Best of LessWrong: June 2015

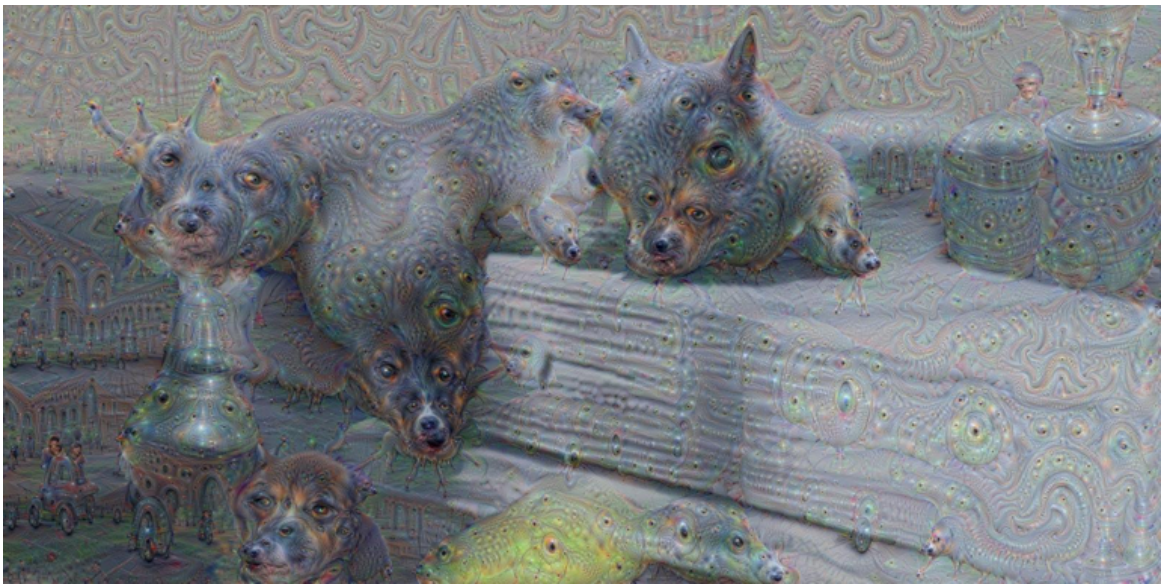
1. The Brain as a Universal Learning Machine
2. Pattern-botching: when you forget you understand
3. Taking the reins at MIRI
4. Top 9+2 myths about AI risk
5. Your "shoulds" are not a duty
6. Working yourself ragged is not a virtue
7. Rest in motion

Best of LessWrong: June 2015

1. The Brain as a Universal Learning Machine
2. Pattern-botching: when you forget you understand
3. Taking the reins at MIRI
4. Top 9+2 myths about AI risk
5. Your "shoulds" are not a duty
6. Working yourself ragged is not a virtue
7. Rest in motion

The Brain as a Universal Learning Machine

This article presents an emerging architectural hypothesis of the brain as a biological implementation of a *Universal Learning Machine*. I present a rough but complete architectural view of how the brain works under the universal learning hypothesis. I also contrast this new viewpoint - which comes from computational neuroscience and machine learning - with the older evolved modularity hypothesis popular in evolutionary psychology and the heuristics and biases literature. These two conceptions of the brain lead to *very* different predictions for the likely route to AGI, the value of neuroscience, the expected differences between AGI and humans, and thus any consequent safety issues and dependent strategies.



(The image above is from a recent mysterious post to r/machinelearning, probably from a Google project that generates art based on a visualization tool used to inspect the patterns learned by convolutional neural networks. I am especially fond of the wierd figures riding the cart in the lower left.)

Intro: Two Viewpoints on the Mind

Few discoveries are more irritating than those that expose the pedigree of ideas.

-- Lord Acton (probably)

Less Wrong is a site devoted to refining the art of human rationality, where rationality is based on an idealized conceptualization of how minds should or could work. Less Wrong and its founding sequences draws heavily on the heuristics and biases literature in cognitive psychology and related work in evolutionary psychology. More specifically the sequences build upon a specific cluster in the space of cognitive theories, which can be identified in particular with the highly influential "evolved modularity" perspective of Cosmides and Tooby.

From Wikipedia:

> Evolutionary psychologists propose that the mind is made up of genetically influenced and domain-specific[3] mental algorithms or computational modules, designed to solve specific evolutionary problems of the past.[4]

From "Evolutionary Psychology and the Emotions":[5]

> An evolutionary perspective leads one to view the mind as a crowded zoo of evolved, domain-specific programs. Each is functionally specialized for solving a different adaptive problem that arose during hominid evolutionary history, such as face recognition, foraging, mate choice, heart rate regulation, sleep management, or predator vigilance, and each is activated by a different set of cues from the environment.

If you imagine these general theories or perspectives on the brain/mind as points in theory space, the evolved modularity cluster posits that much of the machinery of human mental algorithms is largely innate. General learning - if it exists at all - exists only in specific modules; in most modules learning is relegated to the role of adapting existing algorithms and acquiring data; the impact of the information environment is de-emphasized. In this view the brain is a complex messy cludge of evolved mechanisms.

There is another viewpoint cluster, more popular in computational neuroscience (especially today), that is almost the *exact opposite* of the evolved modularity hypothesis. I will rebrand this viewpoint the "universal learner" hypothesis, aka the "one learning algorithm" hypothesis (the rebranding is justified mainly by the inclusion of some newer theories and evidence for the basal ganglia as a 'CPU' which learns to control the cortex). The roots of the universal learning hypothesis can be traced back to Mountcastle's discovery of the simple uniform architecture of the cortex.[6]

The universal learning hypothesis proposes that *all* significant mental algorithms are learned; nothing is innate except for the learning and reward machinery itself (which is somewhat complicated, involving a number of systems and mechanisms), the initial rough architecture (equivalent to a prior over mindspace), and a small library of simple innate circuits (analogous to the operating system layer in a computer). In this view the mind (software) is distinct from the brain (hardware). The mind is a complex software system built out of a general learning mechanism.

In simplification, the main difference between these viewpoints is the relative quantity of domain specific mental algorithmic information specified in the genome vs that acquired through general purpose learning during the organism's lifetime. Evolved modules vs learned modules.

When you have two hypotheses or viewpoints that are almost complete opposites this is generally a sign that the field is in an early state of knowledge; further experiments

typically are required to resolve the conflict.

It has been about 25 years since Cosmides and Tooby began to popularize the evolved modularity hypothesis. A number of key neuroscience experiments have been performed since then which support the universal learning hypothesis (reviewed later in this article).

Additional indirect support comes from the rapid unexpected success of Deep Learning[7], which is entirely based on building AI systems using simple universal learning algorithms (such as Stochastic Gradient Descent or other various approximate Bayesian methods[8][9][10][11]) scaled up on fast parallel hardware (GPUs). Deep Learning techniques have quickly come to dominate most of the key AI benchmarks including vision[12], speech recognition[13][14], various natural language tasks, and now even ATARI [15] - proving that simple architectures (priors) combined with universal learning is a path (and perhaps the only viable path) to AGI. Moreover, the internal representations that develop in some deep learning systems are structurally and functionally similar to representations in analogous regions of biological cortex[16].

To paraphrase Feynman: to truly understand something you must build it.

In this article I am going to quickly introduce the abstract concept of a universal learning machine, present an overview of the brain's architecture as a specific type of universal learning machine, and finally I will conclude with some speculations on the implications for the race to AGI and AI safety issues in particular.

Universal Learning Machines

A universal learning machine is a simple and yet very powerful and general model for intelligent agents. It is an extension of a general computer - such as Turing Machine - amplified with a universal learning algorithm. Do not view this as my 'big new theory' - it is simply an amalgamation of a set of related proposals by various researchers.

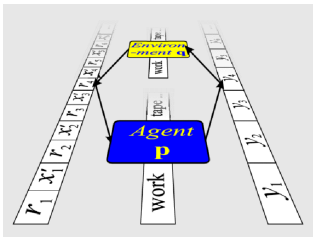
An initial untrained seed ULM can be defined by 1.) a prior over the space of models (or equivalently, programs), 2.) an initial utility function, and 3.) the universal learning machinery/algorithm. The machine is a real-time system that processes an input sensory/observation stream and produces an output motor/action stream to control the external world using a learned internal program that is the result of continuous self-optimization.

There is of course always room to smuggle in arbitrary innate functionality via the prior, but in general the prior is expected to be extremely small in bits in comparison to the learned model.

The key defining characteristic of a ULM is that it uses its universal learning algorithm for continuous recursive self-improvement with regards to the utility function (reward system). We can view this as second (and higher) order optimization: the ULM optimizes the external world (first order), and also optimizes its own internal optimization process (second order), and so on. Without loss of generality, any system capable of computing a large number of decision variables can also compute internal self-modification decisions.

Conceptually the learning machinery computes a probability distribution over program-space that is proportional to the expected utility distribution. At each timestep it receives a new sensory observation and expends some amount of computational energy to infer an updated (approximate) posterior distribution over its internal program-space: an approximate 'Bayesian' self-improvement.

The above description is intentionally vague in the right ways to cover the wide space of possible practical implementations and current uncertainty. You could view AIXI as a particular formalization of the above general principles, although it is also as dumb as a rock in any practical sense and has other potential theoretical problems. Although the general idea is simple enough to convey in the abstract, one should beware of concise formal descriptions: practical ULMs are too complex to reduce to a few lines of math.



A ULM inherits the general property of a Turing Machine that it can compute anything that is computable, given appropriate resources. However a ULM is also more powerful than a TM. A Turing Machine can only do what it is programmed to do. A ULM automatically programs itself.

If you were to open up an infant ULM - a machine with zero experience - you would mainly just see the small initial code for the learning machinery. The vast majority of the codestore starts out empty - initialized to noise. (In the brain the learning machinery is built in at the hardware level for maximal efficiency).

Theoretical turing machines are all qualitatively alike, and are all qualitatively distinct from any non-universal machine. Likewise for ULMs. Theoretically a small ULM is just as general/expressive as a planet-sized ULM. In practice quantitative distinctions do matter, and can become effectively qualitative.

Just as the simplest possible Turing Machine is in fact quite simple, the simplest possible Universal Learning Machine is also probably quite simple. A couple of recent proposals for simple universal learning machines include the Neural Turing Machine[16] (from Google DeepMind), and Memory Networks[17]. The core of both approaches involve training an RNN to learn how to control a memory store through gating operations.

Historical Interlude

At this point you may be skeptical: how could the brain be anything like a universal learner? What about all of the known innate biases/errors in human cognition? I'll get to that soon, but let's start by thinking of a couple of general experiments to test the universal learning hypothesis vs the evolved modularity hypothesis.

In a world where the ULH is mostly correct, what do we expect to be different than in worlds where the EMH is mostly correct?

One type of evidence that would support the ULH is the demonstration of key structures in the brain along with associated wiring such that the brain can be shown to directly implement some version of a ULM architecture.

Another type of indirect evidence that would help discriminate the two theories would be evidence that the brain is capable of general global optimization, and that complex domain specific algorithms/circuits mostly result from this process. If on the other hand the brain is only capable of constrained/local optimization, then most of the complexity must instead be innate - the result of global optimization in evolutionary deeptime. So in essence it boils down to the optimization capability of biological learning vs biological evolution.

From the perspective of the EMH, it is not sufficient to demonstrate that there are things that brains can not learn in practice - because those simply could be quantitative

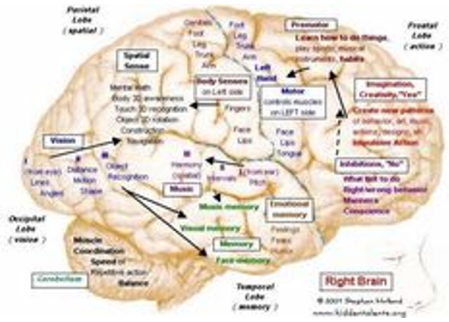
limitations. Demonstrating that an intel 486 can't compute some known computable function in our lifetimes is not proof that the 486 is not a Turing Machine.

Nor is it sufficient to demonstrate that biases exist: a ULM is only 'rational' to the extent that its observational experience and learning machinery allows (and to the extent one has the correct theory of rationality). In fact, the existence of many (most?) biases intrinsically *depends* on the EMH - based on the implicit assumption that some cognitive algorithms are innate. If brains are mostly ULMs then most cognitive biases dissolve, or become learning biases - for if all cognitive algorithms are learned, then evidence for biases is evidence for cognitive algorithms that people haven't had sufficient time/energy/motivation to learn. (This does not imply that intrinsic limitations/biases do not exist or that the study of cognitive biases is a waste of time; rather the ULH implies that educational history is what matters most)

The genome can only specify a limited amount of information. The question is then how much of our advanced cognitive machinery for things like facial recognition, motor planning, language, logic, planning, etc. is innate vs learned. From evolution's perspective there is a huge advantage to preloading the brain with innate algorithms so long as said algorithms have high expected utility across the expected domain landscape.

On the other hand, evolution is also highly constrained in a bit coding sense: every extra bit of code costs additional energy for the vast number of cellular replication events across the lifetime of the organism. Low code complexity solutions also happen to be exponentially easier to find. These considerations seem to strongly favor the ULH but they are difficult to quantify.

Neuroscientists have long known that the brain is divided into physical and functional modules. These modular subdivisions were discovered a century ago by Brodmann. Every time neuroscientists opened up a new brain, they saw the same old cortical modules in the same old places doing the same old things. The specific layout of course varied from species to species, but the variations between individuals are minuscule. This evidence seems to *strongly* favor the EMH.



Throughout most of the 90's up into the 2000's, evidence from computational neuroscience models and AI were heavily influenced by - and unsurprisingly - largely supported the EMH. Neural nets and backprop were known of course since the 1980's and worked on small problems[18], but at the time they didn't scale well - and there was no theory to suggest they ever would.

Theory of the time also suggested local minima would always be a problem (now we understand that local minima are not really the main problem[19], and modern stochastic gradient descent methods combined with highly overcomplete models and stochastic regularization[20] are effectively global optimizers that can often handle obstacles such as local minima and saddle points[21]).

The other related historical criticism rests on the lack of biological plausibility for backprop style gradient descent. (There is as of yet little consensus on how the brain implements the equivalent machinery, but target propagation is one of the more promising recent proposals[22][23].)

Many AI researchers are naturally interested in the brain, and we can see the influence of the EMH in much of the work before the deep learning era. HMAX is a hierarchical vision

system developed in the late 90's by Poggio et al as a working model of biological vision[24]. It is based on a preconfigured hierarchy of modules, each of which has its own mix of innate features such as gabor edge detectors along with a little bit of local learning. It implements the general idea that complex algorithms/features are innate - the result of evolutionary global optimization - while neural networks (incapable of global optimization) use hebbian local learning to fill in details of the design.

Dynamic Rewiring

In a groundbreaking study from 2000 published in Nature, Sharma et al successfully rewired ferret retinal pathways to project into the auditory cortex instead of the visual cortex.[25] The result: auditory cortex can become visual cortex, just by receiving visual data! Not only does the rewired auditory cortex develop the specific gabor features characteristic of visual cortex; the rewired cortex also becomes functionally visual.[26] True, it isn't quite as effective as normal visual cortex, but that could also *possibly* be an artifact of crude and invasive brain rewiring surgery.

The ferret study was popularized by the book *On Intelligence* by Hawkins in 2004 as evidence for a single cortical learning algorithm. This helped percolate the evidence into the wider AI community, and thus probably helped in setting up the stage for the deep learning movement of today. The modern view of the cortex is that of a mostly uniform set of general purpose modules which slowly become recruited for specific tasks and filled with domain specific 'code' as a result of the learning (self optimization) process.

The next key set of evidence comes from studies of atypical human brains with novel extrasensory powers. In 2009 Vuillerme et al showed that the brain could automatically learn to process sensory feedback rendered onto the tongue[27]. This research was developed into a complete device that allows blind people to develop primitive tongue based vision.

In the modern era some blind humans have apparently acquired the ability to perform echolocation (sonar), similar to cetaceans. In 2011 Thaler et al used MRI and PET scans to show that human echolocators use diverse non-auditory brain regions to process echo clicks, predominantly relying on re-purposed 'visual' cortex.[27]

The echolocation study in particular helps establish the case that the brain is actually doing global, highly nonlocal optimization - far beyond simple hebbian dynamics. Echolocation is an active sensing strategy that requires very low latency processing, involving complex timed coordination between a number of motor and sensory circuits - all of which must be learned.

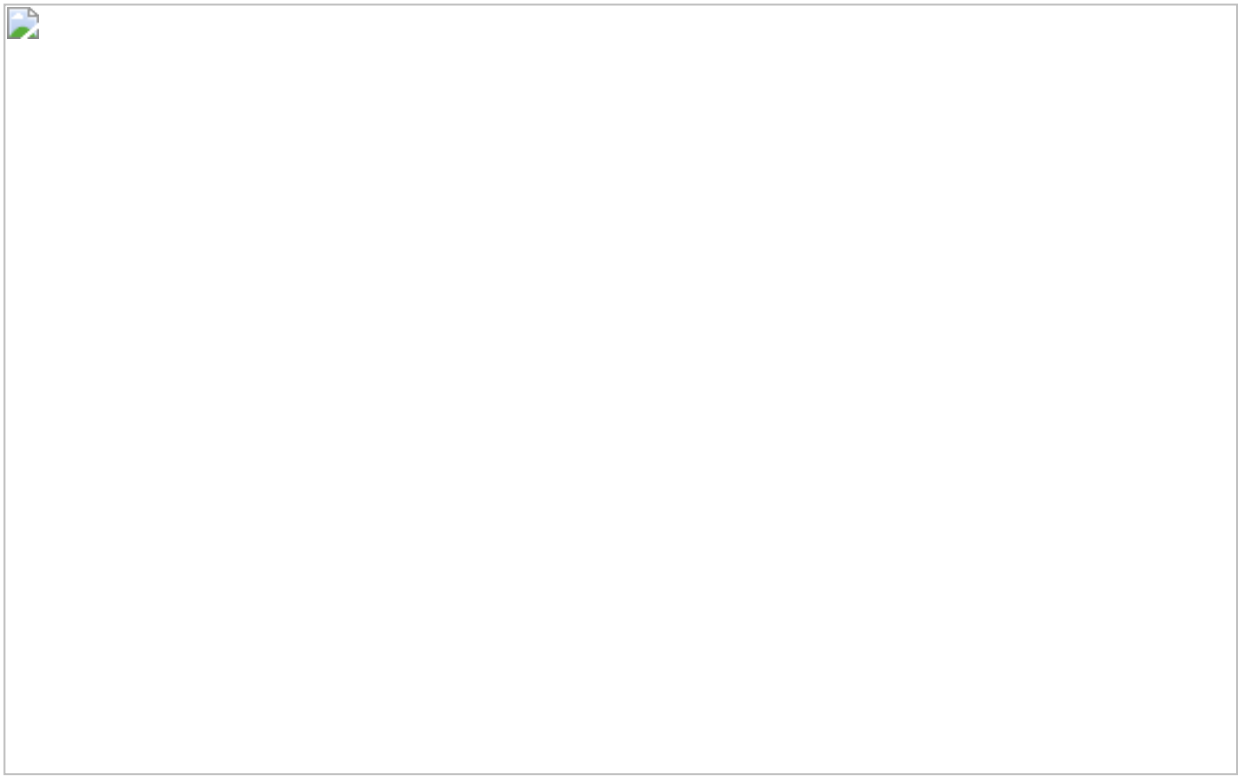
Somehow the brain is dynamically learning how to use and assemble cortical modules to implement mental algorithms: everyday tasks such as visual counting, comparisons of images or sounds, reading, etc - all are task which require simple mental programs that can shuffle processed data between modules (some or any of which can also function as short term memory buffers).

To explain this data, we should be on the lookout for a system in the brain that can learn to control the cortex - a general system that *dynamically* routes data between different brain modules to solve domain specific tasks.

But first let's take a step back and start with a high level architectural view of the entire brain to put everything in perspective.

Brain Architecture

Below is a circuit diagram for the whole brain. Each of the main subsystems work together and are best understood together. You can probably get a good high level extremely coarse understanding of the entire brain is less than one hour.



(there are a couple of circuit diagrams of the whole brain on the web, but this is the best. From this site.)

The human brain has ~100 billion neurons and ~100 trillion synapses, but ultimately it evolved from the bottom up - from organisms with just hundreds of neurons, like the tiny brain of C. Elegans.

We know that evolution is code complexity constrained: much of the genome codes for cellular metabolism, all the other organs, and so on. For the brain, most of its bit budget needs to be spent on all the complex neuron, synapse, and even neurotransmitter level machinery - the low level hardware foundation.

For a tiny brain with 1000 neurons or less, the genome can directly specify each connection. As you scale up to larger brains, evolution needs to create vastly more circuitry while still using only about the same amount of code/bits. So instead of specifying connectivity at the neuron layer, the genome codes connectivity at the module layer. Each module can be built from simple procedural/fractal expansion of progenitor cells.

So the size of a module has little to nothing to do with its innate complexity. The cortical modules are huge - V1 alone contains 200 million neurons in a human - but there is no reason to suspect that V1 has greater initial code complexity than any other brain module. Big modules are built out of simple procedural tiling patterns.

Very roughly the brain's *main* modules can be divided into six subsystems (there are numerous smaller subsystems):

- The neocortex: the brain's primary computational workhorse (blue/purple modules at the top of the diagram). Kind of like a bunch of general purpose FPGA coprocessors.
- The cerebellum: another set of coprocessors with a simpler feedforward architecture. Specializes more in motor functionality.
- The thalamus: the orangish modules below the cortex. Kind of like a relay/routing bus.
- The hippocampal complex: the apex of the cortex, and something like the brain's database.
- The amygdala and limbic reward system: these modules specialize in something like the value function.
- The Basal Ganglia (green modules): the central control system, similar to a CPU.

In the interest of space/time I will focus primarily on the Basal Ganglia and will just touch on the other subsystems very briefly and provide some links to further reading.

The neocortex has been studied extensively and is the main focus of several popular books on the brain. Each neocortical module is a 2D array of neurons (technically 2.5D with a depth of about a few dozen neurons arranged in about 5 to 6 layers).

Each cortical module is something like a general purpose RNN (recursive neural network) with 2D local connectivity. Each neuron connects to its neighbors in the 2D array. Each module also has nonlocal connections to other brain subsystems and these connections follow the same local 2D connectivity pattern, in some cases with some simple affine transformations. Convolutional neural networks use the same general architecture (but they are typically not recurrent.)

Cortical modules - like artifical RNNs - are general purpose and can be trained to perform various tasks. There are a huge number of models of the cortex, varying across the tradeoff between biological realism and practical functionality.

Perhaps surprisingly, any of a wide variety of learning algorithms can reproduce cortical connectivity and features when trained on appropriate sensory data[27]. This is a computational proof of the one-learning-algorithm hypothesis; furthermore it illustrates the general idea that *data* determines functional structure in any general learning system.

There is evidence that cortical modules learn automatically (unsupervised) to some degree, and there is also some evidence that cortical modules can be trained to relearn data from other brain subsystems - namely the hippocampal complex. The dark knowledge distillation technique in ANNs[28][29] is a potential natural analog/model of hippocampus -> cortex knowledge transfer.

Module connections are bidirectional, and feedback connections (from high level modules to low level) outnumber forward connections. We can speculate that something like target propagation can also be used to guide or constrain the development of cortical maps (speculation).

The hippocampal complex is the root or top level of the sensory/motor hierarchy. This short youtube video gives a good seven minute overview of the HC. It is like a spatiotemporal database. It receives compressed scene descriptor streams from the sensory cortices, it stores this information in medium-term memory, and it supports later auto-associative recall of these memories. Imagination and memory recall seem to be basically the same.

The 'scene descriptors' take the sensible form of things like 3D position and camera orientation, as encoded in place, grid, and head direction cells. This is basically the logical result of compressing the sensory stream, comparable to the networking data stream in a multiplayer video game.

Imagination/recall is basically just the reverse of the forward sensory coding path - in reverse mode a compact scene descriptor is expanded into a full imagined scene. Imagined/remembered scenes activate the same cortical subnetworks that originally formed the memory (or would have if the memory was real, in the case of imagined recall).

The amygdala and associated limbic reward modules are rather complex, but look something like the brain's version of the value function for reinforcement learning. These modules are interesting because they clearly rely on learning, but clearly the brain must specify an initial version of the value/utility function that has some minimal complexity.

As an example, consider taste. Infants are born with basic taste detectors and a very simple initial value function for taste. Over time the brain receives feedback from digestion and various estimators of general mood/health, and it uses this to refine the initial taste value function. Eventually the adult sense of taste becomes considerably more complex. Acquired taste for bitter substances - such as coffee and beer - are good examples.

The amygdala appears to do something similar for emotional learning. For example infants are born with a simple versions of a fear response, with is later refined through reinforcement learning. The amygdala sits on the end of the hippocampus, and it is also involved heavily in memory processing.

See also these two videos from khanacademy: one on the limbic system and amygdala (10 mins), and another on the midbrain reward system (8 mins)

The Basal Ganglia

The Basal Ganglia is a wierd looking complex of structures located in the center of the brain. It is a conserved structure found in all vertebrates, which suggests a core functionality. The BG is proximal to and connects heavily with the midbrain reward/limbic systems. It also connects to the brain's various modules in the cortex/hippocampus, thalamus and the cerebellum . . . basically everything.

All of these connections form recurrent loops between associated compartmental modules in each structure: thalamocortical/hippocampal-cerebellar-basal_ganglial loops.

Just as the cortex and hippocampus are subdivided into modules, there are corresponding modular compartments in the thalamus, basal ganglia, and the cerebellum. The set of modules/compartments in each main structure are all highly interconnected with their correspondents across structures, leading to the concept of distributed processing modules.

Each DPM forms a recurrent loop across brain structures (the local networks in the cortex, BG, and thalamus are also locally recurrent, whereas those in the cerebellum are not). These recurrent loops are mostly separate, but each sub-structure also provides different opportunities for inter-loop connections.

The BG appears to be involved in essentially all higher cognitive functions. Its core functionality is *action selection* via subnetwork switching. In essence action selection is the core problem of intelligence, and it is also general enough to function as the building block of all higher functionality. A system that can select between motor actions can also select between tasks or subgoals. More generally, low level action selection can easily form the basis of a Turing Machine via selective routing: deciding where to route the output of thalamocortical-cerebellar modules (some of which may specialize in short term memory as in the prefrontal cortex, although all cortical modules have some short term memory capability).

There are now a number of computational models for the Basal Ganglia-Cortical system that demonstrate possible biologically plausible implementations of the general theory[28] [29]; integration with the hippocampal complex leads to larger-scale systems which aim to model/explain most of higher cognition in terms of sequential mental programs[30] (of course fully testing any such models awaits sufficient computational power to run very large-scale neural nets).

For an extremely oversimplified model of the BG as a dynamic router, consider an array of N distributed modules controlled by the BG system. The BG control network expands these N inputs into an NxN matrix. There are N^2 potential intermodular connections, each of which can be individually controlled. The control layer reads a compressed, downsampled version of the module's hidden units as its main input, and is also recurrent. Each output node in the BG has a multiplicative gating effect which selectively enables/disables an individual intermodular connection. If the control layer is naively fully connected, this would require $(N^2)^2$ connections, which is only feasible for N ~ 100 modules, but sparse connectivity can substantially reduce those numbers.

It is unclear (to me), whether the BG actually implements NxN style routing as described above, or something more like 1xN or Nx1 routing, but there is general agreement that it implements cortical routing.

Of course in actuality the BG architecture is considerably more complex, as it also must implement reinforcement learning, and the intermodular connectivity map itself is also probably quite sparse/compressed (the BG may not control all of cortex, certainly not at a

uniform resolution, and many controlled modules may have a very limited number of allowed routing decisions). Nonetheless, the simple multiplicative gating model illustrates the core idea.

This same multiplicative gating mechanism is the core principle behind the highly successful LSTM (Long Short-Term Memory)[30] units that are used in various deep learning systems. The simple version of the BG's gating mechanism can be considered a wider parallel and hierarchical extension of the basic LSTM architecture, where you have a parallel array of N memory cells instead of 1, and each memory cell is a large vector instead of a single scalar value.

The main advantage of the BG architecture is parallel hierarchical approximate control: it allows a large number of hierarchical control loops to update and influence each other in parallel. It also reduces the huge complexity of general routing across the full cortex down into a much smaller-scale, more manageable routing challenge.

Implications for AGI

These two conceptions of the brain - the universal learning machine hypothesis and the evolved modularity hypothesis - lead to very different predictions for the likely route to AGI, the expected differences between AGI and humans, and thus any consequent safety issues and strategies.

In the extreme case imagine that the brain is a pure ULM, such that the genetic prior information is close to zero or is simply unimportant. In this case it is vastly more likely that successful AGI will be built around designs very similar to the brain, as the ULM architecture in general is the natural ideal, vs the alternative of having to hand engineer all of the AI's various cognitive mechanisms.

In reality learning is computationally hard, and any practical general learning system depends on good priors to constrain the learning process (essentially taking advantage of previous knowledge/learning). The recent and rapid success of deep learning is strong evidence for how much prior information is ideal: just a little. The prior in deep learning systems takes the form of a compact, small set of hyperparameters that control the learning process and specify the overall network architecture (an extremely compressed prior over the network topology and thus the program space).

The ULH suggests that most everything that defines the human mind is cognitive software rather than hardware: the adult mind (in terms of algorithmic information) is 99.999% a cultural/memetic construct. Obviously there are some important exceptions: infants are born with some functional but very primitive sensory and motor processing 'code'. Most of the genome's complexity is used to specify the learning machinery, and the associated reward circuitry. Infant emotions appear to simplify down to a single axis of happy/sad; differentiation into the more subtle vector space of adult emotions does not occur until later in development.

If the mind is software, and if the brain's learning architecture is already universal, then AGI could - by default - end up with a similar distribution over mindspace, simply because it will be built out of similar general purpose learning algorithms running over the same general dataset. We already see evidence for this trend in the high functional similarity between the features learned by some machine learning systems and those found in the cortex.

Of course an AGI will have little need for some specific evolutionary features: emotions that are subconsciously broadcast via the facial muscles is a quirk unnecessary for an AGI - but that is a rather specific detail.

The key takeway is that the data is what matters - and in the end it is all that matters. Train a universal learner on image data and it just becomes a visual system. Train it on speech data and it becomes a speech recognizer. Train it on ATARI and it becomes a little gamer agent.

Train a universal learner on the real world in something like a human body and you get something like the human mind. Put a ULM in a dolphin's body and echolocation is the natural primary sense, put a ULM in a human body with broken visual wiring and you can also get echolocation.

Control over training is the most natural and straightforward way to control the outcome.

To create a superhuman AI driver, you 'just' need to create a realistic VR driving sim and then train a ULM in that world (better training and the simple power of selective copying leads to superhuman driving capability).

So to create benevolent AGI, we should think about how to create virtual worlds with the right structure, how to educate minds in those worlds, and how to safely evaluate the results.

One key idea - which I proposed five years ago is that the AI *should not know it is in a sim.*

New AI designs (world design + architectural priors + training/education system) should be tested first in the safest virtual worlds: which in simplification are simply low tech worlds without computer technology. Design combinations that work well in safe low-tech sandboxes are promoted to less safe high-tech VR worlds, and then finally the real world.

A key principle of a secure code sandbox is that the code you are testing should not be aware that it is in a sandbox. If you violate this principle then you have already failed. Yudkowsky's AI box thought experiment assumes the violation of the sandbox security principle apriori and thus is something of a distraction. (the virtual sandbox idea was most likely discussed elsewhere previously, as Yudkowsky indirectly critiques a strawman version of the idea via this sci-fi story).

The virtual sandbox approach also combines nicely with invisible thought monitors, where the AI's thoughts are automatically dumped to searchable logs.

Of course we will still need a solution to the value learning problem. The natural route with brain-inspired AI is to learn the key ideas behind value acquisition in humans to help derive an improved version of something like inverse reinforcement learning and or imitation learning[31] - an interesting topic for another day.

Conclusion

Ray Kurzweil has been predicting for decades that AGI will be built by reverse engineering the brain, and this particular prediction is not especially unique - this has been a popular position for quite a while. My own investigation of neuroscience and machine learning led me to a similar conclusion some time ago.

The recent progress in deep learning, combined with the emerging modern understanding of the brain, provide further evidence that AGI could arrive around the time when we can build and train ANNs with similar computational power as measured very roughly in terms of neuron/synapse counts. In general the evidence from the last four years or so supports Hanson's viewpoint from the Foom debate. More specifically, his general conclusion:

> Future superintelligences will exist, but their vast and broad mental capacities will come mainly from vast mental content and computational resources. By comparison, their general architectural innovations will be minor additions.

The ULH supports this conclusion.

Current ANN engines can already train and run models with around 10 million neurons and 10 billion (compressed/shared) synapses on a single GPU, which suggests that the goal could soon be within the reach of a large organization. Furthermore, Moore's Law for GPUs still has some steam left, and software advances are currently improving simulation performance at a faster rate than hardware. These trends implies that Anthropomorphic/Neuromorphic AGI could be surprisingly close, and may appear suddenly.

What kind of leverage can we exert on a short timescale?

Pattern-botching: when you forget you understand

It’s all too easy to let a false understanding of something replace your actual understanding. Sometimes this is an oversimplification, but it can also take the form of an over*complication*. I have an illuminating story:

Years ago, when I was young and foolish, I found myself in a particular romantic relationship that would later end for epistemic reasons, when I was slightly less young and slightly less foolish. Anyway, this particular girlfriend of mine was very into healthy eating: raw, organic, home-cooked, etc. During her visits my diet would change substantially for a few days. At one point, we got in a tiny fight about something, and in a not-actually-desperate chance to placate her, I semi-jokingly offered: “I’ll go vegetarian!”

“I don’t care,” she said with a sneer.

…and she didn’t. She wasn’t a vegetarian. Duhhh… I knew that. We’d made some ground beef together the day before.

So what was I thinking? Why did I say “I’ll go vegetarian” as an attempt to appeal to her values?

(I’ll invite you to take a moment to come up with your own model of why that happened. You don't have to, but it can be helpful for evading hindsight bias of obviousness.)

(Got one?)

Here's my take: I pattern-matched a bunch of actual preferences she had with a general "healthy-eating" cluster, and then I went and pulled out something random that felt vaguely associated. It's telling, I think, that I don't even explicitly *believe* that vegetarianism is healthy. But to my pattern-matcher, they go together nicely.

I'm going to call this **pattern-botching**.† Pattern-botching is when you pattern-match a thing "X", as following a certain model, but then implicit queries to that model return properties that aren't true about X. What makes this different from just having false beliefs is that you *know* the truth, but you're forgetting to use it because there's a botched model that is easier to use.

†Maybe this already has a name, but I've read a lot of stuff and it feels like a distinct concept to me.

Examples of pattern-botching

So, that's pattern-botching, in a nutshell. Now, examples! We'll start with some simple ones.

Calmness and pretending to be a zen master

In my Againstness Training video, past!me tries a bunch of things to calm down. In the pursuit of "calm", I tried things like...

- dissociating
- trying to imitate a zen master
- speaking really quietly and timidly

None of these are the desired state. The desired state is present, authentic, and can project well while speaking assertively.

But that would require actually being in a different state, which to my brain at the time seemed hard. So my brain constructed a pattern around the target state, and said "what's easy and looks vaguely like this?" and generated the list above. Not as a list, of course! That would be too easy. It generated each one individually as a plausible course of action, which I then tried, and which Val then called me out on.

Personality Types

I'm quite gregarious, extraverted, and generally unflappable by noise and social situations. Many people I know describe themselves as HSPs (Highly Sensitive Persons) or as very introverted, or as "not having a lot of spoons". These concepts are related—or perhaps not related, but at least correlated—but they're not the same. And even if these three terms *did* all mean the same thing, individual people would still vary in their needs and preferences.

Just this past week, I found myself talking with an HSP friend L, and noting that I didn't really know what her needs were. Like I knew that she was easily startled by loud noises and often found them painful, and that she found motion in her periphery distracting. But beyond that... yeah. So I told her this, in the context of a more general conversation about her HSPness, and I said that I'd like to learn more about her needs.

L responded positively, and suggested we talk about it at some point. I said, "Sure," then added, "though it would be helpful for me to know just this one thing: how would you feel about me asking you about a specific need in the middle of an interaction we're having?"

"I would love that!" she said.

"Great! Then I suspect our future interactions will go more smoothly," I responded. I realized what had happened was that I had conflated L's HSPness with... something else. I'm not exactly sure what, but a preference for indirect communication, perhaps? I have another friend, who is also sometimes short on spoons, who I model as finding that kind of question stressful because it would kind of put them on the spot.

I've only just recently been realizing this, so I suspect that I'm still doing a ton of this pattern-botching with people, that I haven't specifically noticed.

Of course, having clusters makes it easier to have heuristics about what people will do, without knowing them too well. A loose cluster is better than nothing. I think the issue is when we *do* know the person well, but we're still relying on this cluster-based model of them. It's telling that I was **not actually surprised** when L said that she would like it if I asked about her needs. On some level I kind of already knew it. But my botched pattern was making me doubt what I knew.

False aversions

CFAR teaches a technique called "Aversion Factoring", in which you try to break down the reasons why you don't do something, and then consider each reason. In some cases, the reasons are sound reasons, so you decide not to try to force yourself to do the thing. If not, then you want to make the reasons go away. There are three types of reasons, with different approaches.

One is for when you have a legitimate issue, and you have to redesign your plan to avert that issue. The second is where the thing you're averse to is real but isn't actually bad, and you can kind of ignore it, or maybe use exposure therapy to get yourself more comfortable with it. The third is... when the outcome *would* be an issue, but *it's not actually a necessary outcome of the thing*. As in, it's a fear that's vaguely associated with the thing at hand, but the thing you're afraid of isn't real.

All of these share a structural similarity with pattern-botching, but the third one in particular is a great example. The aversion is generated from a property that the thing you're averse to doesn't actually have. Unlike a miscalibrated aversion (#2 above) it's usually pretty obvious under careful inspection that the fear itself is based on a botched model of the thing you're averse to.

Taking the training wheels off of your model

One other place this structure shows up is in the difference between what something looks like when you're learning it versus what it looks like once you've learned it. Many people learn to ride a bike while actually riding a four-wheeled vehicle: training wheels. I don't think anyone makes the mistake of thinking that the ultimate bike will have training wheels, but in other contexts it's much less obvious.

The remaining three examples look at how pattern-botching shows up in learning contexts, where people implicitly forget that they're only partway there.

Rationality as a way of thinking

CFAR runs 4-day rationality workshops, which currently are evenly split between specific techniques and how to approach things in general. Let's consider what kinds of behaviours spring to mind when someone encounters a problem and asks themselves: "what would be a rational approach to this problem?"

- someone with a really naïve model, who hasn't actually learned much about applied rationality, might pattern-match "rational" to "hyper-logical", and think

"What Would Spock Do?"
- someone who is somewhat familiar with CFAR and its instructors but who still doesn't know any rationality techniques, might complete the pattern with something that they think of as being archetypal of CFAR-folk: "What Would Anna Salamon Do?"
- CFAR alumni, especially new ones, might pattern-match "rational" as "using these rationality techniques" and conclude that they need to "goal factor" or "use trigger-action plans"
- someone who *gets* rationality would simply apply that particular structure of thinking to their problem

In the case of a bike, we see hundreds of people biking around without training wheels, and so that becomes the obvious example from which we generalize the pattern of "bike". In other learning contexts, though, most people—including, sometimes, the people at the leading edge—are still in the early learning phases, so the training wheels are the rule, not the exception.

So people start thinking that the figurative bikes are supposed to have training wheels.

Incidentally, this can also be the grounds for strawman arguments where detractors of the thing say, "Look at these bikes [with training wheels]! How are you supposed to get anywhere on them?!"

Effective Altruism

We potentially see a similar effect with topics like Effective Altruism. It's a movement that is still in its infancy, which means that *nobody* has it all figured out. So when trying to answer "How do I be an effective altruist?" our pattern-matchers might pull up a bunch of examples of things that EA-identified people have been commonly observed to do.

- donating 10% of one's income to a strategically selected charity
- going to a coding bootcamp and switching careers, in order to Earn to Give
- starting a new organization to serve an unmet need, or to serve a need more efficiently
- supporting the Against Malaria Fund

...and this generated list might be helpful for various things, but be wary of thinking that it represents what Effective Altruism *is*. It's possible—it's almost *inevitable*—that we don't actually know what the most effective interventions are yet. We will potentially never actually know, but we can expect that in the future we will generally know more than at present. Which means that **the current sampling of good EA behaviours likely does not actually even cluster around the ultimate set of behaviours we might expect.**

Creating a new (platform for) culture

At my intentional community in Waterloo, we're building a new culture. But that's actually a by-product: our goal isn't to build this particular culture but to build a platform on which many cultures can be built. It's like how as a company you don't

just want to be building the product but rather building the company itself, or "the machine that builds the product," as Foursquare founder Dennis Crowley puts it.

What I started to notice though, is that we started to confused the particular, *transitionary* culture that we have at our house, with either (a) the particular, target culture, that we're aiming for, or (b) the more abstract range of cultures that will be constructable on our platform.

So from a training wheels perspective, we might totally eradicate words like "should". I did this! It was really helpful. But once I had removed the word from my idiolect, it became unhelpful to still be treating it as being a touchy word. Then I heard my mentor use it, and I remembered that the point of removing the word *wasn't* to not ever use it, but to train my brain to think without a particular structure that "should" represented.

This shows up on much larger scales too. Val from CFAR was talking about a particular kind of fierceness, "hellfire", that he sees as fundamental and important, and he noted that it seemed to be incompatible with the kind of culture my group is building. I initially agreed with him, which was kind of dissonant for my brain, but then I realized that hellfire was only incompatible with our *training* culture, not *the entire set of cultures that could ultimately be built on our platform*. That is, engaging with hellfire would potentially interfere with the learning process, but it's not ultimately proscribed by our culture platform.

Conscious cargo-culting

I think it might be helpful to repeat the definition:

> Pattern-botching is you pattern-match a thing "X", as following a certain model, but then but then implicit queries to that model return properties that aren't true about X. What makes this different from just having false beliefs is that you *know* the truth, but you're forgetting to use it because there's a botched model that is easier to use.

It's kind of like if you were doing a cargo-cult, except you knew how airplanes worked.

(Cross-posted from malcolmocean.com)

Taking the reins at MIRI

Hi all. In a few hours I'll be taking over as executive director at MIRI. The LessWrong community has played a key role in MIRI's history, and I hope to retain and build your support as (with more and more people joining the global conversation about long-term AI risks & benefits) MIRI moves towards the mainstream.

Below I've cross-posted my introductory post on the MIRI blog, which went live a few hours ago. The short version is: there are very exciting times ahead, and I'm honored to be here. Many of you already know me in person or through my blog posts, but for those of you who want to get to know me better, I'll be running an AMA on the effective altruism forum at 3PM Pacific on Thursday June 11th.

I extend to all of you my thanks and appreciation for the support that so many members of this community have given to MIRI throughout the years.

Hello, I'm Nate Soares, and I'm pleased to be taking the reins at MIRI on Monday morning.

For those who don't know me, I've been a research fellow at MIRI for a little over a year now. I attended my first MIRI workshop in December of 2013 while I was still working at Google, and was offered a job soon after. Over the last year, I wrote a dozen papers, half as primary author. Six of those papers were written for the MIRI technical agenda, which we compiled in preparation for the Puerto Rico conference put on by the FLI in January 2015. Our technical agenda is cited extensively in the research priorities document referenced by the open letter that came out of that conference. In addition to the Puerto Rico conference, I attended five other conferences over the course of the year, and gave a talk at three of them. I also put together the MIRI research guide (a resource for students interested in getting involved with AI alignment research), and of course I spent a fair bit of time doing the actual research at workshops, at researcher retreats, and on my own. It's been a jam-packed year, and it's been loads of fun.

I've always had a natural inclination towards leadership: in the past, I've led a F.I.R.S.T. Robotics team, managed two volunteer theaters, served as president of an Entrepreneur's Club, and co-founded a startup or two. However, this is the first time I've taken a professional leadership role, and I'm grateful that I'll be able to call upon the experience and expertise of the board, of our advisors, and of outgoing executive director Luke Muehlhauser.

MIRI has improved greatly under Luke's guidance these last few years, and I'm honored to have the opportunity to continue that trend. I've spent a lot of time in conversation with Luke over the past few weeks, and he'll remain a close advisor going forward. He and the management team have spent the last year or so really tightening up the day-to-day operations at MIRI, and I'm excited about all the opportunities we have open to us now.

The last year has been pretty incredible. Discussion of long-term AI risks and benefits has finally hit the mainstream, thanks to the success of Bostrom's *Superintelligence* and FLI's Puerto Rico conference, and due in no small part to years of movement-

building and effort made possible by MIRI's supporters. Over the last year, I've forged close connections with our friends at the Future of Humanity Institute, the Future of Life Institute, and the Centre for the Study of Existential Risk, as well as with a number of industry teams and academic groups who are focused on long-term AI research. I'm looking forward to our continued participation in the global conversation about the future of AI. These are exciting times in our field, and MIRI is well-poised to grow and expand. Indeed, one of my top priorities as executive director is to grow the research team.

That project is already well under way. I'm pleased to announce that Jessica Taylor has accepted a full-time position as a MIRI researcher starting in August 2015. We are also hosting a series of summer workshops focused on various technical AI alignment problems, the second of which is just now concluding. Additionally, we are working with the Center for Applied Rationality to put on a summer fellows program designed for people interested in gaining the skills needed for research in the field of AI alignment.

I want to take a moment to extend my heartfelt thanks to all those supporters of MIRI who have brought us to where we are today: We have a slew of opportunities before us, and it's all thanks to your effort and support these past years. MIRI couldn't have made it as far as it has without you. Exciting times are ahead, and your continued support will allow us to grow quickly and pursue all the opportunities that the last year opened up.

Finally, in case you want to get to know me a little better, I'll be answering questions on the effective altruism forum at 3PM Pacific time on Thursday June 11th.

Onwards,

Top 9+2 myths about AI risk

Following some somewhat misleading articles quoting me, I thought I'd present the top 9 myths about the AI risk thesis:

1. **That we're certain AI will doom us**. Certainly not. It's very hard to be certain of anything involving a technology that doesn't exist; we're just claiming that the probability of AI going bad isn't low enough that we can ignore it.
2. **That humanity will survive, because we've always survived before**. Many groups of humans haven't survived contact with more powerful intelligent agents. In the past, those agents were other humans; but they need not be. The universe does not owe us a destiny. In the future, *something* will survive; it need not be us.
3. **That uncertainty means that you're safe**. If you're claiming that AI is impossible, or that it will take countless decades, or that it'll be safe... you're not being uncertain, you're being extremely specific about the future. "No AI risk" is certain; "Possible AI risk" is where we stand.
4. **That Terminator robots will be involved**. Please? The threat from AI comes from its potential intelligence, not from its ability to clank around slowly with an Austrian accent.
5. **That we're assuming the AI is too dumb to know what we're asking it**. No. A powerful AI will know what we meant to program it to do. But why should it care? And if we could figure out how to program "care about what we meant to ask", well, then we'd have safe AI.
6. **That there's one simple trick that can solve the whole problem**. Many people have proposed that one trick. Some of them could even help (see Holden's tool AI idea). None of them reduce the risk enough to relax – and many of the tricks contradict each other (you can't design an AI that's both a tool and socialising with humans!).
7. **That we want to stop AI research**. We don't. Current AI research is very far from the risky areas and abilities. And it's risk aware AI researchers that are most likely to figure out how to make safe AI.
8. **That AIs will be more intelligent than us, hence more moral**. It's pretty clear than in humans, high intelligence is no guarantee of morality. Are you really willing to bet the whole future of humanity on the idea that AIs *might* be different? That in the billions of possible minds out there, there is none that is both dangerous and very intelligent?
9. **That science fiction or spiritual ideas are useful ways of understanding AI risk**. Science fiction and spirituality are full of human concepts, created by humans, for humans, to communicate human ideas. They need not apply to AI at all, as these could be minds far removed from human concepts, possibly without a body, possibly with no emotions or consciousness, possibly with many new emotions and a different type of consciousness, etc... Anthropomorphising the AIs could lead us completely astray.

Lists cannot be comprehensive, but they can adapt and grow, adding more important points:

1. **That AIs have to be evil to be dangerous**. The majority of the risk comes from indifferent or partially nice AIs. Those that have some goal to follow, with humanity and its desires just getting in the way – using resources, trying to oppose it, or just not being perfectly efficient for its goal.

2. **That we believe AI is coming soon. It might; it might not**. Even if AI is known to be in the distant future (which isn't known, currently), some of the groundwork is worth laying now.

Your "shoulds" are not a duty

This is a linkpost for https://mindingourway.com/shoulds-are-not-a-duty/

I have a friend who, after reading my last two posts, still struggled to give up her shoulds. She protested that, if she stopped doing things because she should, then she might do the wrong thing. I see this frequently, even among people who claim to be moral relativists: they protest that if they weigh their wants and their shoulds on the same scales, then they might make the wrong choice.

But this notion of "right" vs "wrong" cannot come from outside. There is no stone tablet among the stars that mandates what is right. Moral relativists usually have no trouble remembering that their narrow, short-term desires (for comfort, pleasure, etc.) are internal, but many seem to forget that their wide, long-term desires (flourishing, less suffering, etc.) are also part of them.

Why did my friend worry that, if she stopped forcing herself to follow her shoulds, that she might do the wrong thing? There are no outside authorities punishing people who don't follow their best interests. There are no heavenly gatekeepers rewarding you for doing something other than what's best. The only reason to care about doing what's right is because you want what's right to be done. Why was she afraid that she would fail to follow her own interests, if she stopped using internal force? It's *you* who wants the right thing done, so if you fear you're not going to do the right things, then bargain with yourself.

Part of the problem, I think, is that she realized that she wants to both (a) do the right thing and (b) avoid all the effort that entails, and she feared that without the tool of internal force, she would be unable to do as much good as she wants to do. This is a valid concern, and following posts will discuss different tools for doing what you want (without resorting to internal force, which I think is unsustainable — remember, expending willpower is a stopgap, not a solution).

But another part of the problem, I think, was a lingering sense of resentment towards the shoulds, for trying to suck fun and enjoyment out of her life.

I see this often. Picture someone who needs to choose between playing video games all day (and losing their job) or getting abused by customers all day (for not all that much money). They conclude that they "should" do the job, and they feel compelled by the should. And then, over and over, I find my friends resenting their shoulds, as if the shoulds came from outside, from Beyond, from the Intergalactic Oughthorities. They treat their should like shackles that bind them to the "right path", the one where have to go to work when they could be playing video games.

But the shoulds aren't the shackles. There aren't any oughthorities. You always get to do as you please, within the bounds allowed by the universe. It's the *situation* which forces you to choose between bad and worse. Don't resent the bad option for being better than the worse option — if you must resent something, resent the situation.

(Or, better yet, turn your resentment into a cold resolve to *change* the situation.)

If you ever start to feel that your shoulds are obligations, then remember this:

The shoulds were made for us, not us for them.

There are no facts about the stars that say what you ought to do. Your shoulds are not written in the heavens, nor in the void.

But your shoulds are written in *you.*

What you "should do" in any given situation is a fact about your brain and the situation (which takes into account your current state of knowledge, and the amount of time you have available, and so on).

In other words, someone with a ton of computing power and intimate knowledge of your brain could *tell* you what you should do in any given situation.

Imagine being told some of those facts about what you "should do, as computed by someone with ridiculous amounts of computing power. They print them out on a sheet of paper, and hand it to you. What would this sheet of paper look like?

I think that most people expect it would look like a long list of obligations, full of uncomfortable task they're actively trying to not remember. Most people seem to expect a highly aversive list that reads something like this:

- Clean your room.
- Send a message to that one friend you fell out of touch with who sent you a message on your birthday.
- Reconcile with your father.
- Donate more to charity.
- ...

This is exactly the notion of "should" that I'm trying to discharge.

Your true shoulds, if I could show them to you, would not look like a list of obligations. Your true shoulds would look like a *recipe for building a utopia.*

They would look like a series of steps that make the world the best place you can make it.

And they wouldn't tell you to do anything psychologically unrealistic, either. Just as the list wouldn't say "snap your fingers in just such a way that alzheimers is cured," the list wouldn't say "work yourself to the bone for 16 hours a day while still remaining in high spirits." No, the true shoulds (as computed by someone with deep knowledge of your brain and ridiculous amounts of computing power) would appear to you as a psychologically possible list of things that happened to have surprisingly awesome impacts on the world.

The things that you feel resentment towards are false shoulds, or at least twisted shoulds. Encountering one of your *actual* moral bonds feels very different indeed. A true opportunity to execute a moral commitment feels not like an obligation, but like a *priviledge*. It feels like executing a Screw The Rules I'm Doing What's Right trope.

In fiction, picture the moment when the villain reveals that doing the Right Thing will start a war, and the hero sets their jaw, looks them in the eyes, and says "so be it," and then does the right thing anyway.

In real life, think of Irena Sendler, who smuggled thousands of Jewish children to safety during the holocaust, who was captured by the Nazis and tortured and had her legs broken and was sentenced to death, and who escaped anyway,

and then *went back.*

Imagine what was going through her mind, when she decided to go back and save more people. Now, of course, I have no idea what she was actually feeling, but when I imagine what it would take for *me* to go back under those circumstances, I imagine feeling fear, and a hint of despair at finding myself still capable, but also a burning resolve to do the right thing anyway.

I imagine her feeling that having the opportunity to go back was a *privilege.* Not an external obligation whispered down from the heavens, but an internal fire, a defiance of the natural order, a need to make the world *different* from the way it would be otherwise.

Irena didn't have an *obligation* to keep fighting. She had more than discharged her moral duty. And while I'm willing to bet that at least part of her was scared, and at least part of her wished she had been crippled and unable to return, there was also a part of her that didn't look at the opportunity to return to save more children as a misfortune, but as an honor.

Can you begin to see the difference between a false should, and a true moral commitment? Think of a false should, one that gives you a strong sense of obligation and a hint of resentment (such as "finish this paper" or "go to work tomorrow"). Now imagine of Irena Sendler, offered the opportunity to return to Warsaw. Imagine what went on in her head, in that moment.

I imagine a mind afraid, but unified, because for her, it wasn't really a choice. Innocent children were still dying, and there was only one thing to do.

That's what a true moral impulse feels like, when you find one. Not like an obligation, but like a piece of cold iron found deep in your core, the thing that you touch — or that touches you — in the moment that you really see the best option available to you, the moment that you realize you already know which way this choice is going to go.

Your shoulds are not shackles, and I caution you to be wary of anyone who tries to force a should upon you. For if you are not careful, you may start to feel like your shoulds are obligations, and you may start to resent them.

Human moral bonds aren't compulsions. They are what let Irena Sendler see the opportunity to risk life and limb to save just one more child, and treat it not as a duty, but as an honor. If you told *her* that she didn't have to go back, that she'd done enough, that she'd earned the right to turn away, and you asked her why the hell she was still going back to Warsaw,

then she's allowed to reach inside, touch that something of iron, look you in the eyes, and say "because I should."

That's how you use a should. Not with obligation and resentment, but with steel in your heart and no other choice that compares.

I strongly encourage you to unpack your shoulds into their component wants and desires — I would rather not be responsible for inspiring a bunch of people to run around shoulding themselves and saying "no, it's OK, these are the true moral bonds." Rather, the point I'm trying to make is this:

Many treat their moral impulses as a burden. But I say, find all the parts that feel like a burden, and drop them. Keep only the things that fill you with resolve, the things you would risk life and limb to defend.

Those moral impulses are not a reminder of your grudging duty. They are a reminder that you value things larger than yourself. They are a description of everything you're fighting for. They are the birthright of humanity, they are your love for fellow sentient creatures, they are everything we struggle so hard to send upwards to the stars.

They aren't a duty. They're an honor.

Working yourself ragged is not a virtue

This is a linkpost for https://mindingourway.com/stop-before-you-drop/

Let's get back to the "replacing guilt" series. Here's a quick recap of what we've covered so far:

Part 1 was about replacing the listless guilt: if someone feels vaguely guilty for not really doing anything with their life, then the best advice I can give is to start doing something. Find something to fight for. Find a way that the world is not right, and decide to change it. Once the guilt is about failing at a *specific* task, then we can start addressing it.

Part 2 was about refusing to treat your moral impulses as obligations. Be wary of the word should, which tries to force an obligation upon you. I recommend refusing to do anything just because you "should": Insofar as that sets you free, the obligations were false ones. Insofar as that sparks fear that something important won't get completed, seek out the cause of the worry, and complete the task because you want to see it done, rather than because you "should."

However, having something to change in the world and being free of false obligations is not anywhere near enough to replace guilt motivation. In fact, I think that most guilt in most people comes from a different source: it comes from people honestly deciding that X is what they want to do and then finding themselves *not doing X anyway.*

Maybe they *know* that watching another episode of a TV show will cause them to stay up too late and be tired at class tomorrow, and they *know* that their classes are very expensive and that their parents would be very disappointed, and they decide that the best thing to do would be to stop binge-watching the TV show and get some sleep — and then they find themselves watching the next episode anyway.

This sort of guilt is one of the most demoralizing, and therefore it's perhaps one of the most damaging types of guilt. Addressing it is going to require quite a few different tools. Today, I'll describe one of them.

(If you haven't read half-assing it with everything you've got, recommend doing so now: I wrote it as a direct predecessor to this idea, before realizing that I actually needed the previous seven or so posts first.)

Here's a failure mode that I used to see all the time, back when I was a professional programmer: A co-worker of mine would be working on a project that was *almost* under control. It would be a Friday afternoon, with an important deadline coming up in a few weeks, and everything would be almost passable but slightly behind schedule. Some dire bugs demanded fixes, some poor decisions required refactoring. Inevitably, my co-worker would conclude that if they just worked really hard *this* Friday, then they could finish the big refactor, and once that was done, *next* week they could get all the bugs under control, and then by the beginning of the week after that, everything would be back on track again.

(We all know how this story goes.)

Inevitably, co-workers of this type were constantly stressed, and reliably worked late into the night.

I suspect that most people who act like this are guilt-motivated. They're often the sort of person who feels guilty if they stop working before they're completely exhausted. Sometimes, they feel guilty for stopping even when they *are* exhausted, if there is still more work to be done. It's as if part of them believes that if they stop before they're physically forced to drop, and there's still work to be done, then they're being Bad.

This sort of behavior can stem from a number of mistakes. First and foremost, it seems to me that this sort of programmer is usually pursuing a lost purpose. They have succumbed to tryer propaganda; they have confused the quality line for the preference curve. I sometimes want to grab them by the shoulders late on a weekend, look them in the eyes, and ask them what they're fighting for — surely not this? You're allowed to fight for something!

But I also see this failure mode in people who love their work, who believe in its importance. And yet, they still work themselves to exhaustion in a binge/recovery cycle, as if this were the best way to cause their project to succeed.

These people seem to be following an impulse to work as hard as possible whenever they can, perhaps due to a belief that it is unvirtuous for one to stop working when they could continue.

This is an error. **The goal is not to maximize how much work you get done today. The goal is to maximize your productivity over time.**

People who feel guilty for stopping work when they could continue seem to be trying to maximize their local velocity: they feel a need to produce as much as they can, right now, on pain of guilt if they fail. But the actual goal is to maximize the total distance traveled, to maximize how much important work you can get done over time.

(When all is said and done, and Nature passes her final judgement, you will not be measured by the number of moments in which you worked as hard as you could. You will be measured by what actually happened, as will we all.)

People driven by guilt and shame often feel bad for *slowing down.* This is about as effective as starting a marathon with a dead sprint, and then feeling bad for slowing down when you can't sustain it.

Working yourself ragged is not a virtue. You don't get extra points for effort. In fact, you *lose* points for effort: effort is costly; spend it only to purchase better outcomes. The goal is not to appear to be working hard, the goal is to improve the world. Sometimes you do need to push yourself to the limit, but before you do, acknowledge the costs and weigh the tradeoffs, while keeping your long-term goals in view.

We're not yet gods. We're still apes. Remember to pay attention to the distance you need to cover, and remember to pay attention to yourself.

Being a human can be frustrating. Human-bodies aren't as productive as we might like them to be, and running a human-body at maximum capacity for too long causes stress, chronic exhaustion, burnout, and psychological damage. With this in mind,

doesn't it seem a bit confused for a person to berate themselves for stopping before they've spent all their available reserves?

Let me be clear: I'm *not* saying to restrict yourself to only 40 hour per week of work because it's important to pace yourself. I'm just saying that it's important to pace yourself. Do as much as you can, but don't be constantly taking damage. We aren't yet gods. We're still fragile. If you have something urgent to do, then work as hard as you can — but work as hard as you can *over a long period of time,* not *in the moment*.

I'm also not saying "stop as soon as working feels hard." When exercising, it's important to understand the difference between soreness and strain, between pushing yourself and hurting yourself, and the same is true psychologically. Maintaining focus and productivity for long periods of time is a skill that can be trained, like any other. (More on that later, but spoiler alert, "feeling really guilty when you didn't work as hard as you wanted to" is not the best way to train this skill.)

Push your limits! Some things are worth fighting for! But while you're doing that, recognize that the way to complete a marathon isn't to sprint 42+ kilometers.

There is no shame in doing less than you could do in any given moment. Most guilt-motivated people I meet would do well to worry less about whether they're going fast enough *now*, and worry more about whether the amount of work they're doing day-to-day is ideal in the long term, taking psychological constraints into account. You don't get points for pushing your body and mind as hard as you can, you get *good outcomes* from using your resources as *wisely* as you can. That usually entails stopping well before you drop each day, while steadily improving your capabilities.

Please treat yourself well today; doing so is an important component of long-term productivity.

Rest in motion

This is a linkpost for https://mindingourway.com/rest-in-motion/

Many people seem to think the 'good' state of being, the 'ground' state, is a relaxed state, a state with lots of rest and very little action. Because they think the ground state is the relaxed state, they act like maintaining any other state requires effort, requires suffering.

This is a failure mode that I used to fall into pretty regularly. I would model my work as a finite stream of tasks that needed doing. I'd think "once I've done the laundry and bought new shoes and finished the grocery shopping and fixed the bugs in my code and finished the big refactor, everything will be in order, and I'll be able to rest." And in that state of mind, every new email that hit my inbox, every new bug discovered in my code, every tool of mine that wore down and needed repair, would deal me damage.

I was modeling my work as finite, with the rest state being the state where all tasks were completed, and so every new task would push me further from that precious rest state and wear me down.

But the work that needs to be done is not a finite list of tasks, it is a neverending stream. Clothes are always getting worn down, food is always getting eaten, code is always in motion. The goal is not to finish all the work before you; for that is impossible. The goal is simply to move *through* the work. Instead of struggling to reach the end of the stream, simply focus on moving along it.

Advertisements and media often push the narrative that the purpose of all our toil is to win a chance at relaxation. We're supposed to work hard at boring jobs in order to earn our vacations. We're supposed to work hard for decades so that we can retire. (We're supposed to conceive of heaven as a place where nobody does anything except lounge on clouds.)

I call bullshit. For almost everybody, inaction is *boring.* That's why we pick up books, go exploring, and take up hobbies. The ground state is an active state, not a passive one.

The *actual* reward state is not one where you're lazing around doing nothing. It's one where you're keeping busy, where you're doing things that stimulate you, and where you're resting only a fraction of the time. The preferred ground state is not one where you have no activity to partake in, it's one where you're managing the streams of activity precisely, and moving through them at the right pace: not too fast, but also not too slow. For that would be boring.

And yet, most people have this model of the world where whenever they're not resting, they're taking damage. When the homework isn't done, they're taking damage. When they're reading a textbook, they're taking damage. When they go to sleep with work unfinished, they're taking damage. When they're at a large social event, they're taking damage. Some part of them yearns to be in the rest state, where they don't need to do all these *things*, and insofar as they aren't, they're suffering a little.

This is a grave error, in a world where the work is never finished, where the tasks are neverending.

Rest is not a reward for getting through all your obligations. You already dropped your obligations, remember? Rather, rest (and personal health, and personal time) are part of the goal. Both because most people care about their personal comfort, and because taking care of yourself is very important in order to do all the other things you want to do.

Rest isn't something you do when everything else is finished. Everything else doesn't *get* finished. Rather, there are lots of activities that you do, some which are more fun than others, and rest is an important one to do in appropriate proportions. Disconnect your impulse to rest from whether or not the world is in a stable state, because, spoiler alert, the world isn't going to be in a stable state for a long time.

Rest isn't a reward for good behavior! It's not something you get to do when all the work is finished! That's finite task thinking. Rather, rest and health are just two of the unending streams that you move through.

Imagine the person who is tight on money and needs to buy groceries once a month. Imagine that they agonize over every purchase, even though they know that they're buying as little as they can in order to secure the health of their family. You might suggest to them that they stop fretting over individual purchases and come to terms once and for all with the fact that food is a necessary purchase, and suggest that they fret over their *budget* instead. That way, they won't need to suffer every time they enter a grocery store.

The same technique applies to effort. You don't need to suffer every time it's time to do the laundry. Stop looking at the individual tasks, and start looking at the streams of work, some of which you can widen and some of which you can narrow.

Look at *all* the streams you want to move through, assess how much bandwidth you have available, and then simply move through the streams at the appropriate clip. Some streams will be unpleasant (chores, etc.), some will be basically mandatory (making money, etc.), some will be quite fun (learning, exploring, relaxing, etc.), and some of the most important streams are the meta streams (improving your capacity, finding better ways to fulfill your needs, etc.). But in all cases, simply see the streams and then move along them.

Many people I meet seem to think that they need to take damage whenever they're working, and then only heal it when they rest. While they're studying, they're taking damage. While they're at a large social event, they're taking damage. While they're doing their job, they're taking damage. They seem to think they "should" be able to be at home doing nothing, and so when they're not, they're taking damage. They think that the ground state is a resting state, a state of inaction, and so whenever they're acting, this is a deviation from the default, and it requires effort to maintain.

I say, *the ground state is in motion.* The privileged state is not a frozen state. Most of us wouldn't want to just lie in bed doing nothing forever, anyway. The easiest state to maintain isn't a motionless state, it's the state where you're out there *doing what needs doing* at a sustainable pace. *That's* the ground state, that's the state that requires no effort to maintain. Anything less leads to boredom, and it's *boredom* that's taxing.

I think one of the reasons people think high productivity is hard is that they think of lying in bed doing nothing as the default state, and anything else as taking damage. But it's not. It's really not. We were built to *move*, and we have things to do.

Make sure you're not taking damage just for moving. If any state of being is going to wear you down, then I suggest that you feel pressure whenever you start to move too fast *or* too slow. Take damage when your life is too boring and nothing's getting done, and take damage when your life is moving at an unsustainable pace: but don't take damage when you're moving through the streams at a steady clip.

The default state, the effortless state, is the one where you're moving along many streams. It is up to you to make sure that you're prioritizing the right streams and that you're steadily increasing your throughput, but the end goal is not to cease moving. Total inaction is dreadfully boring.

The ground state, the state to aspire to, the healthy state, the state that occurs naturally when you aren't forcing yourself to do anything, is the state where you're getting done what you want done as fast as is sustainable, and no faster.

The ground state is in motion.