic. I contacted the companies that made our mattresses, and they're fine.

Ratings for chemical safety of children's car seats here.

Thoughts on IQ

A lot of people, when I start talking like this, say things like "Well, I lived in a house with lead paint/played with mercury/etc. and I'm still alive." And yes, I played with mercury as a child, and Jeff is still one of the smartest people I know even after getting acute lead poisoning as a child.

But I do wonder if my mind would work a little better without the mercury exposure, and if Jeff would have had an easier time in school without the hyperactivity (a symptom of lead exposure). Given the choice between a brain that works a little better and one that works a little worse, who wouldn't choose the one that works better?

We'll never know how an individual's nervous system might have been different with a different childhood. But we can see population-level effects. The Environmental Protection Agency, for example, is fine with calculating the expected benefit of making coal plants stop releasing mercury by looking at the expected gains in terms of children's IQ and increased earnings.

Scott writes:

> *A 15 to 20 point rise in IQ, which is a little more than you get from supplementing iodine in an iodine-deficient region, is associated with half the chance of living in poverty, going to prison, or being on welfare, and with only one-fifth the chance of dropping out of high-school ("associated with" does not mean "causes").*

Salkever concludes that for each lost IQ point, males experience a 1.93% decrease in lifetime earnings and females experience a 3.23% decrease. If Lily would earn about what I do, saving her one IQ point would save her $1600 a year or $64000 over her career. (And that's not counting the other benefits she and others will reap from her having a better-functioning mind!) I use that for perspective when making decisions. $64000 would buy a lot of the posh prenatal vitamins that actually contain iodine, or organic food, or alternate housing while we're fixing up the new house.

Conclusion

There are times when Jeff and I prioritize social relationships over protecting Lily from everything that might harm her physical development. It's awkward to refuse to go to someone's house because of the chemicals they use, or to refuse to eat food we're offered. Social interactions are good for children's development, and we value those as well as physical safety. And there are times when I've had to stop being so careful because I was getting paralyzed by anxiety (literally perched in the rocker with the baby trying not to touch anything after my in-laws scraped lead paint off the outside of the house).

But we also prioritize neurological development more than most parents, and we hope that will have good outcomes for Lily.

Wear a Helmet While Driving a Car

A 2006 study showed that “280,000 people in the U.S. receive a motor vehicle induced traumatic brain injury every year” so you would think that wearing a helmet while driving would be commonplace. Race car drivers wear helmets. But since almost no one wears a helmet while driving a regular car, you probably fear that if you wore one you would look silly, attract the notice of the police for driving while weird, or the attention of another driver who took your safety attire as a challenge. (Car drivers are more likely to hit bicyclists who wear helmets.)

The $30+shipping Crasche hat is designed for people who should wear a helmet but don’t. It looks like a ski cap, but contains concealed lightweight protective material. People who have signed up for cryonics, such as myself, would get an especially high expected benefit from using a driving helmet because we very much want our brains to “survive” even a “fatal” crash. I have been using a Crasche hat for about a week.

The Unfriendly Superintelligence next door

Markets are powerful decentralized optimization engines - it is known. Liberals see the free market as a kind of optimizer run amuck, a dangerous superintelligence with simple non-human values that must be checked and constrained by the government - the friendly SI. Conservatives just reverse the narrative roles.

In some domains, where the incentive structure aligns with human values, the market works well. In our current framework, the market works best for producing gadgets. It does not work so well for pricing intangible information, and most specifically it is broken when it comes to health.

We treat health as just another gadget problem: something to be solved by pills. Health is really a problem of *knowledge*; it is a computational prediction problem. Drugs are useful only to the extent that you can package the results of new knowledge into a pill and patent it. If you can't patent it, you can't profit from it.

So the market is constrained to solve human health by coming up with new patentable designs for mass-producible physical objects which go into human bodies. Why did we add that constraint - thou should solve health, but thou shalt only use pills? (Ok technically the solutions don't have to be ingestible, but that's a detail.)

The gadget model works for gadgets because we know how gadgets work - we built them, after all. The central problem with health is that we do not completely understand how the human body works - we did not build it. Thus we should be using the market to figure out how the body works - completely - and arguably we should be allocating trillions of dollars towards that problem.

The market optimizer analogy runs deeper when we consider the complexity of instilling values into a market. Lawmakers cannot program the market with goals directly, so instead they attempt to engineer desireable behavior by ever more layers and layers of constraints. Lawmakers are deontologists.

As an example, consider the regulations on drug advertising. Big pharma is unsafe - its profit function does not encode anything like "maximize human health and happiness" (which of course itself is an oversimplification). If allowed to its own devices, there are strong incentives to sell subtly addictive drugs, to create elaborate

hyped false advertising campaigns, etc. Thus all the deontological injunctions. I take that as a strong indicator of a poor solution - a value alignment failure.

What would healthcare look like in a world where we solved the alignment problem?

To solve the alignment problem, the market's profit function must encode long term human health and happiness. This really is a mechanism design problem - its not something lawmakers are even remotely trained or qualified for. A full solution is naturally beyond the scope of a little blog post, but I will sketch out the general idea.

To encode health into a market utility function, first we create financial contracts with an expected value which captures long-term health. We can accomplish this with a long-term contract that generates positive cash flow when a human is healthy, and negative when unhealthy - basically an insurance contract. There is naturally much complexity in getting those contracts right, so that they measure what we really want. But assuming that is accomplished, the next step is pretty simple - we allow those contracts to trade freely on an open market.

There are some interesting failure modes and considerations that are mostly beyond scope but worth briefly mentioning. This system probably needs to be asymmetric. The transfers on poor health outcomes should partially go to cover medical payments, but it may be best to have a portion of the wealth simply go to nobody/everybody - just destroyed.

In this new framework, designing and patenting new drugs can still be profitable, but it is now put on even footing with preventive medicine. More importantly, the market can now actually allocate the correct resources towards long term research.

To make all this concrete, let's use an example of a trillion dollar health question - one that our current system is especially ill-posed to solve:

What are the long-term health effects of abnormally low levels of solar radiation? What levels of sun exposure are ideal for human health?

This is a big important question, and you've probably read some of the hoopla and debate about vitamin D. I'm going to soon briefly summarize a general abstract theory, one that I would bet heavily on if we lived in a more rational world where such bets were possible.

In a sane world where health is solved by a proper computational market, I could make enormous - ridiculous really - amounts of money if I happened to be an early researcher who discovered the full health effects of sunlight. I would bet on my theory simply by buying up contracts for individuals/demographics who had the most health to gain by correcting their sunlight deficiency. I would then publicize the theory and evidence, and perhaps even raise a heap pile of money to create a strong marketing engine to help ensure that my investments - my patients - were taking the necessary actions to correct their sunlight deficiency. Naturally I would use complex machine learning models to guide the trading strategy.

Now, just as an example, here is the brief 'pitch' for sunlight.

If we go back and look across all of time, there is a mountain of evidence which more or less screams - proper sunlight is important to health. Heliotherapy has a long history.

Humans, like most mammals, and most other earth organisms in general, evolved under the sun. A priori we should expect that organisms will have some 'genetic programs' which take approximate measures of incident sunlight as an input. The serotonin -> melatonin mediated blue-light pathway is an example of one such light detecting circuit which is useful for regulating the 24 hour circadian rhythm.

The vitamin D pathway has existed since the time of algae such as the Coccolithophore. It is a multi-stage pathway that can measure solar radiation over a range of temporal frequencies. It starts with synthesis of fat soluble cholecalciferiol which has a very long half life measured in months. [1] [2]

The rough pathway is:

- Cholecalciferiol (HL ~ months) becomes
- 25(OH)D (HL ~ 15 days) which finally becomes
- $1,25(OH)_2$ D (HL ~ 15 hours)

The main recognized role for this pathway in regards to human health - at least according to the current Wikipedia entry - is to enhance "the internal absorption of calcium, iron, magnesium, phosphate, and zinc". Ponder that for a moment.

Interestingly, this pathway still works as a general solar clock and radiation detector for carnivores - as they can simply eat the precomputed measurement in their diet.

So, what is a long term sunlight detector useful for? One potential application could be deciding appropriate resource allocation towards DNA repair. Every time an organism is in the sun it is accumulating potentially catastrophic DNA damage that must be repaired when the cell next divides. We should expect that genetic programs would allocate resources to DNA repair and various related activities dependent upon estimates of solar radiation.

I should point out - just in case it isn't obvious - that this general idea does not imply that cranking up the sunlight hormone to insane levels will lead to much better DNA/cellular repair. There are always tradeoffs, etc.

One other obvious use of a long term sunlight detector is to regulate general strategic metabolic decisions that depend on the seasonal clock - especially for organisms living far from the equator. During the summer when food is plentiful, the body can expect easy calories. As winter approaches calories become scarce and frugal strategies are expected.

So first off we'd expect to see a huge range of complex effects showing up as correlations between low vit D levels and various illnesses, and specifically illnesses connected to DNA damage (such as cancer) and or BMI.

Now it turns out that BMI *itself* is also strongly correlated with a huge range of health issues. So the first key question to focus on is the relationship between vit D and BMI. And - perhaps not surprisingly - there is pretty good evidence for such a correlation [3] [4] , and this has been known for a while.

Now we get into the real debate. Numerous vit D supplement intervention studies have now been run, and the results are *controversial*. In general the vit D experts (such as my father, who started the vit D council, and publishes some related research[5]) say that the only studies that matter are those that supplement at high doses sufficient to elevate vit D levels into a 'proper' range which *substitutes for sunlight*, which in general requires 5000 IU day on average - depending completely on genetics and lifestyle (to the point that any one-size-fits all recommendation is probably terrible).

The mainstream basically ignores all that and funds studies at tiny RDA doses - say 400 IU or less - and then they do meta-analysis over those studies and conclude that their big meta-analysis, unsurprisingly, doesn't show a statistically significant effect. However, these studies still show small effects. Often the meta-analysis is corrected for BMI, which of course also tends to remove any vit D effect, to the extent that low vit D/sunlight is a cause of both weight gain and a bunch of other stuff.

So let's look at two studies for vit D and weight loss.

First, this recent 2015 study of 400 overweight Italians (sorry the actual paper doesn't appear to be available yet) tested vit D supplementation for weight loss. The 3 groups were (0 IU/day, ~1,000 IU / day, ~3,000 IU/day). The observed average weight loss was (1 kg, 3.8 kg, 5.4 kg). I don't know if the 0 IU group received a placebo. Regardless, it looks promising.

On the other hand, this 2013 meta-analysis of 9 studies with 1651 adults total (mainly women) supposedly found no significant weight loss effect for vit D. However, the studies used between 200 IU/day to 1,100 IU/day, with most between 200 to 400 IU. Five studies used calcium, five also showed weight loss (not necessarily the same - unclear). This does not show - at all - what the study claims in its abstract.

In general, medical researchers should not be doing statistics. That is a job for the tech industry.

Now the vit D and sunlight issue is complex, and it will take much research to really work out all of what is going on. The current medical system does not appear to be handling this well - why? Because there is insufficient financial motivation.

Is Big Pharma interested in the sunlight/vit D question? Well *yes* - but only to the extent that they can create a patentable analogue! The various vit D analogue drugs developed or in development is evidence that Big Pharma is at least paying attention. But assuming that the sunlight hypothesis is mainly correct, there is very little profit in actually *fixing* the *real problem*.

There is probably more to sunlight that just vit D and serotonin/melatonin. Consider the interesting correlation between birth month and a number of disease conditions[6].

Perhaps there is a little grain of truth to astrology after all.

Thus concludes my little vit D pitch.

In a more sane world I would have already bet on the general theory. In a really sane world it would have been solved well before I would expect to make any profitable trade. In that rational world you could actually trust health advertising, because you'd know that health advertisers are strongly financially motivated to convince you of things actually truly important for your health.

Instead of charging by the hour or per treatment, like a mechanic, doctors and healthcare companies should literally invest in their patients long-term health, and profit from improvements to long term outcomes. The sunlight health connection is a trillion dollar question in terms of medical value, but not in terms of exploitable profits in today's reality. In a properly constructed market, there would be enormous resources allocated to answer these questions, flowing into legions of profit motivated startups that could generate billions trading on computational health financial markets, all without selling any gadgets.

So in conclusion: the market could solve health, but only if we allowed it to and only if we setup appropriate financial mechanisms to encode the correct value function. This is the UFAI problem next door.

Don't steer with guilt

I've spoken at length about shifting guilt or dispelling guilt. What I haven't talked about, yet, is guilt itself.

So let's talk about guilt.

Guilt is one of those strange tools that works by *not* occurring. You place guilt on the branches of possibility that you don't want to happen, and then, if all goes well, those futures don't occur. Guilt is supposed to steer the future towards non-guilty futures; it's never supposed to be instantiated in reality.

Guilt works by the same mechanism as threats: imagine the tribesperson who precommits to breaking the legs of anyone who steals their food. If this precommitment works, then it never needs to be carried out: violence is a dangerous business, and the tribesperson would much rather that they never need to break legs at all. The threat is something that the tribesperson places on possibilities that they disprefer, in attempts to ensure that they never come to be.

Imagine, by contrast, the tribesperson who threatens to breaking the legs of anyone who looks at them funny: they might find themselves attempting violence every single day, and this likely makes their life unpleasant, to say the least. In this case, I would argue that they're using their threats poorly. I would say that, if you keep finding yourself carrying out a threat, then you really need to consider whether or not your threats are really capable of steering the future in the way you hoped.

Guilt is the same way: *if you find yourself regularly experiencing guilt, then you're using guilt incorrectly.*

Guilt works only when you wield it in such a way that it *doesn't happen.*

Guilt is costly when deployed. Once activated, it's usually strongly demotivating, and can easily lead to failure spirals or vicious cycles of depression.

As far as I can tell, the way that guilt-motivated people tend to operate is by working fervently in attempts to avoid the scourge of guilt. This may be effective when it works, but as soon as it starts to fail, the failure often cascades into a full-blown failure spiral (you're guilty that you're not working, which makes you feel bad, which makes it hard to work, which makes you guiltier, which you feel worse, which makes it harder to work, …). As a result, guilt motivation often results in a boom/bust productivity/depression cycle that, as far as I can tell, results in people feeling quite bad about themselves and being much less effective than they would be if they could maintain a steady pace.

Some might argue that the boom is worth the bust, that the productivity is worth the depression. This seems straight up false to me (and I have some relevant experience): the frantic productivity fueled by fear of guilt doesn't seem more effective (and often seems *less* effective) than intrinsically motivated productivity, and that's *before* we count the losses from periodic failure spirals. As far as I can tell, intrinsic motivation is just straight up more effective.

(This is something you have to accept before I can help you remove your guilt: it's much harder to remove guilt if you don't want to.)

Guilt is very costly when activated, so if it's getting activated regularly, then you're placing it on the wrong branches of possibility.

You might protest, "but then what do I *do* in the unsatisfying branches of reality? I need to find *some* way to prevent me from chasing short-term satisfaction at the expense of long-term benefits." If you regularly finding yourself binging netflix TV shows, and you would rather not find yourself regularly binging netflix TV shows, then shouldn't you feel guilty whenever you do?

No! If the situation occurs regularly, then guilt is not the tool to use! You're welcome to feel guilty if you ever kidnap a baby or punch a homeless person, and you can tell that the guilt is working in those cases because you *never do those things*. But if you repeatedly find yourself in a situation that you disprefer, then guilt is just not the tool to use. That's not where it's useful.

If you want to figure out how to avoid a certain recurring situation, then there's a different tool that *is* appropriate, that's much more effective at figuring out how to steer the future towards better places: Science!

When you find yourself binging netflix, don't heap loads of guilt on yourself post-binge. That sort of thing clearly doesn't prevent the binge. Instead, say to yourself, "huh, I appear to netflix-binge under certain conditions, despite the fact that I'd rather not. I wonder which conditions, specifically, led to that binge! What were the triggers? How could they have been avoided? What methods might help me avoid binging in the future?"

And then treat it like an experiment! Write up your hypotheses. Experiment with many different ways to fix your glitches. Write postmortems when you fail. If you attempt a fix and then find yourself binging *again*, then don't heap loads of guilt on yourself! *That still doesn't help.* Instead, say "Aha! So *that* attempted fix didn't work. I wonder if I can figure out why?" Cross a hypothesis or two off your list. Refine your models. Expand your hypothesis space. Gather more data.

Do science to it.

Don't bemoan individual failures. That's finite-task thinking. Instead, acknowledge that there's an unlimited number of changes you'd like to make to your behavior, and that some of them are more important than others, and that some of them are more costly than others, and that they all take time to fix. See the infinite stream of self-improvement that lies before you, add it to all the other streams you're optimizing, and then simply navigate the streams as quickly as you are able.

Don't feel terrible whenever you do something you wish you hadn't! That is a poor mechanism by which to steer the future. Instead, when you do something you wish you hadn't, identify the *pattern of behavior* that led to this, and add addressing *that* to your todo list. Then weigh the time you're losing against the time it would take to change the pattern, and weigh that against the other priorities that are vying for your attention, and then do what needs doing.

Sometimes you'll ignore a pattern of failure. Maybe the failures are relatively cheap and the pattern is hard to change, and fixing the pattern simply isn't worth your attention. In this case, when the failure occurs, there is no need to feel guilty: the failures are the price you pay for time spent not fixing them. You can't simply teleport to a new pattern of behavior, and so if you lack the time to change the pattern, then

the occasional failure is a fair price. Trust yourself to fix the pattern if the costs ever get too high, trust yourself to understand that investing in yourself is important, and if fixing the pattern *still* isn't at the top of your todo list, then don't worry about the individual failures. You have bigger things on your plate.

Other times, you'll decide that the pattern needs changing. Five minutes per day is thirty hours per year, and investing in yourself pays dividends. In this case, treat addressing the pattern of failure like a science project. Every new individual failure is data point about what doesn't work. Every avoided failure is a data point about what does. Heaping guilt on yourself whenever you hit a new failure would be nonsense — fixing the *pattern* is a science experiment, and individual successes or failures are your data points.

Most people use their individual failures as a signal to themselves that it's time to feel terrible. It is much more effective, I think, to use your individual failures as a chance to update your tactics.

This, in my experience, is the head-on cure for guilt: Don't treat the individual failures like a burden; treat changing the pattern like a science experiment.

Shifting guilt

This is a linkpost for https://mindingourway.com/shifting-guilt/

The posts so far have been less about confronting guilt, and more about different tools for shifting it. This is a valuable skill to generalize.

The posts in this series have developed three such tools for shifting guilt. In this post, I'll recast those three tools as members of the same family, so that you can start to see the pattern, and develop similar tools from the same family as you need them.

The tools that I have described so far shift guilt to one particular place: guilt about being unable to act as you desire. This is intentional — that is the one place that I know how to confront guilt head-on.

The first tool for shifting guilt is the tool of *refinement*. This tool is used on listless guilts in need of pointing.

Imagine finding yourself feeling vaguely guilty the morning after a party, having slept in longer than you intended, your head aching from a slight hangover. Imagine a vague guilt making your body feel heavier. Perhaps it whispers that the night was senseless. Perhaps it murmurs that you're wasting your life away. This is the sort of guilt that's amenable to refinement: ask the guilt what, precisely, it would have had you do instead of what you did. (It is important, when refining, to also possess the virtue of concreteness: do not settle for "I should have been studying." Demand a specific action: Which book? Which chapter?)

Sometimes, when asking the guilt what you could have done instead, you will remember that none of the alternatives were compelling. Maybe the party was for an old friend who you only see once every few years, and fulfilling the social obligation was better than the alternative. Maybe you were exhausted from a day of studying, low on human contact, and needed the party to reinvigorate you. When using the tool of refinement, the guilt sometimes simply disappears.

But often, the guilt gets more pointed. Perhaps you conclude you should have been working overtime so you could donate the money to a worthy cause. Perhaps you conclude that you had the opportunity and the stamina, but simply not the willpower. This is good! This is a success! The refinement has succeeded, and the guilt has come into more focus.

But more often than not, when you succeed at refining guilt, you find yourself left with an obligation ("I should have drank less" or "I should have studied" or "I should have worked overtime.") This has not yet shifted enough to be confronted. For obligations, you need the second tool.

The second tool for shifting guilt is the tool of *internalization*. This tool is used on guilts that stem from neglected obligations.

I strongly recommend that you staunchly refuse to bow to any guilt forced on you from the outside. You say you "should have" studied more, instead of going to the party? Says who? Cash out the should. Again, it is critically important have virtue of concreteness when cashing out a should: do not say "it would have been better for me to study more;" for this has not removed the should, it has simply hidden it inside the

word "better." The way to cash out a should (and, thus, the way to use the tool of internalization) is to ask yourself whether or not it would be OK to drop the obligation entirely.

What would happen if you decide to never study that textbook again? Is it a relief? If so, then drop the obligation, and relinquish the guilt. You probably just accidentally confused someone's quality line with your preference curve. Sometimes, when attempting internalization, the guilt simply disappears. (Other times, part of the guilt disappears, and your find yourself again facing a vague, unfocused guilt. This is fine, and indeed quite normal — just apply the tool of refinement again, and repeat.)

But more often than not, when you threaten to drop an obligation entirely, some part of you protests. Imagine you're feeling terrible for failing to work overtime and donate money. If you ask yourself "what if I just never donate money to those worse off ever again?," then most likely, some part of you will protest "but that would be bad!"

This is good. It means you have your *own* reasons for wanting to donate, which means you can drop the external obligation and do it because *you* want to.

Why would it be bad to stop donating? Don't settle for answers like "because then I would be a bad person" — that's replacing one obligation with another. If you get an answer like that, ask yourself, "why would it be bad for me to be a bad person?" Remember that concreteness is a virtue. Don't settle for an externalized answer (such as "because then people wouldn't like me"); push on until you get an internalized answer ("because I prefer worlds where _____").

Keep in mind that there may be many different parts to the answer: if you use the tool of internalization and get an answer that feels uncompelling, such as "because I prefer worlds where my friends think that I am generous," then ask yourself something like "OK, let's say that my friends were guaranteed to think that I am generous regardless of how much I donate to people worse off than me, *then* is it OK for me to never donate to people worse off than me ever again?" — You can keep doing this until you uncover all the reasons behind your desires.

This is how the tool of internalization shifts guilt: it forces the guilt to either resolve itself, or reveal itself to you in terms of your own desires. It shifts the guilt to a place where the thing the guilt demands are things you want for yourself, rather than things you want because you think you should.

So perhaps now you feel guilty for not working overtime to earn money to give to those less well-off than yourself (which is something you desire due to a deep dissatisfaction with the unfairnesses of the modern world). This, again, is progress: the guilt is now focused and internalized. This is exactly the sort of guilt that the third tool addresses.

The third tool for shifting guilt is the filter of *realism.* Look at your guilt, and ask it whether its demands are realistic.

Ask whether you really could have worked harder and done something else, while remembering that you are in fact mortal. You are no more able to work 20 hour days at peak capacity than you are able to cure Alzheimer's disease with a snap of your fingers. Look not to whether you were moving as fast as you physically could. Instead, look to the streams you need to move through in order to achieve your goals while remembering that two of the most important streams are maintaining health and motivation.

Do not ask, "could I have skipped the party and worked more?" Ask, instead, "am I traversing the work streams at the fastest sustainable pace?" Check whether the task the guilt demands is realistic. Remember that working yourself ragged is not a virtue. When keeping the filter of realism in mind, many guilt simply fail to materialize in the first place.

But some guilts do pass the filter of realism, and leave you lamenting a flaw in your process, an inability to do what you think is best. Perhaps you will notice that you attend parties far more often than you prefer, due to peer pressure. Perhaps you will notice that you actually find parties draining, and that you were only attending this one in hopes of finding a date, which you could have done in a less costly manner if you were really trying. Perhaps you will realize that you've been adrift, that you've lost focus, and you'll feel guilty for failing to maintain your drive.

And this is right where we want the guilt. If you must feel guilty, I recommend feeling guilty not about what you did or didn't do, but about the *pattern of behavior* that corresponds to acting against your will. Don't feel guilty for going to *this* party, feel guilty for the general pattern of giving into peer pressure, or misjudging how much fun you'll have, or overindulging. Because *this* is the sort of guilt that I know how to address head-on.

The three tools of refinement, internalization, and realism, are, in my case, effectively universal: I can use them to shift any hint of guilt up to specific, internalized guilt about a realistic concern at the process level. I am sure, though, that for many of you, there will be other forms of guilt that these three tools do not cover.

This is why I make the tools explicit here: so that you can see how they work and see what they share, and then construct your own variants that work on whatever other guilts you tend to encounter.

As you hone those tools, I recommend you seek a similar endpoint: shift the guilt away from the misstep and onto the systemic flaw in your footwork. Shift guilt from the instance to the pattern. Bring your guilt to this battleground, and I will show you how to defeat it.

Update from the suckerpunch

This is a linkpost for https://mindingourway.com/update-from-the-suckerpunch/

The most common objection I hear when helping people remove their guilt is something along the lines of "Hey wait! I was using that!"

Believing this (or really any variant of "but guilt is good for me!") makes it fairly hard to replace guilt with something more productive.

I've met some people who complain that if they didn't have guilt then they'd do horrible things. I think this is fairly unlikely, and I file it right next to the arguments that say that if they didn't believe in God then they'd do horrible things. Even after dropping your obligations, you will still have something to fight for. Your *reasons* for not doing things you'd rather not do will remain even after the guilt is replaced.

Others I have met protest that guilt is useful in order to ensure that they won't repeat their failures. Without guilt, how would they learn their lesson? To which I generally say, that's fine, but if it keeps happening then you aren't learning, and it's time to use a different tool instead.

That said, there *are* lessons that need learning, and there *is* something sort of like 'guilt' that can help you learn them.

But you can use it even while completely replacing your guilt motivation.

Once upon a time, I had a loose date planned with a girlfriend. She was going to drop by around 21:00 to hang out. I had something else planned at 19:00 that I didn't expect to take too long; it ended up taking many hours longer than expected. There was no particularly convenient point along the way to step out and call my girlfriend and tell her I'd be late… so I didn't. I simply got home at 23:00 at night, opened the door, and saw my girlfriend sitting worried on the bed.

There's a very distinct type of feeling that I experienced, there, which you might call "guilt." Seeing her sitting there on the bed, I suddenly remembered that the anxiety and dejection that she went through was far worse than the slight awkwardness I would have incurred to call her. A compartmentalization in my head broke down, and the part of me that had *known* she'd been feeling terrible suddenly came into mental focus. My error became obvious. The feeling was something like being punched in the gut.

Afterwards, I *also* had the opportunity to feel a lingering sense of regret for days.

When I suggest removing guilt, I suggest removing the latter — but not the former. The former is quite useful.

If you worry that, by removing guilt, you will lose your ability to update when you mess up, then I say: update on the suckerpunch. Trust me, it's strong enough. Update *immediately* when you realize where you failed, and use the terrible feeling to make sure you *don't do that again.*

Update fully on the suckerpunch, and there will be no need for that lingering regret. Skip to the end, immediately; update as far as you can, the moment that you realize

your error. Moping for days doesn't make things better. Updating your behavior does.

There are those who still protest that the lingering regret is useful: if you hurt your friend, you may think that they need to see you spending days filled with regret, or otherwise they will think less of you. You may think that others find it disconcerting to see you update immediately and continue without missing a beat. Some people want to see penance done.

If that is your protest, then I have little to offer you. I can only note that I have seen many groups of friends form a tacit pact of non-excellence, where each individual in the group is reluctant to outperform the others, in fear that high performance will be punished with ostracization. Many have condemned themselves to a life of dissatisfaction thanks to a non-excellence pact. I say: better to inspire your friends than validate their mediocrity.

It can give some people whiplash, to see you update quickly, but I much prefer friends and lovers that encourage skipping to the end rather than those who feel a need to extract their pound of flesh whenever you err. For me, the social cost of updating quickly is well worth the ability to move faster. Your experience, of course, may differ.

Just remember that you won't be able to replace guilt-based motivation before giving yourself permission to do so. For so long as you view your guilt as an aid rather than a burden, for so long as you view it as right and necessary, I cannot help you remove it.

But I can tell you this:

Almost all emotions, I have found a place for. I have long looked upon Spock and Jedi with some dissatisfaction: I am not one to advocate suppressing emotion. Anger has its place and time, as does joy, as does sadness. Awe and fear and cold resolve, I have found a use for.

I have even found a use for that suckerpunch that occurs when you learn you have made a mistake, that you might label 'guilt.'

But the lingering, drawn-out guilt, the persistent regret that drives one to work in fear of it?

I have never once found a use for that.

Be a new homunculus

This is a linkpost for https://mindingourway.com/be-a-new-homunculus/

Here's a mental technique that I find useful for addressing many dour feelings, guilt among them:

When you're feeling guilty, it is sometimes helpful to close your eyes for a moment, re-open them, and pretend that you're a new homunculus.

A "homunculus" is a tiny representation of a human, and one classic fallacy when reasoning about how brains work is the homunculus fallacy, in which people imagine that "they" are a little homonculus inside their head looking at an image generated by their eyes.

It's an easy fiction to buy into, that you're a little person in your head that can move your hands and shape your mouth and that decides where to steer the body and so on. There is, of course, no homunculus inside your head (for if *you* are steered by a homunculus, then how is the homunculus steered?), but it can be quite fun to pretend that you are a homunculus sometimes, mostly because this allows you to occasionally pretend you're a *new* homunculus, fresh off the factory lines, and newly installed into this particular person.

Close your eyes, and pretend you're arriving in this body for the very first time. Open them and do some original seeing on this person you now are. Rub your hands together, look around, and take stock of your surroundings. Do some internal checks to figure out what this body values, to figure out what it is you're fighting for. Check the catalog of plans and upcoming actions. Check the backlog of memories and obligations.

There will probably be some housecleaning to do: homunculi are known to get a little careless as they age, and the old homunculus that you replaced probably let a bunch of useless tasks accumulate without realizing it. As a new homunculus you have the privilege of pruning the things that obviously need pruning. Maybe you'll look and say "Ah, yes, we're going to cancel lunch with *that* person; this body was secretly dreading it. I also see that this body is currently spending a lot of cycles feeling guilty about a date that went poorly last week; we can dismiss that, it's no longer useful for *this* homunculus. And also, "exercise" doesn't seem to be on today's schedule at all! How strange. This body definitely intended to exercise today; somehow it fell off the list. I'll put it back on."

It can be quite liberating to be a new homunculus, without any obligation to propagate the errors of the old one.

This is, in fact, a common technique for dealing with the sunk cost fallacy (also known as the "pretend you're a teleporting alien that just teleported into your body" technique). This is useful for avoiding sunk costs because the *new* homunculus has no reason to honor the old homunculus' sunk costs.

Say the old homunculus bought plane tickets which would let you travel to Texas tomorrow (and return in a week), and that the ticket is non-refundable. The old homunculus may well have an attachment to the "go to Texas" plan, and may try to

convince themselves to go even when it becomes clear that the trip won't be worth the time. The new homunculus, however, has no such loyalty to the sunk costs: *it* can just evaluate whether or not to go on the trip regardless of how much the tickets costed.

This is also a technique that works quite well for managing guilt: it's often easy for the new homunculus to recognize lingering guilt as a bodily response marking malcontent about something that was done in the past, by the old homunculus. The best action for the new homunculus to take, usually, is to check what regretted action caused the guilt, check what pattern of behavior led to the regretted action, mark down a note about which cognitive pattern needs to be reprogrammed, and then dismiss the guilt (which has now served its purpose).

As a matter of fact, guilt and sunk cost fallacy are closely related: both are about suffering for costs that were paid in the past. The only difference is that guilt carries with it a lesson, an instruction to alter your environment and your mind so that similar actions don't occur in the future. With practice, it is possible to *reflexively* treat the initial gut-wrenching guilt as an instruction to update your behavioral patterns, and then dismiss the lingering guilt immediately. (Cognitive patterns, after all, take some time to train.)

In the interim I suggest pretending you're a new homunculus. If you start to feel guilt, then close your eyes and re-open them as a brand new homunculus. Notice the guilt, listen to the message it bears, and *actually write down* the behavioral pattern that you wish to change. Then spend five minutes (a full five minutes, by the clock) brainstorming ways that you might change the pattern and start retraining your mind. Then thank the guilt for carrying you this message, and dismiss it.

Eventually, this can become reflexive. Until then, I suggest occasionally becoming a new homunculus. In fact, I often use something like this myself, even though I've been immune to guilt for quite some time: it's a great way to see the world and yourself with fresh eyes, and that can be invaluable.