Value Lock-in Notes

Author: C. Jess Riedel
Released: July 25, 2021

This document[1] is an incomplete collection of notes and arguments concerning the possibility of the values determining the long-run future of Earth-originating life getting "locked in". The author especially thanks Nick Beckstead, Paul Christiano, Wei Dai, Luke Muehlhauser, and Carl Shulman for valuable input. This incomplete document is being made public in case others find it useful, but it is not endorsed by the author (or anyone else). If you find it useful or would like to see this document completed, please consider letting me know at jessriedel@gmail.com.

Summary

Consider this main claim[2]:

> *Given machines capable of performing almost all tasks at least as well as humans, it will be technologically possible, assuming sufficient institutional cooperation, to irreversibly lock-in the values determining the future of earth-originating intelligent life*.

If this claim is true, it has significant implications for the long-termist worldview. This document reviews existing evidence and arguments for and against the claim, and in particular is informed by conversations with Paul Christiano, Carl Shulman, and Wei Dai, though this document doesn't necessarily represent their views.

I present no novel ideas. This document is likely most useful to readers unfamiliar with, or who dispute, these points:

- Permanent value stability looks extremely likely for intelligent artificial systems that have digitally specified values – if such systems can exist – allowing for perfect copying and fault tolerance. However there is some reason to think that values will not be "hard-coded" but instead emerge at a more abstract level from lower-level computations, and such emergent values need not be fully defined by dutch-book arguments.
- Similarly, it is more difficult to rule out value instability with high confidence in artificial systems that are fundamentally analog (e.g., analog neuromorphic chips).

---

[2] See below for definitions of the terms in this summary section.

- The manner in which digital fault tolerance (e.g., error correction) contributes to increased value stability in intelligent artificial systems (relative to humans) is closely connected to the orthogonality thesis.

Ultimately, I consider the main claim to be quite likely (~84%) and conclude that its failure to hold would imply fairly surprising facts about the nature of intelligent systems.

Sketch of the argument

My argument proceeds as follows: Goals/values are one of the most effective means for understanding a world that is populated by one or more intelligent agents. Human behavior can be predicted in terms of values only crudely, but artificial intelligent systems could possibly be very well described as pursuing explicit goals. Unlike biological systems, artificial intelligences could use error correction to ensure their goals are very stable over arbitrarily long times. Relative to humans, artificial intelligences could also be more resilient to harm, project power further, and coordinate more tightly across vast distances. If such systems acquire global influence, the future evolution of the world (and accessible universe) could be largely dictated by these stable goals. Though not necessarily a likely outcome, I consider it at least as possible that this occurs as the chance of nuclear winter from large-scale nuclear war this century. This suggests it is worth preparing carefully for the development of such intelligences due to vast and irreversible consequences.

Cruxes

Here are some key questions on which the main claim may hinge:

- Is it possible for an AGI to have stable (non-trivial and non-universal[3]) goals in the sense that its behavior can be thoroughly explained as pursuing these goals given the resources available?
    - Or conversely, must *all* sufficiently intelligent systems behave similarly on long enough time scale, or exhibit behavior that defies explanation in terms of goals?
- Is it possible for the goals of an AGI to be encoded in digital information (whether explicitly as a bit string or implicitly in software)?
    - Or conversely, must the goals of an AGI *necessarily* depend on non-digital input?
- Given an AGI with stable values but with many of the same cognitive limitations as humans, is is possible to establish a singleton composed of many copies of the AGI?
    - Or conversely, are the coordination frictions between AGI copies sufficient to prevent the creation of a stable singleton even if the other various human limitations – unstable/misaligned values and biological fragility – are eliminated?

[3] Here we mean that those goals are not universally shared by all AGI with stable goals. (It is occasionally argued that all sufficiently intelligent agents will converge on universal goals e.g., a "universal moral internalism" that is contrasted with the orthogonality thesis.)

- Can an AGI with stable goals maintain coordination over interstellar distances assuming it can do so over a planet?
    - Or conversely, do speed-of-light constraints over interstellar distances, or sheer volume considerations, introduce novel difficulties with singleton stability?

I claim the answer to all these top-level questions is “yes, probably”.

Preliminaries

In this preliminary section, I lay out the motivations, list key definitions, delineate my scope, and describe the structure of this document.

Motivation

> *Summary: Irreversible lock-in of the values pursued by intelligent being might be the highest leverage event in the history and future of the universe. The in-principle technical feasibility of such an unprecedented and initially implausible event is an important premise to assess for the long-termist worldview .*

My starting point is the long-termist worldview, under which the moral import of our actions are overwhelmingly driven by their impact on the far future. The philosophical basis for this position is laid out by Bostrom[4], Beckstead[5], and MacAskill[6], among others.

Anticipating the impact of our actions on the far future is very hard. The most reliable and powerful principles for analysis are likely

- *Physics*: The basic constraints imposed by the known fundamental laws of physics, plus some robust corollaries in areas like chemistry, information theory, and astrophysics.
- *Evolution*: The broadest implications of Darwinian evolution (when life is present), especially with regard to growth and equilibrium.[7]
- *Optimization*: The goals/values of powerful intelligent agents (when they are present); this can sometimes allow us to predict in which direction the world will be steered even if we don't know how it will be accomplished.

An important observation for understanding the last century or two of human history is the escape from the Malthusian trap: human behavior, originally selected for maximizing reproductive success in the ancestral environment, are currently driving significantly smaller population growth than resources allow. The failure of evolved creature to be maximally fit following a rapid change in their environment is by no means uniquely caused by human

---

[4] Nick Bostrom, "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards" [PDF], *Journal of Evolution and Technology*, Vol. 9, (2002). Nick Bostrom, "Existential Risk Prevention as Global Priority" [PDF], *Global Policy*, Vol 4, Issue 1, p. 15-31 (2013).

[5] Nicholas Beckstead, "On the Overwhelming Importance of Shaping the Far Future" [PDF] dissertation (2013).

[6] William MacAskill, forthcoming.

[7] For long-term predictions from evolutionary considerations, see for instance Robin Hanson, "Chapter 8 : The Rapacious Hardscrapple Frontier", [PDF] in *Year Million: Science at the Far Edge of Knowledge*, pp. 168-192, ed. Damien Broderick, Atlas Books, (2008).

intelligence,[8] and it is still very possible that evolutionary considerations will dominate in the long-term. However, the comparative speed and coordination of intelligence in thwarting slower evolutionary process (e.g., through contraception) raises the possibility that the goals of intelligent agents could surpass evolution as informative principles for predicting the future.

Indeed, there are compelling reasons to think the development of intelligent machines, on par with or exceeding the capabilities of the human mind, would dramatically reshape the world and irreversibly shift the balance of power from evolution to the goals of intelligent beings[9] (though see counterarguments[10]). Moreover, Bostrom, Yudkowsky, Christiano, and others have argued that this is likely to happen this century. The main claim of this document – that the values dictating the future of Earth-originating life could jettison evolutionary pressures and become permanently fixed – figures crucially into these arguments. The events leading up to and influencing such an irreversible development would have profound implications for the distant future,[11] and so might give extreme leverage to long-termist interventions.[12]

As further discussed in the subsection on scope, I do not attempt to assess here whether this lock-in scenario is in fact *likely*. I focus only on whether it is *possible*.

Analogously, when weighing the threat posed by nuclear weapons to humanity, one might first ask whether there are enough nuclear weapons on the planet to exterminate the species *even in principle*, e.g., assuming an extreme scenario in which the Presidents of the United States and Russia went crazy and conspired together to cause maximum damage. The advantage of this approach is that it brackets many complicated political issues and allows one to rely on comparatively more robust principles of physics and ecology. Afterwards, with the maximum possible damage established, one might then assess more likely futures like accidental nuclear war.

For similar reasons, gathering existing evidence and arguments for the main claim is an important first step in checking the robustness of the idea that longtermists should work hard to positively influence the development of machine intelligence.

---

[8] For instance, species that manage to colonize extremely isolated islands can stay at relatively low levels of fitness *on* those islands for millions of years, as evidenced by the success of invasive species later brought by humans.

[9] Eliezer Yudkowsky, “Intelligence Explosion Microeconomics” [PDF], manuscript (2013).

[10] Robin Hanson, “Chapter 8 : The Rapacious Hardscrapple Frontier”, [PDF] in *Year Million: Science at the Far Edge of Knowledge*, pp. 168-192, ed. Damien Broderick, Atlas Books, (2008). Robin Hanson, *Age of Em* [URL], Oxford University Press (2016).

[11] Paul Christiano, “Machine intelligence and capital accumulation” [URL], blog post (2014). Paul Christiano, “We can probably influence the far future” [URL], blog post (2014).

[12] This paragraph needs many more citations at multiple places.

Definitions

> *Summary: I define the following important terms:* narrow*,* general*,* super-*, and* artificial intelligence*;* agents*,* instrumental *and* terminal goals*, and* values*;* fault*,* fault tolerance*,* bit flip*, and* digital error correction*;* singleton*,* unipolar*, and* multipolar*;* value stability *and* drift*, and* global value lock-in*; and the* orthogonality thesis*.*

Readers can safely skip to the next section if they are familiar with the terms in the above summary.

General intelligence and superintelligence

Narrow intelligence is a capacity to complete tasks in a certain restricted domain[13], like chess, especially when that involves a strong degree of optimization, i.e., reaching a small target in a large space of possibilities. **General intelligence** is[14] "a central cognitive capability that lets humans learn a huge variety of different domains without those domains being specifically preprogrammed as instincts" (in contrast to animals). In our usage, intelligence is merely the ability of a system to complete tasks (e.g., optimize), and does not necessarily imply self-awareness, consciousness, moral patienthood, moral culpability, or other features that might be associated with intelligence in humans.

Artificial intelligence (**AI**) obtains when non-biological systems behave intelligently, and may be either narrow or general. **Superintelligence** is[15] general intelligence that obtains "strongly superhuman, or else at least optimal, [performance] across all cognitive domains" (or essentially all cognitive domains).

Artificial general intelligence (**AGI**) can in principle be inferior to humans on some tasks, but in this document I will reserve it for systems that are at least as capable as humans in essentially all domains. (However, I do not assume it is superintelligent unless stated explicitly.) **Agents** are physical systems exhibiting general intelligence, including both AGIs and humans.

Goals and values

Insofar as an agent's behavior can be understood as optimizing something to reach a particular target, that target is said to be (one of) the agent's **goals**. The meaningfulness of ascribing goals to particular agents will be discussed further in the subsection on expected utility maximization.

---

[13] See here for more precise definitions of words like 'task' and `human-level AI'.
[14] Here we use the definition of general intelligence from Arbital. This is probably the most vague definition in this document, and mostly amounts to pointing at the thing humans are doing differently from animals.
[15] Here we use the definition of superintelligence from Arbital.

When the agent pursues a goal only *conditional* on its usefulness to reach another goal, I say the first goal is **instrumental**. If not – that is, if the agent pursues the goal unconditionally – I say the goal is **terminal**. I will refer to an agent's terminal goals as its **values** for short. For agents without well-defined goals[16], such as humans, "values" can be generalized to the terminal goals an agent "should" adopt according to some proposed prescription, e.g., an equilibrium reached after substantial philosophical reflection.

Fault tolerance and error correction

When a component of a machine fails to behave as desired (e.g., because the component was poorly manufactured, or because of unusual outside interference), this is a physical **fault.** When a machine is capable of correctly performing its intended task even in the presence of faults, it is said to be **fault tolerant** with respect to that task and the set of possible faults. In the special case of information processing machines like digital computers, a particularly notable form of fault is a **bit flip**, when the physical component for representing one bit of memory (0 or 1) erroneously changes state (to 1 or 0). **Error correction** is a class of strategies for encoding data into memory such that the data remains uncorrupted with high confidence even in the presence of a limited number of bit flips; it is a crucial technique for building fault tolerant computers.

An introduction to the basic principles of error correction appears in the next section.

Multipolar, unipolar, and singleton

I use the term **singleton** as originally introduced by Nick Bostrom[17]:

> *[A singleton] refers to a world order in which there is a single decision-making agency at the highest level. Among its powers would be (1) the ability to prevent any threats (internal or external) to its own existence and supremacy, and (2) the ability to exert effective control over major features of its domain...A democratic world republic could be a kind of singleton, as could a world dictatorship. A friendly superintelligent machine could be another kind of singleton, assuming it was powerful enough that no other entity could threaten its existence or thwart its plans.*

Future scenarios featuring a singleton are **unipolar**, and those that don't are **multipolar**.

Some authors appear to refer to scenarios as multipolar so long as they feature vigorous economic competition between agents, even in cases where there is a central authority enforcing ground rules.[18] In contrast Bostrom's notion of unipolar appears to be satisfied so long as agents can summon the minimal cooperation to avoid the most destructive outcomes in the

---

[16] Here we are following Eliezer Yudkowsky's use on Arbital (link).
[17] See first page of Nick Bostrom, "What is a Singleton?" [URL], *Linguistic and Philosophical Investigations*, Vol. 5, No. 2 (2006): pp. 48-54.
[18] For instance, see Katja Grace's discussion of multipolar scenarios on Less Wrong.

context of his “vulnerable world hypothesis”.[19] For my purposes, the important distinction is between those values that are subject to competitive (Darwinian) selection, and those that are robustly preserved through coordination of the intelligent agents. These later values are the ones that are potentially locked in.

Value stability and lock-in

If the values of an intelligent system do not change with time, those values are said to be **stable**. Otherwise, the system undergoes value **drift**, which may be random or predictable. (Depending on how human values are defined, examples of drift might include increased risk aversion in older people.) If there is an AGI singleton with stable values I say those values are **locked-in**.[20] I sometimes characterize this as **global** to emphasize that it applies to all Earth-originating life.

Orthogonality

The **orthogonality thesis** is this assertion[21]:

> *Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal.*

For defense and criticism of the thesis, see Bostrom[22], Yudkowsky[23], Armstrong[24], and <insert others>.

---

[19] Nick Bostrom, “The Vulnerable World Hypothesis” [PDF], working paper (2018).
[20] We generally assume an AGI singleton is powerful enough to avoid extinction or other disruption to the pursuit of its values with high probability. Per Bostrom’s definition, an AGI that encountered uncooperative aliens (which might pose a threat) would not qualify as a singleton.
[21] See page 3 of Bostrom’s “The Superintelligent Will: Motivation and Instrumental Rationality in Advance Artificial Agents.” [PDF], Minds and Machines, 22(2), 71-85 (2012). Likewise, Eliezer Yudkowsky on Arbital (link) defines it as the assertion that “there can exist arbitrarily intelligent agents pursuing any kind of goal”.
[22] Nick Bostrom, “The Superintelligent Will: Motivation and Instrumental Rationality in Advance Artificial Agents.” [PDF], Minds and Machines, 22(2), 71-85 (2012). Nick Bostrom, “Superintelligence” [URL], (2014).
[23] Eliezer Yudkowsky, “Artificial Intelligence as a Positive and Negative Factor in Global Risk” [PDF], p. 308 in *Global Catastrophic Risks*, ed. Nick Bostrom, Milan M. Ćirković, Oxford University Press (2008).
[24] Stuart Armstrong, "General Purpose Intelligence: Arguing the Orthogonality Thesis." [PDF] *Analysis & Metaphysics*, 12 (2013).

Scope

Summary: I delineate the scope of this document. In particular, this document does not attempt to address the physical limits to intelligence, AGI take-off speed, unipolar vs. multipolar scenarios, the practical feasibility of lock-in, the imminence of AGI, or any normative questions.

In this document, I don't try to answer the question of how likely it is that global value lock-in *actually* occurs in the future. To see how this differs from our main claim, consider a widely discussed global lock-in scenario that can be decomposed into these three parts:[25]

1. It is technologically feasible for people to build an AGI with stable goals.
2. Due to factors like recursive self-improvement and hardware overhangs, the first AGI will become much more powerful than any humans or other (near-AGI) systems that potentially oppose it. If that AGI desired to take control of the world, it would.
3. If an AGI with stable values takes control of the world, it will be able to maintain control indefinitely if it desires.

The second item is an extremely complicated assertion.[26] It depends sensitively on many contentious technical and economic issues related to whether AGI capabilities will accelerate rapidly after achieving parity with human on some key tasks, a scenario known as the *intelligence explosion* or "AI-Foom".<cite> It is also affected by social and geopolitical issues regarding the feasibility of such an agent gaining a *decisive strategic advantage* over competing human institutions.<cite> I will not attempt to address any of these issues in this document.

Rather, this is only an *equilibrium* analysis; I try to determine whether there exists a possible stable state of the universe, at some point in the future, characterized by global value lock-in. I will assume, when necessary, that any AGI being discussed is able to initially reach a position of great power, e.g., by being purposefully constructed by a totalitarian dictator.

More exhaustively, I do not take a position on these considerations:

- *Physical limits to intelligence*: Once artificial systems become at least as good at humans at all tasks, can they eventually become superintelligent, i.e., much better than humans (or optimal) at essentially all of those tasks?

[25] Of course, even if our second item is false, there are other ways an AGI might eventually obtain global power.

[26] The resulting future would look markedly different from a "star wars" universe of many competing factions continuously bubbling up, merging, and dissolving, which is a popular default vision of the future obtained by extrapolating human history to an era of interstellar colonization.

- *Take-off speed*: If superintelligence is technologically possible (i.e., if humans are far less intelligent than allowed by physical limits), how quickly will this be achieved following the arrival of AGI?
- *Unipolar vs. multipolar scenarios*: Given facts about take-off speed and world institutions, will a singleton AGI with stable values arise? Or will the powers of intelligent actors remain sufficient balanced to ensure competition even as life expands beyond Earth?
- *Imminence*: When might global lock-in occur?
- *Wisdom*: Would the permanent global lock-in of values be a moral disaster? Or is it necessary to prevent disaster?

Document structure

The rest of the document has three sections:

- First, I describe the two key considerations for machine intelligence that will drive my main claim: fault tolerance and agents goals.
- Subsequently, I argue that these considerations make it plausible that it's possible to build an AGI with stable values.
- Finally, I argue that, given the existence of an AGI with stable values, it would be *technologically* possible to lock-in the values steering the future of humanity and its descendants.
- In the appendix, I give an account of my process for writing this document, the next steps I would take to improve it, and an Error correction example.

Most sections and subsections of this document will begin with a short summary of their contents. These function as local abstracts to aid the reader in consuming the document efficiently.

Subsections marked with an asterisk are incomplete in the current version of this document. They contain incomplete sentences, non-sequiturs, etc. Reader is advised to just read the subsection summary.

Key considerations

Summary: I give important background information about two semi-technical ideas that are crucial to my argument: (a) digital error correction and (b) goals of intelligent agents.

Consider the following methods by which agents might influence the values pursued in the future:[27]

1. Ancient hunter-gatherer parent advising their children.
2. 18th-century farmer sending their children to formal schooling.
3. Near-future service worker selecting an embryo with a polygenic score associated with high conscientiousness.
4. Same as above, except using more advanced genetic engineering to design a particular cognitive profile.
5. Future businessman achieves longevity escape velocity, and maintains his fraction of total world wealth in perpetuity.
6. Future adolescent prodigy destructively scanned to become immortal em template for notable fraction of total em population.
7. Future authoritarian ruler of China gets decisive strategic advantage and creates em totalitarian state.
8. World government creates robustly stable AGI singleton with hard-coded (digitally represented) values.

In the first example, the hunter-gatherer parent lives in the ancestral environment for which humans were evolved, and essentially acts as a tool of evolution to further the goal of improved fitness. If the parent gives bad advice, the child may be slightly less fit, but if so they could be out-competed, in expectation, by a child whose parent gives better advice. In this sense, the evolutionary (quasi-)equilibrium continues, and the content of any particular piece of advice has little causal influence on the values of the future. Values drift, but in a way dictated by evolution and random chance.

The other examples exhibit progressively stronger influence by the current generation on future values. The last (eighth) example – if actually possible – depicts a world in which values stop drifting and become locked-in basically forever (say, until the universe ceases to support intelligent life). Note, though, that this list is not ordered by the rate of value change. Values are changing much faster during the industrial era than they did in the ancestral environment, and than they would in the case of lock-in. It's obvious that permanent lock-in requires something more than merely the interference of agents with the goals of evolution.

[27] I thank Paul Christiano for suggesting this framing in terms of a spectrum.

Many have argued that the jump from present-day biological intelligent systems (humans) to artificial general intelligence opens up the technological possibility of global value lock-in. This would mean that if all possible resistance were temporarily suppressed (say, by a totalitarian regime), then an AGI system could be created that would pursue some fixed set of values forever. In this document, I emphasize that, although not always stated explicitly, the key considerations that open this up as a technical possibility are

- the creation of intelligent systems whose values are ultimately determined by (or generated by, or unfolding from) discrete information, i.e., digital data and software that is untethered from specific hardware; and
- the perfect copying of digital information, or more generally the technique of digital fault tolerance.

As we will explain more thoroughly, the crucial importance of perfect copying can be seen in future ultra-stable em-society scenarios, which are among the scenarios least sensitive to unknown facts about AGI, where refreshing from identical copies of ems plays a key role.[28] Locking in values by making them fault tolerance is just a slightly more sophisticated implementation of this idea.

In the remainder of this section I give brief introductions to fault tolerance and to agent goals.

Redundancy and fault tolerance

Summary: I introduce some basic facts about computational redundancy and fault tolerance that will be relevant for my argument. I discuss the extent of error correction in existing artificial and biological systems, but do not actually consider AGI or the main claim.

Introduction[29]

At the platonic level, a computation is a sequence of mathematical operations on numbers. In order for a physical device to perform a computation, it encodes those numbers in physical degrees of freedom, like the position of an abacus bead or the amount of electrical current flowing through a wire.

A computing machine is ultimately made of atoms, and its behavior is understood and predicted with the use of mathematical models that summarize the enormous number of possible microscopic configurations with a much smaller set of parameters that describe the machine up

[28] Carl Shulman, "Whole brain emulation and the evolution of superorganisms" [PDF], Machine Intelligence Research Institute, working paper (2010). Robin Hanson, *Age of Em* [URL], Oxford University Press (2016).

[29] Some text in this section has been adapted from a report on quantum computing I drafted in 2017 in coordination with Carl and Luke.

to some level of resolution. Some of those parameters will encode the numbers of the platonic computation. The model also includes (possibly probabilistic) rules that specify how these parameters are expected to change, mostly based on observed empirical regularities. As the microscopic configuration of the machine evolves in time according to the laws of physics, a good model will predict parameter changes that correspond to the actual microscopic configuration with very high likelihood, and hence a correct physical instantiation of the computation.

However, all models are imperfect, and there will always be cases where the model breaks down, i.e., where the machine evolves such that its true microscopic configuration no longer corresponds to the parameters predicted by the model. This is known as a fault, and it may be due to foreseeable or unforeseeable effects**.** As machines become more complicated, requiring an ever larger number of parameters to usefully describe, the chance that *all* parameters correctly track the physical configuration of the machine usually becomes small. In the cases where N possible faults are statistically independent and occur with probability p, the chance that no fault occurs is $(1-p)^N$, which, for fixed non-zero p, is driven exponentially to zero by large N.

The simplest form of fault tolerance is global (end-to-end) redundancy. For instance, if an abacus is a bit shaky and the beads tend to slide out of position from time to time, we might implement a long computation by doing it simultaneously on three abacuses. If at any time we notice that a bead on one abacus is configured differently than the other two, we reason that the minority device is likely to have experienced a fault and we correct it by adjusting it to be identical to the others. Even though it means performing the computation three times and thus exposing the calculation to a larger absolute number of possible faults, this increases the reliability of our overall calculation since now *two* faults must occur, each in a different computational path, in order for the majority vote to be erroneous. More generally,[30] if R identical computations are run with fault probability p, with R odd, the chance that a majority of the computational paths give the incorrect answer scales[31] roughly like $p^{(R+1)/2}$.

The cost of this simple strategy is to increase computational resources linearly in the redundancy R, but more sophisticated fault tolerance mechanisms exist with a much smaller overhead that nevertheless assure that the chance of an uncorrected fault is negligible. The simplest one is described as an Error correction efficiency example in an appendix. The bottom line is that, unlike mechanical processes, digital information processing can perfect, allowing for trillion-year Jupiter-brain computations that have negligible chance of making a mistake.

However, the “digital” qualifier is key, and analog processing is not similarly robust, i.e., the information must be discrete like an integer rather than continuous like a physical length.

---

[30] The Space Shuttle used five separate computers, allowing for two independent faults without a safety-critical failure. See for instance Computers in Spaceflight: The NASA Experience.

[31] Formally this is something like (R choose $(R+1)/2$)$p^{(R+1)/2}$ plus higher order corrections in p, but the specifics aren’t important.

Low-overhead fault tolerance schemes can be very sophisticated, but they all rely on the ability to make identical copies of the information. No method is known for making identical copies of continuous quantities.

Statistical independence[32]

Besides discreteness, these very strong reliability claims rely crucially on our assumption that faults in separate locations are approximately uncorrelated, i.e., if the probability of fault at one place is p, then the probability that faults occur at two separate places is p^2. This is necessary for achieving exponentially low rates of uncorrected faults, so that (say) trillion-year computations are reliable. One can consider more general "error models", with corresponding more sophisticated fault-tolerance schemes, but they essentially all rely on low correlations between faults that are sufficiently well-separated in space and time.

Such assumptions are physically well motivated: if stray cosmic rays occasionally flip bits in your hard drive, the chance a bit flips on one side of your hard drive is usually independent of whether a bit on the other side flips. This is why the error correction on your hard drive can never actually drive the chance of an error below the rate at which a nearby gamma-ray burst will scramble *most* the bits all at once, or the rate at which people break into your house and smash your hard drive with a hammer; in those cases, the error model is breaking down because the errors in different bits are correlated.

Analogously, consider the protection against extinction provided by human colonization. Although colonization increases the total number of people, the much more important thing is that most risks to the population become decorrelated over long distances. This is just a way of formalizing the intuitively obvious idea of robustness: waving a magic wand to double the population of the Earth would not be nearly as effective for reducing existential risk as creating a second populated Earth-like planet around another star.

Current usage in artificial and biological systems

Modern computer memory has such low fault rates (on the order of 1 fault per 10 billion bit-hours)[33] that error-correcting codes (ECCs) are used in some but not all consumer products like hard drives. However, ECCs are used widely in communication equipment and enterprise applications that demand high reliability like finance.[34]

There's nothing intrinsic about silicon that allows error-correction, and it's natural to wonder if biological cells employ it to prevent the accumulation of errors as genetic information is passed from one generation to the next. Although most species (including all eukaryotes) have

---

[32] Some text in this section has been adapted from some comments I wrote on Holden's document "Transformative technologies and irreversible shaping of society".

[33] See Wikipedia: ECC memory and the abstract of Bianca Schroeder et al "DRAM Errors in the Wild: A Large-Scale Field Study" [PDF], *SIGMETRICS/Performance* (2009), which notes up to 70,000 errors per billion device hours per megabit.

[34] See, for example, "Enterprise-class versus desktop-class hard drives" [PDF], Intel Server Boards and systems (2016); Dan Luu, "Why use ECC?" [URL], blog post (2015); and Wikipedia: Error detection and correction.

mechanisms to *physically* detect DNA damage, this mechanism is imperfect and does not operate on the DNA as a pure sequence of T, C, A, and G symbols.[35,36] Indeed, research suggests that cells do not employ any error-correcting codes[37] that could drive error rates to essentially zero. As a result of this lack of perfect fault tolerance, of order 40 *de novo* mutations are introduced[38] in each new generation of human.[39] Later we will briefly discuss why this may be.

Goals of agents

> *Summary: I describe arguments for and against the meaningfulness of describing agents in terms of goals, and reasons this might continue to be true even for future AGI.*

In this subsection I assume the reader is familiar with the basic arguments that artificial intelligent agents will adopt a framework of expected utility maximization (EUM), as originally described by von Neumann & Morgenstern; Savage; Anscombe & Aumann; and Jaynes.<cites> A more recent and intuitive summary of some of these argument in the context of AI was given

---

[35] Mispairings during DNA synthesis can be corrected first by proofreading and then by mismatch repair, but these rely on *physical cues* (e.g., methylation) to distinguish the old (presumed correct) strand from the new (presumed erroneous) strand. Thus, this is not error correction in the information-theoretic sense.

[36] The choice of nucleotides corresponding to the A, C, G, T/U symbols of the coding alphabet can be very roughly interpreted as a sort of error check at the level of atoms, explaining (to some extent) why evolution uses these four species for DNA and RNA. See Dónall A. Mac Dónaill, "A parity code interpretation of nucleotide alphabet composition" [URL], *Chemical Communications* 18, 2062-3 (2002) and Tsvi Tlusty, "A colorful origin for the genetic code: Information theory, statistical mechanics and the emergence of molecular codes" [PDF], *Physics of Life Reviews* 7, 362-376 (2010). However, the strong correlations between atomic dynamics within a single nucleotide molecule prevents this from providing proper error correction.

[37] L.S. Liebovitch, Y. Tao, A. T. Todorov, and L. Levine, "Is there an error correcting code in the base sequence in DNA?" [URL], Biophysics Journal, 1996 Sep; 71(3): 1539–1544. This result is discussed on the first page of Manish Gupta, "The Quest for Error "Correction in Biology" [PDF], IEEE Engineering in Medicine and Biology Magazine 25, 46 (2006) who cites Liebovitch et al. and notes that their technique is "very basic", but whose survey of the literature also fails to find any examples of true error correction. Likewise, a quick skim of I.S. Mian and C. Rose, "Communication theory and multicellular biology" [URL], *Integrative Biology* 3, 350-36 (2011) did not turn up any examples. I could pin this down better, but it's not important enough right now.

[38] "...the average de novo mutation rate is 1.20×10^{-8} per nucleotide per generation": Augustine Kong et al., "Rate of de novo mutations and the importance of father's age to disease risk", [URL] *Nature* 488, 471–475 (2012). Given the 3.23 billion base pairs in the human genome, this "works out to around 38 mutations genome-wide per offspring", according to Dan Koboldt of the Institute of Genomic Medicine at Nationwide Children's Hospital.

[39] In plants and animals, most but not all of these mutations occur in non-protein coding regions of the genome, whose influence on biological function is difficult to detect. See Wikipedia.org: Non-coding DNA and the sections "Mutation rates in higher eukaryotes based on specific" and "Evolutionary forces shaping mutation rates" in J.W. Drake et al., "Rates of spontaneous mutation" [PDF], *Genetics* 48, 1667–1686 (1998).

by Omohundro.[40] I will, with a few exceptions, ignore the extent to which EUM is a richer and more precise concept than the intuitive idea of goals (such as the explicit and fundamental role of probability) since those difference do not seem crucial to the idea of value lock-in. Instead I concentrate merely on the arguments for why goals are thought to be useful for understanding (even superintelligent) AGI.

Explanatory and predictive power of goals

Earlier we said that, insofar as an agent's behavior can be understood as optimizing something to reach a particular target, that target is said to be (one of) the agent's goals. Here we begin by emphasizing these caveats:

- Claims about agents having goals have content only insofar as the goals are in some sense simple or reasonable. Trivially, any physical systems could be said to have the goals "execute behavior X_1 at time T_1, then behavior X_2 at time T_2, …" where the behaviors X_n and times T_n are taken from observation. (Similarly for statements about the system's putative utility function.)
- An agent's behavior may not be adequately explained by any set of reasonable goals, in which case the agent's goals are simply ill-defined. The degree of well-defined-ness can lie on a spectrum depending roughly on the parsimony of the explanation.
- Whether an agent acts in accordance with goals can be environment dependent. A given nematodes species, like most species, will be highly optimized to achieving the goal of proliferation in its normal environment. But experiments in artificial environments of course reveal that worms contain a lot of response-stimulus machinery which does pursue any environment-independent goal.
- An agent's observed behavior may be equally consistent with multiple goals, and so not uniquely defined by observation.

Of course, those caveats are to be weighed against these points, which are intended to go beyond the EUM theorems that hold under ideal assumptions:

- Despite the fact that goals might not never be defined with mathematical precision, they seem critical to understanding a future influenced by intelligent agents. We routinely reason about the (not mathematically precise) values of other humans to predict their behavior, often very successfully. Even when we employ non-goal-based explanations to predict the behavior of humans (e.g,, a salesman exploiting framing effects in his customer), these are arguably mere *perturbations* to a strongly goal-based predictive framework. Indeed, we seem to have no viable substitute for this sort of reasoning where it works.
- The cognitive machinery for seeking abstract goals seems to be resource-expensive for animals. The largest animals (e.g., over 10 kg), for whom a fixed amount of resource

[40] See the appendix of Stephen Omohundro, "The Nature of Self-Improving Artificial Intelligence" [PDF], manuscript (2007).

cost is a smaller percentage, tend to exhibit more sophisticated goal-explained behavior than the smallest animals (e.g., under 100 mg).[41] Thus we can weakly expect that agents constructed with access to plentiful resources can be more goal-driven than survive in the ancestral environment (including agents that can project power over longer ranges, giving them a larger effective size).

The above list is incomplete.

Goal specification

In light of the ideal EUM theorem and the above arguments, we have good reasons to think that AGI will be at least approximately, and perhaps very precisely, well-described as having goals. A somewhat vague but important crux driving a divergence of research approaches in AI safety is the degree to which these goals will be internally specified. This is a spectrum roughly stretching between these extremes for AGI design:

- *Digitally specified*: Goals are hard-coded in AGI software, and match up well with the (possibly hypothetical, not necessarily observed) revealed preferences of the AGI. This might arise (somehow) from a literal, explicit list of goals, or it might robustly emerge from lower-level mechanisms that don't look like goal-pursuit. But crucially, the thing that determines AGI goals are digital bits (software and data), and is fully independent of the outside world. AIXI is an infeasible example of the case of explicit goals.
- *Fully implicit*: Nothing interior to the AGI specifies its goals as defined by its revealed preferences. Its behavior emergently approximates goal pursuit (at least to an accuracy that's necessary in light of EUM theorems, efficiency concerns, and other considerations), but those goals depend on properties of the world outside the AGI.

Neither of these extreme poles are necessarily endorsed by any researchers, and indeed it's even disputable whether this is a meaningful distinction. (Internal goals may refer to external facts.) Still, the opposing directions are sometimes associated, respectively, with MIRI's "Embedded Agency" research agenda[42] and Paul Christiano's approach with "Approval-directed Agents"[43].

This spectrum appears important to the question of whether the existence of fault tolerance should give us strong confidence in goal stability, or whether we must rely on weaker arguments. MIRI researchers have emphasized the dangers of naive approaches to

---

[41] Just basing this on common sense for now. Would need to confirm and add cites in the future.
[42] For an accessible introduction, see the overview by Abram Demski and Scott Garrabrant (also available on LessWrong).
[43] Paul Christiano, "Approval Directed Agents" (presumed title) [URL], blog post (2014). Paul Christiano, "ALBA: An explicit proposal for aligned AI" (presumed title) [URL], blog post (2016)

hard-coding AGI utility functions,[44] if that is in fact possible. Christiano has argued that his approach (where goals are more implicit) may avoid issues with locking in misspecified values or other AGI design decisions.[45]

We emphasize that digitally specified values are not in conflict with an AGI systems using non-digital computational machinery, such as analog (e.g., neuromorphic[46]) or quantum chips, just as any AGI system will certainly using non-digital machinery for non-computational tasks like manipulating the world with robotic arms or storing bulk resources in imperfect containers. Indeed, this non-digital computational machinery might very well be crucial for an AGI system's continued existence (e.g., performing the computational necessary for maintenance, for collecting more resources, etc.). But the values pursued by the AGI are (implicitly) assumed to be *determined* by a digital core in the sense that these values are independent of the non-digital machinery, i.e., the values would not be perturbed if the non-digital hardware were replaced with other hardware that, by assumption, could only be the equivalent up to some finite level of resolution.[47] Otherwise, there would be an opportunity for value drift due to hardware decay, and arguments for being able to limit the degree of value drift would be substantially less air-tight in light of our current ignorance about the nature of AGI architecture.

Other considerations

Here are some related concerns that require more investigation:

- We expect agents to accumulate physical resources and knowledge continuously, and that data has to be encoded somewhere. For a useful notion of stable values, it seems necessary that these goals be kept separate (and bounded in size) from the resources and knowledge which increase essentially unboundedly. How obvious is it that such a separation *must* occur with intelligences?
- It is unclear how to interpret modern designs of artificial intelligent systems at evidence that AGI will be best described as having explicit goals. (And it depends in part on

[44] See for instance the introduction to Nate Soares, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong, "Corrigibility" [PDF] In AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, January 25–26, 2015.
[45] See section "Avoiding lock-in" in Paul Christiano, "Approval Directed Agents" (presumed title) [URL], blog post (2014) and the section "Contrast with the usual futurist perspective" in Paul Christiano, "Corrigibility", [URL], blog post (2017).
[46] Some examples: "We believe that this demonstration is an important step toward the effective analog-hardware implementation of much more complex neuromorphic networks – first of all, multilayer-perceptron classifiers with deep learning, and eventually also much more elaborate CrossNet-based cognitive systems." - M. Prezioso et al., "Training and operation of an integrated neuromorphic network based on metal-oxide memristors" [PDF], *Nature* 521, 61–64 (2015). Giacomo Indiveri and Rodney Douglas "Neuromorphic Vision Sensors" [PDF], *Science* 288, 1189-1190 (2000). Robert D. Hof, "Neuromorphic Chips" [URL] *Technology Review* (2014).
[47] I also note that the assumption of a digital core is compatible with an AGI re-writing its core, although in that case value stability would obviously depend on showing that its values survives the re-writing process unscathed.

controversial questions about how close modern systems are to AGI in terms of basic architecture.) Many modern systems have explicit goals, but many do not. Furthermore, the fact that much modern systems are described by goals could be evidence *against* explicit goals in AGI if you think modern systems are far away from general intelligence (and the conventional framing in terms of goals might be a reason why).

- Humans are plausibly the most goal-driven animal in the sense that they perform very complicated sequences that can often be explained in terms of the pursuit of abstract goals. However, it's widely argued that they consistently fail to be utility maximizers (under reasonable, non-question-begging definitions of utility).[48] Humans side-step EUM theorems by "patching" vulnerabilities (e.g., recognizing that they are being dutch-booked and declining to bet). Are there strong reasons, beyond elegance/simplicity, to expect that AGI will eschew patches and incorporate utility at a more fundamental internal level?
- Despite their key role in discussion of AGI, values and value stability remain frustratingly slippery concepts to say anything concrete about. As examples, here are a couple unusual scenarios that seem very unlikely intuitively but are taken seriously and interfere with attempts to discuss an AGIs goals:
    - Optimization daemons[49]: "If you subject a dynamic system to a large amount of optimization pressure, it can turn into an optimizer or even an intelligence.In the case of AGIs deploying powerful internal and external optimization pressures, we'd very much like to not have that optimization deliberately or accidentally crystallize into new modes of optimization, especially if this breaks goal alignment with the previous system or breaks other safety properties...When heavy optimization pressure on a system crystallizes it into an optimizer - especially one that's powerful, or more powerful than the previous system, or misaligned with the previous system - we could term the crystallized optimizer a "daemon" of the previous system. "
    - "[Suppose] value-stable AGI are possible, but only for values that are boring in some sense (e.g., paperclips). Perhaps after a lot of philosophical reflection we decide the best AGI to build is one that, it turns out, happens to be difficult to put in a goal-pursuing framework. For instance, the revealed preferences of this AGI might be something like "maximize how interesting I am, including my goals", and the function you'd need to maximize is complicated in some deep sense that frustrates any attempt to model it as an expected-utility maximizer."

[48] Behavioral economics is one notable field of study related to these observations.
[49] Eliezer Yudkowsky, "Optimization daemons" [URL], webpage

AGI value stability

In this section I argue for the technical feasibility of value stability in an AGI system assuming that an AGI can be built. (Establishing how such a system could be used to lock-in global values is addressed to the next section.)

Given the vast unknowns about such a hypothetical system, the most robust argument strategy has two main part:

- Begin with human capabilities as a reliable baseline. Human brains are clear proofs of concept about what sorts of generally intelligent systems are physically possible.
- Make a minimum number of assertions about a few capabilities AGI agents will have that exceed human agents with high confidence. These should follow from only modest assumptions regarding its non-biological nature, and not from more contentious claims about ways AGI might be vastly more capable than humans.

In other words, we are going to argue for AGI value stability by making the smallest possible jump from humans, who do not exhibit value stability. The argument is as follows, with each point corresponding to a subsection:

1. Instability of DNA and biological values: Human are only imperfectly described as pursuing values, and those values drift over time, but these properties are very likely contingent on the evolutionary origin of the human species rather than intrinsic to their general intelligence capabilities. The lack of perfect error correction in human genetics – but its inevitable future presence in AGI systems – makes it reasonable to consider the possibility that AGI values could be stable despite the instability in human values.
2. Predictability of biological versus artificial systems: It is routinely demonstrated that, within a given narrow cognitive domain, artificial systems with narrow intelligent can produce much more reliable output than humans. This is further evidence that human unreliability in cognitive tasks is not an intrinsic property of human intelligence, and can be avoided by machines that perform those tasks.
3. Nearly digital values: Unlike individual pieces of physical hardware that may decay, digital information can be made essentially eternal through the use of redundancy and – more generally and with greater efficiency – fault tolerance. Any method that can anchor an AGI system's values to digital information is automatically a method for suppressing value drift. Although it is conceivable that there exist surprising facts about general intelligence that prevent it from anchoring to digital information to arbitrarily high accuracy, none are currently known and circumstantial evidence suggests they do not exist.
4. AGI effective immortality: AGI systems with digital values can be perfectly copied and hence propagated into the future indefinitely. Thus AGI systems are not tied to physical

hardware that may degrade in time. It also means that, if necessary and desired, such systems can avoid even speculative sources of value drift by performing complete resets to initial configurations.

These points together lend relative confidence, conditional on the existence of AGI, to the technical feasibility of AGI systems with stable values over arbitrarily long time scales.

I now argue for each of these points in turn. (Readers who accept any given point can skip the corresponding subsection.)

Instability of DNA and biological values

Summary: DNA mutations contributes to value drift in human society even if all other sources of instability were to be eliminated, and in principle could be eliminated with full DNA error correction. This source of drift may be unavoidable for biological agents in the ancestral environment, but is unnecessary for an AGI.

Earlier I noted that fluctuations in human values in the ancestral environment have little long-run impact; their influence tends to decay over a few generations as the environment applies selective pressure to individuals. In addition to their genes, humans influence future generations through some accumulation of cultural information, but most of this transmission is very noisy. Even after the development of agriculture, humans had few institutions that could preserve substantial amounts of information.

The human genome is about three gigabits in size, and the vast majority (~99.8%) of it is shared by all members of the species.[50] It is the closest humans have to a preserved source (or seed) of values, but as mentioned it nevertheless accrues mutations each generation. Since the machinery necessary to implement error-correcting codes (ECCs) does not seem out of reach for biological cells in light of the vastly complicated machinery those cells are known to employ on microscopic levels, and since there is selective pressure on genes to propagate themselves, one might wonder whether the lack of ECCs for the genome – the most tightly conserved source of values – is somehow necessary for general intelligence. After all, humans are our only existing example of general intelligence.

The favored explanation is not surprising: although a species that perfectly suppressed all mutations would avoid the (much more common) fitness-decreasing mutations, it would be forever cut off from rare fitness-increasing mutations and would eventually go extinct.[51] The

[50] "...variations were differences of a single letter in the approximately three billion letters that make up a person's DNA...The typical person's genome differs from the agreed-upon "reference human genome" in 4.1 million to five million places". Francie Diepsep, "Scientists quantify how different humans are from each other genetically" [URL], *Pacific Standard* (2015).

[51] This paragraph would require additional research or expert interviews to be thoroughly convincing, but I'm sufficiently confident in the answer that it's not worth the time investment.

natural environment appears to apply selective pressure toward some optimal mutation rate that low but not zero.[52]

Although it's conceivable that this sort of reasoning might also apply to an AGI system,[53] I find it unlikely. The key difference seems to be that once intelligence appears, it can purposefully adapt to threats without relying on random mutations and the resulting risk to its values; adaptation and value drift can be separated and only the former preserved. This makes it more plausible that values of future artificial systems could be perfectly preserved through error correction indefinitely even though we do not find this feature in biological humans.

*Predictability of biological versus artificial systems

> *Summary: The fact that contemporary (non-generally-intelligent) artificial systems, within their domain, are more reliable and predictable than biological systems is weak evidence that AGI systems could have stable values even though humans do not. It looks very possible, in principle, to significantly increase stability of values in individual humans with sufficient effort.*

<Incomplete>

Here I briefly draw attention to the intuitive idea that machines built by humans can often be more predictable and stable than biological systems performing a similar role. <Insert examples.> To be sure, this is partially because artificial systems are simpler than essentially all biological systems. But it is also because humans designers often *desire* a degree of predictability (e.g., to simplify planning) that goes well beyond the simplicity imposed by their limited designing abilities.[54] This is especially true of artificial systems that process information (computers) which can exhibit much higher degrees of predictability than manual human calculations.

It's not that biological systems *must* be unpredictable or unstable, only that the ones in the real world are. It seems quite possible that biological systems could be modified to increase predictability (e.g., by adding additional DNA error correction), and these might function perfectly well for quite long timescales (although they might be less adaptive as a species on the very long timescales of evolutionary importance).

---

[52] This citation is not a complete justification of this claim, but it is one starting point: John W. Drake et al., "Rates of spontaneous mutation" [PDF], *Genetics* 48, 1667–1686 (1998).

[53] For instance, an AGI might encounter a threat to which it can't adapt to while maintaining its values, such as a more powerful alien species that has diametrically opposed goals and refuses to risk co-existence with human-descended intelligent agents. But in that case it means humans goals would be doomed anyways, so it seems moot from an AGI design standpoint. Maybe there are other scenarios.

[54] Even very simple physical systems can behave chaotically, and any Turing machine can behave in a way that can only be predicted by running its code.

Reduced value drift as humans age

The human brain is a clear demonstration that compact physical systems can exhibit general intelligence,[55] but it is not a clear demonstration that generally intelligent agents have stable values. <Behavioral economics, bounded rationality, etc.>

Examples of predictable human values drift:

- Entering college
- Having children
- Increasing risk aversion between adolescents and middle age.
- < Changing religion probably not a good example since this could be just updating on new info>

<But note dependency on the definition of values>

<Highly speculative> Within the spectrum of values/goals that humans pursue, the goals of humans -- insofar as they are well-defined -- do drift over a lifetime, but the speed of change is commonly said to slow down. (Certainly, the willingness of humans to change opinion on topics goes down as they age, but of course we expect a rational AGI to never stop correctly updating on new evidence about empirical questions, and it's very plausible we expect, insofar as this is a sensible thing for AGIs to do, that it change its mind on moral questions in response to good ethical arguments.)

The speed at which the values of a human drifts over a lifetime appears to be due to specific details about human biological lifetimes rather than anything about the timescale on which human brains reasons (much less anything intrinsic about universe) That is, the ratio of reasoning (~seconds) to drift (~years, plausibly) is in the millions, and there doesn't seem to be a reason it couldn't be much higher.

Suppose it becomes possible to extend human lifetime to arbitrary lengths[56], perhaps at great expense.[57] It seems reasonable to expect that the values of such a person might change no faster than a normal middle-age human, and that his or her values would be much more stable than a succession of humans with normal life spans (even genetic clones).

---

[55] This is certainly *not* obvious from basic knowledge of physics and computer science, if we could imagine knowing those things but not knowing/believing that our brains were physical.

[56] Note that this does not necessarily require biological decay/disease/accident to cease, only that these can be repaired with sufficiently high fidelity to the original. A classic car kept running with recently fabricated parts and historical fuel-mixture ratios would maintain very similar driving performance as when the car originally appeared.

[57] This needn't be economically feasible for the entire human population, e.g., if it were used only by a totalitarian dictator; see below.

*Nearly digital values

Summary: Unlike individual pieces of physical hardware that may decay, digital information can be made essentially eternal through the use of redundancy and – more generally and with greater efficiency – fault tolerance. Any method that can anchor an AGI system's values to digital information is automatically a method for suppressing value drift. Although it is conceivable that there exist surprising facts about general intelligence that prevent it from anchoring to digital information to arbitrarily high accuracy, none are currently known and circumstantial evidence suggests they do not exist.

<Incomplete>

Much speculation about AGI values assumes, explicitly or implicitly, that AGI will have digitally specified values. Recall having values digitally specified means that the part of an AGI system that is responsible for determining its values exists as discrete information – ultimately, a bit string – possibly including both software and data. This assumption of digital values must be made any time copying of an AGI is asserted without discussion of the quality of the copy, i.e.,

As we've seen, the most basic form of fault tolerance takes the form of redundancy, i.e., *back-up* copies. Likewise, the minimal assumption that is actually needed is that the value-storing core of an AGI can be copied. This is a key conjectured property of whole-brain emulations (ems) on which much discussion has been built in spite of the necessarily few concrete things we know now about ems would be achieved technically.[58] Note that if, contrary to our expectations, human brains make crucial use of hyper computations or other tricks that violate the Church-Turing thesis, then em are no longer obviously possible. In particular, if the human brain somehow maintains long-lived quantum coherence that is crucial to its continued operation (in spite of much evidence to the contrary), then copying humans into ems is impossible.

Assuming AGI agents can be digitally copied, the risk of physical destruction of the agent is vastly reduced since it can always be restored from off-site back-ups.[59] In contrast, humans are vulnerable to all sorts of catastrophic risks like illness, accident, and attack.

<Need to argue that even the em scenario is all about error correction><The ability to make copies the heart of error correction, and copying appears as a primary ability in the ultra-stable em scenario. Other scenarios, like immortal biological humans, do not use copying but seem

[58] Robin Hanson, *Age of Em* [URL], Oxford University Press (2016). Carl Shulman, "Whole brain emulation and the evolution of superorganisms" [PDF], Machine Intelligence Research Institute working paper (2010).

[59] Although terrestrial backups cannot avoid the risk of disasters that affect the entire planet, space colonization could avoid even these.

much less likely to never age. The ways in which we imagine they might keep their values forever in the face of random brain fluctuations looks something like error correction, e.g., scanning their brain and fixing chance faults that develop.>

The most obvious way this could be wrong is if AGI fundamentally requires analog hardware or quantum hardware that resists copying, but this seems very unlikely, as it breaks the Church-Turing thesis and related ideas at the foundations of computation. (This hardware must not just be used as a *tool* of the AGI, but must in fact be the source of AGI values in order for this possibility to interfere with our argument.)

Of course, since our main claim is conditioned on the existence of any AGI, that claim would also fail if such hardware were necessary for the operation of that particular AGI even if, in principle, digital AGI was possible.

<Hard coded unnecessary. The goals might be high-level abstract so long as they are generated by a digital source (“seed”) that is preserved.>

Digital information and the software-hardware separation

<Incomplete>

There is a close connection betweeb the software-hardware distinction and meaningful digital information. <Tentative:> Technical impossibility of AGI value stability would strongly imply that every possible AGIs would be intrinsically tied to a specific physical object.

<Reliable porting of static software from one piece of hardware to the next. Software can be essentially immortal.>

<If AGI is a finitely specifiable algorithm that needs to be run on some minimal level of hardware to be useable, but likely would want to be upgraded to more advanced hardware as it accumulates resources, error correction allows perfect transmission from one hardware instantiation to the next.>

Weak independence of intelligence and values

<Incomplete>

There is a connection between value stability and a weak form of the orthogonality thesis: Value lock-in, like orthogonality, is only meaningful insofar as goals could have been different, and many arguments for orthogonality involve an “insertion point” for arbitrary goals that could be digitized and hence error-corrected.

<Lots more needs to be said about relation to orthogonality thesis>

We basically need a highly *weakened* version of the orthogonality thesis. For lock-in to be impossible, it requires something like

- High intelligence implies convergence toward some intrinsic goals (not just convergence toward instrumental goals valuable for most intrinsic goals). But this seems unlikely since the very most intelligent humans, which have vastly more in common with each other than any artificial system need to have, still clearly have different goals, and there's not much reason to think that human are just below the convergence threshold. (And there's no reason that AGI needs to keep progressing; if it's built by someone who hard-codes explicit goals, that might keep it below some level of intelligence, but that's not a big deal if it's a singleton or in a multipolar world of other such AGI who are similarly hobbled.)
- High intelligence implies inherent instability of goals. (E.g., sufficient sophisticated goals must be abstract high-level things that emerge from low-level things, and that emergence frustrates analysis/stability/transparency.)

<Incorporate discussion from Paul's links on "long-term" (reflective/self-improving fixed point?) value stability: a,b>

<Incorporate my statement from Wei Dai interview>

- It seems to me that the key feature of artificial systems over biological ones that enables value lock-in is error correction. It's obvious that any given bit string can be preserved into the indefinite future. What's less obvious is that an AGI can be built whose values are encoded in a bit string, i.e., whose values are "hard-coded".
- The alternative would be an AGI whose values instead only emerge abstractly from its behavior (as revealed preferences), with its behavior being driven both by the original code and by new observations/data/hardware. In that case, a different argument needs to be made for the feasibility of value stability, and it's not clear it will be driven by error correction.
- The connection to the orthogonality thesis is just that hard-coded values clearly imply (a weak version of) orthogonality. That a bit string could be different, leading to different values, is exactly what we mean for the values to be encoded *in* the bit string. "Value lock-in" is only meaningful insofar as things could be different.

<Incorporate notes on Armstrong's defense of orthogonality thesis>

*AGI effective immortality

Summary: Digital AGI is free from the limitations of hardware decay and can be copied with perfect fidelity indefinitely. Even for hypothetical AGI like whole-brain emulations which may unavoidably exhibit some value drift during operation, restarting from back-up copies plausibly allows them to pursue goals over extremely long periods of time.

<incomplete>

It's digital, duh.

Hypothetical development

<incomplete>

Here I describe two (necessarily highly speculative) scenarios for how AGI might arise. I emphasize that these are merely examples and not intended to cover a large fraction of the space of possibilities.

Consider:

- Em AGI: Can always reset them. And terrifying perfect monitoring of thoughts is possible. Not that this will happen, but it is notable that it could even in a future scenario that contains the most human-like AGI, you still can get stability. <Flesh out and reference Age of Em.>
- De novo AGI: Values/goals are a very useful concept for understanding intelligent systems. Along with hard physical constraints, they are one of the few tools for predicting the behavior of hypothetical systems that are more intelligent than humans. In a system that is understood from the group up, a lot of folks find it likely that simple explicit values could be loaded.
- Coherent extrapolated volition (CEV): <Discuss CEV as something like locking in values, but at a higher meta level? Specification of humans constitutes seed of values?>

Global value lock-in

In the previous subsection I argued that, conditional on the technological ability to create AGI systems, such agents could be created with values that are stable over arbitrary long timespans. Here I argue that, assuming such stable AGI agents can be created, then in principle they could tightly constrain or fix the overall future of earth-originating intelligent life.

My argument proceeds along these main points:

1. Irreversible outcomes: Global value lock-in may initially appear unprecedented and unbelievable, and worrying about it might seem alarmist or premature. However, there are several examples of irreversible and globally important outcomes of the physical world, increasing the initial plausibility of value lock-in.
2. Long-lived human influence: There are example of very long-lived human institutions that managed to transmit goals over a millennium, of individual humans (dictators) who managed to control a large fraction of the economic output of very large nations, and of very long-lived human conventions and ideologies that were historically contingent. These examples are ultimately bounded, but the time and impact scales to which they are limited can be traced specific aspects of human beings rather than anything intrinsic to intelligent agents.
3. Power projection and coordination: Due merely to a small number of advantages AGI systems are highly likely to enjoy over humans, such as copying, AGI systems will be largely free of the constraints that ultimately limited the previous examples of long-lived human influence.
4. Concrete scenarios: There exist scenarios by which global lock-in could be achieved that are very likely technically allowed. These are not necessarily likely to play out in the future, but their in-principle feasibility establishes our main claim. Carl Shulman's ultrastable em superorganism is the most concrete such scenario.
5. Stability over astronomical distances: It appears technologically feasible for humanity and its descendants to colonize other stars. Introducing astronomical distances between agents and the corresponding speed-of-life limitations for coordination does not strongly affect my argument.

I now argue for each of the above points in turn. (Readers who accept any given point can skip the corresponding subsection.)

Irreversibile outcomes

Summary: There are several examples of irreversible and globally important outcomes of the physical world. This should increase the initial plausibility of other irreversible and globally important outcomes, like value lock-in.

Values are abstract and potentially fuzzy. Here I discuss some irreversible *physical outcomes* that pose real, albeit possibly very small, risks to humanity.

- Extinction: In 1600, it was impossible for any group of human to destroy the world. Humanity was just too widely dispersed, and the methods of influencing the world were too crude and underpowered. But in 1985, near the peak of nuclear arms build up, an accidental nuclear war could have been catastrophic (though unlikely to end the species). A concerted effort by a crazy dictator in control of the US and Russian arsenals might have been able to cause extinction. Human extinction would be irreversible.
- Earth climate catastrophe: Some climate scientists have argued for the risk of surprisingly strong negative feedback effects sparked by carbon emissions, leading to relatively fast and dramatic climate changes. If prolonged, this could lead to the extinction of the species, or it could prevent recovery for long enough that the remaining human populations evolve to favor significantly different values. <cite>
- Earth resource exhaustion: Humans were present on the planets for a million years, and farming for many millennia, before industrializing. Historically, industrialization was heavily dependent on bountiful and easily accessible sources of energy, especially coal and oil. Essentially all of the easily accessible coal and oil has been extracted, such that the remaining (much larger) natural stocks require highly specialized equipment to recover which was not available prior to industrialization. It is possible that, were a shock to reduce humanity back to a strictly agrarian subsistence, industry might never be recovered due to the absence of accessible energy. <cite>
- Cosmic resource exhaustion: If space colonization is possible, and if an initial wave of colonization were done by agents not aligned with our values (e.g., alien life, or a future human dictator), the natural resources of the cosmos might be partially or wholly consumed in the pursuit of goals we find neutral or negative.

Long-lived human influence

Summary: There are examples of very long-lived human institutions that managed to transmit goals over a millennium, of individual humans who managed to control a large fraction of the economic output of very large nations, and of very long-lived human conventions and ideologies that were historically contingent.

Here are a few institutions that have maintained some aspects of their original mission for more than a millennium:

- University of Al Quaraouiyine (859 AD; 1,159 years old).[60]
- Imperial Monarchy of Japan (539 AD; ~1,500 years old).[61]
- Catholic Church (~350 AD, ~1700 years old).[62]

This was achieved despite constant churn among the humans composing the institutions and a changing economic and political environment.

Even given the significant constraints humans face in projecting power (discussed below), there have been a few cases in history where a dictator took control a significant fraction of total world economic output, or realistically might have done so conditional only a few chance historical events going differently:[63]

- Hitler, especially if Germany wins World War II.[64]
- Stalin, especially if he lived longer and the USSR and Warsaw Pact expands enough to outmatch NATO military.

The reign of dictators seems to have been ended by external forces (military defeat and illness) rather than any countering force that reliably exists in human organizations.[65] More generally, we claim that, though the above examples are ultimately bounded, the time and impact scales to which they are limited can be traced specific aspects of human beings rather than anything intrinsic to intelligent agents.[66]

Although human values are fuzzy, and although these values are only transmitted from one generation to the next with significant noise, there are nevertheless cases where contingent

---

[60] See Wikipedia: List of oldest universities in continuous operation. The longest operated university proper is the University of Bologna (1088 AD; 930 years old). The University of Al Quaraouiyine probably did not qualify as a "university", in terms of the type of degrees awarded, until relatively recently.
[61] Traditionally dated back to 660 BC, but robust records go back only to 539 AD. Arguably ended at conclusion of World War II. See Wikipedia: Imperial House of Japan.
[62] Date depends on what you consider to be earliest instutitions that has sufficient continuity of values with the modern Catholic Church. See Wikipedia: History of the papacy.
[63] All the realistic examples are going to be in the past century or two. Earlier in history, it wasn't feasible for technological and economic reasons for centralized economic control even in cases where there was nominal political control (e.g., Genghis Khan). This is clear from the much smaller fraction of GDP controlled by governments prior to the 19th century. <I don't have a good cite for this.> (E.g., US government spending was ~7% of GDP in 1902.) Mao did not qualify because China only had a few percent of gross world product at the height of his powers.
[64] Germany had the second highest GDP in the world during much of World War II, and if occupied territory is included, it reached within ~20% of the leader (United States). See Wikipedia: Military production during World War II # GDP.
[65] Bryan Caplan, "The Totalitarian Threat" [DOC], p. 498 in *Global Catastrophic Risks*, ed. Nick Bostrom, Milan M. Ćirković, Oxford University Press (2008).
[66] Spelling this out more convincingly is left a future version of this document.

historical details influenced values and conventions across many (e.g., a dozen) generations. Some examples:[67]

- Tonic water[68]
- The US Constitution
- Driving on the right vs. left
- Marxism
- QWERTY keyboards
- Christianity

We would not naturally expect these to be irreversible, in the sense that values in the distant future (e.g., thousands of generations) would probably be uncorrelated with those historical details. Nonetheless, this sort of path dependence over centuries is an imperfect version of the value lock-in considered in this document.

*Power projection and coordination

> *Summary: Under very mild assumptions, AGI systems will not be bound by constraints that currently limit humans from coordinating to achieve global value lock-in.*

<Incomplete>

All humans share many limitations that tend to suppress the domination of any single human, regardless of value stability:

- Limited input/output bandwidth: reading speed, eye resolution, typing speed, speaking speed. (Might be augmented in future with brain-computer interfaces, but seems difficult to vastly improve without major changes to brain structure.)
- Physical fragility (e.g., illness, assassination).
- Lifespan cap.

---

[67] Several of these are examples were taken from pages 7-8 of Nicholas Beckstead, "On the Overwhelming Importance of Shaping the Far Future" [PDF] dissertation (2013): "A classic example of path dependence is our use of QWERTY keyboards…Some political scientists have argued that path dependence is very common in politics…There have been other events that were historically contingent, and changed the course of history significantly. Potential examples include: the rise of Christianity, the creation of the US Constitution, and the influence of Marxism. Various aspects of Christian morality influence the world today in significant ways, but the fact that those aspects of morality, in exactly those ways, were part of a dominant world religion was historically contingent. And therefore events like Jesus's death and Paul writing his epistles are examples of trajectory changes. Likewise, the US Constitution was the product of deliberation among a specific set of men, the document affects government policy today and will affect it for the foreseeable future, but it could easily have been a different document."

[68] Tonic water was originally developed as a method for making more palatable the consumption of quinine, a prophylactic against malaria. It dates back at least to the 19th century, being associated especially with British colonization of India. The widespread consumption of tonic water for the past 100 years, especially in the gin & tonic cocktail, plausibly does not occur in a counterfactual world without British involvement in India.

- Epistemic and cognitive limitations. For one human to accomplish much, they generally must direct large organizations of other humans. In particular:
    - Most decision making of human organizations must be delegated with imperfect oversight even when a single person nominally has absolute authority.
    - Potential rebels can remain discrete and coordinate.

The distinction between one actor and multiple actors may even be an artifact of the discrete nature of human actors. It's conceivable that the intelligent actors in the world long after AGI appears exhibit aspects of both competition/conflict and uniformity/coordination.

Likewise, even when groups of humans coordinate to pursue a shared goal, there are limitations on the effectiveness of these groups:

- Value drift of the organization as new humans are recruited to replace existing ones, or to grow the organization
- <more>

AGI will not be limited in the above ways:

- <list>

<Ways in which a single AGI will be better at controlling more of the world than a single human, and better at coordinating with other AGIs than humans do with humans>
<Singleton discussion "Superintelligence" by Bostrom>

<Surveillance>

<Multiple distributed, perfectly aligned copies>

- One might conjecture that there are surprising constraints on the amount of general intelligence that can be "well coordinated", so that an AGIs can't get much smarter than humans before it is better regarded as multiple AGIs that need to imperfectly coordinate with each other (just like any task with cognitive demandes exceeding a human's necessarily today requires multiple humans that imperfectly coordinate with each other).
    - But identical copy AGIs could coordinate super well!

Of course, we concentrate mainly on whether this totalitarian em scenario is reasonably possible, not whether it's likely. Hanson has discussed how em democracy might look, and changes it might make to be successful.[69]

[69] See the subsection "Governance" in Chapter 16: Conflict in *Age of Em*. <Should insert page number from published version of book.>

<This might occur because global control was acquired by a single unified agent or because multiple such agents coordinate with stability and precision; I argue that both of these possibilities are vastly more feasible for AGI agents than for humans on very long timescales.>

*Concrete scenarios

> *Summary: I sketch a scenario by which global lock-in could be achieved if previously mentioned technical assumptions hold. For the most concrete details, I rely on Carl Shulman's ultrastable em superorganism concept.*

<Incomplete>

...

Here I sketch some hypothetical scenarios with global value lock-in that could plausibly exist.

- Totalitarian super-surveillance police state controlled by effectively immortal bad person. Only tech needed is one immoral immortal dude (to prevent value drift) and then some mundane surveillance stuff. Effective immortality (see earlier) might come from <insert>. See Caplan.[70]
    - Biological life extension
    - Lots of em copies that periodically get reset.
    - Notes follow
    -
- <Decisive strategic advantage><Read DSA discussion "Superintelligence" by Bostrom>
- <Think of more minimal examples>

Insofar as humans have well-defined values, the timescale on which those values persist seems to be contingent on the details of human evolution and in particular *not* on anything fundamental to intelligence.[71] This perspective suggests that AGI values could easily be stable on much longer timescales.

---

[70] Bryan Caplan, "The Totalitarian Threat" [DOC], p. 498 in *Global Catastrophic Risks*, edited by N. Bostrom and M.M. Ćirković. Oxford University Press, 2008, 554 pages; ISBN 978-0198570509.

[71] One can imagine counter-hypotheses but they aren't very compelling, e.g., any intelligent system that wants to learn and react on some characteristic timescale might, for as-yet unknown technical reasons, necessarily experience value-drift on (some fixed multiple of) that timescale, so that anything that wants to be competitive in our environment must drift roughly as fast as humans.

Stability over astronomical distances

Summary: Astronomical distances and the resulting speed-of-light constraints do not pose strong unique problems for maintaining lock-in compared to an Earth-bound scenario. Interstellar colonization probably slightly increases the likelihood of lock-in by increasing robustness to planet-wide risks.

It's likely that humanity or its descendants will be able to colonize other star systems.[72] Here I review how it interacts with the potential for value lock-in, such as affecting the probability that lock-in is thwarted by certain natural risks. This probably does not appreciably change the basic arguments regarding the technical feasibility of lock-in

Some notable features introduced by astronomical distances[73] are

- Speed-of-light constraints: An AGI system cannot react instantly on one planet in response to events on another. Note, though, that speed-of-light constraints will already be noticeable for a rapidly evolving post-AGI civilization on a single planet. It takes 0.04 seconds for light to cross the diameter of the Earth, and many interesting things may happen during that time on an em Earth.[74]
- New robustness to risks from correlated threats occurring on scales as large as a planet but smaller than the whole civilization. Examples:
    - Nearby supernovae[75]
    - Environmental catastrophe
    - Aliens. For instance, if the Aliens are at intermediate scale (more powerful than a planet but less powerful than the Earth-originating AGI) or the Earth-originating AGI would need to possess a strong defensive advantage on account of its civilization size (e.g., it's expanding at near light speed, or is passed the cosmic causal horizon, and can't be caught)

---

[72] Nicholas Beckstead, "Will we eventually be able to colonize other stars? Notes from a preliminary review." [URL], webpage (2014). Stuart Armstrong and Anders Sandberg, "Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox" [PDF], *Acta Astronautica* 89, 1-13 (2013).

[73] Here is a brief summary of the basic distance scales involved in colonization (adapted from earlier assignment): The distance from the Earth to the Sun defines 1 astronomical unit, about 150 million kilometers or 8 light-minutes. The farthest planet in the Solar system is Neptune, which is always about 30 astronomical units away. The distance to the nearest star is 4.3 light-years or 270,000 astronomical units. The typical spacing between stars is ~3 light-years at our distance from the center of the Milky Way. The Milky Way is roughly disc-shaped, with a thickness of about 1k light years and a diameter of 100k light-years. It contains about 500 billion stars. The nearest comparably sized galaxies are about 3 million light-years away, although there are other (smaller) dwarf galaxies that are closer.

[74] Robin Hanson, *Age of Em* [URL], Oxford University Press (2016).

[75] Incidentally, supernovae are the most serious existential risk (that I can think of) against which 100-lightyear-scale colonization provides significant protection. However, supernovae close enough to endanger Earth occur every few hundred million years at most, so are negligible as an extinction risk.

- Vast volumes and resources supporting a huge number of events, making extremely improbable events less so. This is important when there is some lock-in failure mechanism that can spread. The chance of such an event on any given planet might be so small as to be very likely during the entire lifetime of the universe, while the chance that it occurs on any of a large number of planets might be large. Examples:
    - Optimization daemons. For instance, disconnected evolutionary dynamics of non-intelligent life on separate planets would be an example; if intelligent life evolves on one planet, it might then spread to the others.
    - Physics disaster. For instance, maybe the creation of very unusual physical condition (e.g., extremely high-energy particle collisions) sparks vacuum decay[76], overturning all the known effective physical laws that supports life as we know it.

Note that an AGI overseeing an interstellar civilization has the option to administer each planet (or each star systems) independently with little or no communication. If lock-in can be achieved on a single planet, it can likely be achieved independently on *N* planets, and usually this would qualify as global lock-in across all planets. The primary ways I can imagine this failing are

- The previously mentioned issue with unlikely but contagious lock-in failure.
- The choice to maintain no communication between planets is incompatible with the values values of the AGI.

Otherwise, the additional robustness provided by being spread over multiple star systems appears to only increase the case of lock-in.

[76] See Wikipedia: False_vacuum # Vacuum_decay.

References

<Purposefully in 10-point font to simplify copy-pasting into footnotes.>

Stephen Omohundro, “The Nature of Self-Improving Artificial Intelligence” [PDF], manuscript (2007).

Stephen Omohundro, “The Basic AI Drives” [PDF] p. 483 in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference,* edited by P. Wang, B. Goertzel, S. Franklin (2008).

Robin Hanson, “Chapter 8 : The Rapacious Hardscrapple Frontier”, [PDF] in *Year Million: Science at the Far Edge of Knowledge*, pp. 168-192, ed. Damien Broderick, Atlas Books, (2008).

Robin Hanson, *Age of Em* [URL], Oxford University Press (2016).

Edwin Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press (2003).

Nick Bostrom, “Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards” [PDF], *Journal of Evolution and Technology*, Vol. 9, (2002).

Nick Bostrom, “Existential Risk Prevention as Global Priority” [PDF], *Global Policy*, Vol 4, Issue 1, p. 15-31 (2013).

Nick Bostrom, “What is a Singleton?” [URL], *Linguistic and Philosophical Investigations*, Vol. 5, No. 2 (2006): pp. 48-54.

Nick Bostrom, “The Superintelligent Will: Motivation and Instrumental Rationality in Advance Artificial Agents” [PDF], Minds and Machines, 22(2), 71-85 (2012).

Nick Bostrom, “Existential Risk Prevention as Global Priority” [PDF], *Global Policy,* Volume 4, Issue 1. (2013).

Nick Bostrom, *Superintelligence* [URL], (2014).

Nick Bostrom, “The Vulnerable World Hypothesis” [PDF], working paper (2018).

Bryan Caplan, “The Totalitarian Threat” [DOC], p. 498 in *Global Catastrophic Risks*, ed. Nick Bostrom, Milan M. Ćirković, Oxford University Press (2008).

Eliezer Yudkowsky, “Artificial Intelligence as a Positive and Negative Factor in Global Risk” [PDF], p. 308 in *Global Catastrophic Risks*, ed. Nick Bostrom, Milan M. Ćirković, Oxford University Press (2008).

Eliezer Yudkowsky, “Intelligence Explosion Microeconomics” [PDF], manuscript (2013).

Nate Soares, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong, “Corrigibility” [PDF], In AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, January 25–26, 2015.

Stuart Armstrong, "General Purpose Intelligence: Arguing the Orthogonality Thesis" [PDF], *Analysis & Metaphysics*, 12 (2013).

Stuart Armstrong, “Satisficers want to become maximisers” [PDF], blog post (2011).

Stuart Armstrong and Anders Sandberg, “Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox” [PDF], *Acta Astronautica* 89, 1-13 (2013).

Nicholas Beckstead, “On the Overwhelming Importance of Shaping the Far Future” [PDF], dissertation (2013).

Nicholas Beckstead, “Will we eventually be able to colonize other stars? Notes from a preliminary review.” [URL], webpage (2014).

Augustine Kong et al., “Rate of de novo mutations and the importance of father’s age to disease risk”, [URL] *Nature* 488, 471–475 (2012).

Dónall A. Mac Dónaill, “A parity code interpretation of nucleotide alphabet composition” [URL], *Chemical Communications* 18, 2062-2063 (2002).

L.S. Liebovitch, Y. Tao, A. T. Todorov, and L. Levine, “Is there an error correcting code in the base sequence in DNA?” [URL], Biophysics Journal, 1996 Sep; 71(3): 1539–1544.

John W. Drake et al., “Rates of spontaneous mutation” [PDF], *Genetics* 48, 1667–1686 (1998).

M. Prezioso et al., “Training and operation of an integrated neuromorphic network based on metal-oxide memristors” [PDF], *Nature* 521, 61–64 (2015).

Paul Christiano, “We can probably influence the far future” [URL], blog post (2014).

Paul Christiano, “Machine intelligence and capital accumulation” [URL], blog post (2014).

Paul Christiano, “Approval Directed Agents” (presumed title) [URL], blog post (2014).

Paul Christiano, “ALBA: An explicit proposal for aligned AI” (presumed title) [URL], blog post (2016).

Paul Christiano, “Corrigibility”, [URL], blog post (2017).

Carl Shulman, "Whole brain emulation and the evolution of superorganisms" [PDF], Machine Intelligence Research Institute working paper (2010).

Appendix

Process

My academic training is in quantum information, which has given me only brief exposure to the basics of digital error correction. I have no expertise on artificial intelligence and only a light familiarity with the discussion of AGI within the EA and rationality communities. My process for writing this document was as follows:

- Claim chosen because if its key role in long-termist planning, the relative vagueness and shallowness of existing discussion, and my personal interest.
- Before reviewing existing discussion on this topic, I sketched an argument from my personal understanding and past reading.
- I asked Carl Shulman, Paul Christiano, Dan Selsam, and Nick Beckstead for suggested references. I re-read relevant sections of Bostrom's *Superintelligence* and Hanson's *Age of Em*. I searched Less Wrong, the SL4 archives, the EA forum, the Intelligent Agent Foundations Forum, and Arbital for the terms "error correction", "irreversibility", and "path dependence" (and minor modifications), especially for discussion by Eliezer Yudkowsky, Nick, Paul, and Wei Dai.
- This led to less than I expected. Most discussion was informal: on blogs more than books, and in comments rather than main posts. In particular, though it is generally acknowledged that values might not be stable under recursive self-improvement of AGI, the value stability of non-improving (stagnant) AGI was generally assumed to be possible with the help of error correction, which I think may be non-obvious to many outside the EA and rationality communities.
- The most important thing that came from the initial reading was
    - The close connection between value stability and the orthogonality thesis.
    - The lack of discussion of how value stability, insofar as it is supposed to be enabled by error correction of computer memory, interacts with the extent to which values are hard-coded in that memory (vs. very abstract and emergent).
- Discussion with Luke helped clarify that my interest was mostly in value stability of individual AGI systems (both whether this was a coherent topic and whether it was likely to be achieved) and less in the possibility of global coordination/lock-in given those systems (which I take to be more intuitive for outsiders). I broke my argument up into those two corresponding pieces.
- I then bracketed the questions of whether it was *likely* that global lock-in would occur, as described in the section on Motivation.
- Interviewed Carl Shulman, Paul Christiano, and Wei Dai.
- Due to time limitations and the evaluation goals of this assignment, I stopped digging in to the speculative research on AI, which is a bit unbounded, and concentrated on

bringing some concrete technical ideas (error correction, astronomical distances) to bear.

Next steps

If I were to improve this document in the future, I would follow these next steps:

- Interview more experts, especially skeptical ones outside the EA and rationality communities. Rodney Brooks might be good, especially with regard to whether he thinks goals or expected-utility maximization are ultimately good bets even though he thinks AGI is quite distant.
- Pull out more detail from Carl's superorganism paper and draw out the connection to copying.
- Consider introducing a dedicated section on the most likely ways the argument could fail and the main claim could be false.
- Fill in many missing cites, attach probabilities to more things, etc.

Error correction efficiency example

In our introduction to fault tolerance, we demonstrated only the simplest possible sort of fault tolerance: redundancy. The cost of redundancy schemes is to increase computational resources proportionally to the redundancy *R*, but more sophisticated schemes exist with a much smaller overhead that nevertheless assure that the chance of an uncorrected fault is negligible.

To get a flavor for how this works, consider the special case of error correction in memory: we have *B* bits in memory, with each bit of course being either a 0 or 1, and during a certain period of time each bit has a chance of being flipped by a passing cosmic ray. A simple redundancy code would be to record each bit *R* times when the memory is originally written. Then, when the memory is accessed, the system reports each bit as the majority vote of its *R* copies. For *R*=3 odd, such a system can tolerate a fault at any single memory location without performing incorrectly; however, as noted, this triples the memory hardware requirements.

Instead, an error-correcting code (ECC) can be used that can also tolerate a fault at any single memory location but without the same inefficiencies as using plain redundancy. Suppose[77] we have 4 bits of data, $d_1d_2d_3d_4$, with $16=2^4$ possible configuration (0000, 0001, 0010, …). Now, our trick is write those 4 data bits to memory and along with 3 "parity bits" $p_1p_2p_3$ defined as follows:

[77] Here I am describing the code known as Hamming(7,4). Many other codes exist depending on the desired properties. The most famous class of codes may be the Reed-Solomon codes, which enable optical discs like DVDs to be read perfectly in spite of scratches.

- p_1=0 if d_2+d_3+d_4 is an even number. Otherwise, p_1=1.
- p_2=0 if d_1+d_3+d_4 is an even number. Otherwise, p_2=1.
- p_3=0 if d_1+d_2+d_4 is an even number. Otherwise, p_3=1.

This takes a total of 7 bits of memory. Although there are 128=2^7 possible configurations of that memory, we are only using 16 of them since the parity bits are completely determined by the data. Here are what all 16 possibilities look like:

Data	d_1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
	d_2	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
	d_3	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
	d_4	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
Parities	p_1	0	1	1	0	1	0	0	1	0	1	1	0	1	0	0	1
	p_3	0	1	1	0	0	1	1	0	1	0	0	1	1	0	0	1
	p_3	0	1	0	1	1	0	1	0	1	0	1	0	0	1	0	1

The parity bits were chosen to have this property (although we won't prove it rigorously): given any two possible configurations of the data that differ by at least a single bit (e.g., "1**0**10" and "1**1**01"), the joint configuration of data and parities differs by at least *three* bits (e.g., "1**0**10-**10**1" and "1**0**11-**00**1"). This means that, if a single bit is flipped by a fault in the memory, it's always possible to distinguish what the data configuration originally was. (That is, the memory could not end up in a configuration that is consistent with either original data being *X* and the *n*th bit flipping *or* the original data being *Y* and the *m*th bit flipping.) Thus, the error can be identified and corrected.

Although there is some neat math that can be used to show this very concisely, it's probably easiest for the reader to convince themselves by randomly selecting a few pairs of columns and checking that no pair is related by two bit flips or less. (A complete demonstration can be found at the Wikipedia page on the Hamming(7,4) code.)

This example of encoding 4 bits of data in 7 memory bits has a fractional overhead cost of (7-4)/4 = 75%. This is the best scheme possible for 4 bits, but for large amounts of data more sophisticated error correction schemes exist that drive this fractional overhead cost to arbitrarily low levels. Furthermore, this can be done while remaining tolerant to multiple faults so that, assuming fault independence, the probability that the memory experiences any *uncorrected* faults is exponentially small.

More generally, there exist methods for correcting fault not just in memory storage but also during logical manipulations (including the error-correcting process itself) such that it's possible

to perform arbitrarily large and lengthy computations in a fully fault tolerant way, with exponentially small chances of failure at modest overhead cost.

Notes

[This section, and all following sections, contain working notes and are not useful to read.]

Error correction and value stability

LessWrong links

Nick Beckstead, "A Proposed Adjustment to the Astronomical Waste Argument".

- "In this post, I argue that:
    - Though Bostrom's argument supports the conclusion that maximizing humanity's long term potential is extremely important, it does not provide strong evidence that reducing existential risk is the best way of maximizing humanity's future potential…
    - A version of Bostrom's argument better supports a more general view: what matters most is that we make path-dependent aspects of the far future go as well as possible…
    - The above points favor very broad, general, and indirect approaches to shaping the far future for the better…"

SL4 links

(No relevant ones found yet)

EA Forum links

(No relevant ones found yet)

MIRI Agent Foundations Forum / Arbital Links

Just the Paul thing

- https://arbital.com/p/orthogonality/
- https://agentfoundations.org/item?id=1290

Other

<It's obvious we can keep shakespeare around by making copies. This is redundancy code>
<Mention that error correction (w/ Hamming codes) is just part of general "fault tolerance" in the special case where bits flip>

- <Maybe talk in “Explanatory and predictive power of goals” about the issues with defining goals in humans, using examples from subsection “Reduced value drift as humans age”>

<DNA repair><Reddit comments 1, 2.><Wiki><Kahn Academy><Scitable>

More path-dependent examples. From Jacy Reese?

Goals

- <When goals are abstract and arise from software see, we will [or won’t?] fail to distinguish between the digital source and the generated values except where necessary.>
- EUMs with explicit goals look dangerous and/or fragile. <Cite/explain>
    a. Wireheading? (But humans do it too?...sort of)
    b. Ontological crisis

Counterarguments:

- Counterargument: Regardless of vNM theorem and other arguments for utility functions, there’s no reason that AGI will operate *internally* according to a utility function (even if it generally must act externally in this way to avoid getting dutch-booked). Therefore, there doesn’t have to be anything to lock in. AGI could be messy like people
- Counterargument: Real humans don’t seem to worry too much about value drift. Why should we expect real humans to design machines that do?
    - Well, humans *do* care about the value drift in their subordinates.

  It may be true that lock-in is bad/inefficient/uncompetitive and that human won’t choose to build value-locked AI. Even then, however, lock-in is an attractor state for AGI (but not biological AI), in the same way that extinction is an attractor state. What reason do we have to think that this is the most important century, since there will just be lots of chances for lock-in in the future? Basically, you have to think that the chance of lock-in is super low *and* not decreasing for several future centuries in order for this to not be the most likely century for values to get locked in.

Following Carl-Nick and Armstrong, do a “mirror argument” for hypothetical constraints that might arise on goals: We can hypothesize that the goals of all intelligent agents are pushed toward in some direction (objective morality, or whatever), but why not the opposite direction?

- Also, add that constraints can appear from weird unexpected places. The constraints on computation from General Relativity would have been extremely difficult to foresee.

<Superintelligence, Nick Bostrom defines it this way: “Intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal.”>

<The autogenerated Table of Contents is pretty long and ungainly. Probably should use only if we collapse it to multiple columns>

<Possible cites/discussion to incorporate>:

- <Nick on path dependence, both thesis and Less Wrong: Nick Beckstead, “A Proposed Adjustment to the Astronomical Waste Argument”.>
- <Paul Christiano, “Machine intelligence and capital accumulation” [URL]. Cristiano reviews (and an informal/intuitive level) how values imperfectly propagate from one human generation to the next, and how machine intelligence could change this (*assuming* machines can built with fixed values), with large effects on the distributions of resources and the stability of dynasties.>
- <Paul Christiano, “We can probably influence the far future” [URL], blog post (2014). General possibility of influencing the long-term future.>

Other considerations:

- Existential risk from *natural* disasters is unlikely to contribute to making this century more important than previous ones since humanity is more robust now than in the past. (Bad counter argument about increased interconnectedness and hence less resiliency today?) (This is necessarily not in conflict with giving a high expected value to reducing x-risk from natural disasters in light of the fact that such risks should be much lower in the future.)
- Even in multipolar scenarios, and even in unstable ones, it’s possible to get lock-in with goals that are shared by all agents.
- Nick’s angle (inspired by Paul): We cede power to machines who basically obey us, but we lose track of most of what’s going on. Bureaucracy spirals out of control. Even if we do ok, most future value is lost. Distinct from normal human situation; although it’s true our human descendants might not share our values, at least they share our genes. We just don’t have as much in common with machines, and they could pursue really weird stuff by our lights. Might be because AGI doesn’t really keep us informed well (slight misaslignment)?
- <I think this is right, but not important>”Within quantum computing, people generally use the term “error correction” for correcting errors in memory using computations (bit manipulations) that are assumed to be perfect and the term “fault tolerance” for the more general case where errors can occur during the computations but the computation still proceeds reliably. (In the case of quantum computing, these can apparently be shown to be distinct; there exist computation errors that cannot feasibly be fixed/avoided with mere error correction. I think the idea is that even if you do 3 or more identical computations in parallel and then take a majority vote at the end, the majority voting process itself could be so noisy that it’s worse than just doing the computation once and crossing your fingers.)”

Links

- Wei Dai on Paul vs. MIRI (which mentions “error correction”, but only in a metaphorical sense) and is the best thing I could find on the MIRI agent foundations forum, but wasn’t helpful.
- Preserving the utility function doesn’t help if the probabilities can be manipulated: https://arbital.com/p/actual_effectiveness/
- Alexey Turchin advocates AI Nanny with narrow AI to prevent emergence of unfriendly AGI
  https://www.lesswrong.com/posts/7ysKDyQDPK3dDAbkT/narrow-ai-nanny-reaching-strategic-advantage-via-narrow-ai

Summaries and Excerpts

Expected utility maximization

Omohundro’s defense (2008)

Stephen M. Omohundro
“The Nature of Self-Improving Artificial Intelligence”
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.137.1199&rep=rep1&type=pdf

- “The basic theory was created by John von Neumann and Oskar Morgenstern in 1944 [4] for situations with objective uncertainty and was later extended by Savage [5] and Anscombe and Aumann [6] to situations with subjective uncertainty.”
  - [4] J. von Neumann and O. Morgenstern, Theory of Games and Economic Behavior. Princeton University Press, 60th anniversary commemorative edition ed., 2004.
  - [5] L. J. Savage, Foundations of Statistics. Dover Publications, 2nd revised ed., 1954.
  - [6] F. J. Anscombe and R. J. Aumann, “A definition of subjective probability,” Annals of Mathematical Statistics, vol. 34, pp. 199–205, 1963.
- “Why should a self-improving system behave according to this deliberative prescription?...The argument is more compelling if we can identify explicit negative consequences for a system if it fails to follow the axioms. We call potential negative consequences “vulnerabilities”.”
- The vulnerabilities he lists are the usual ones: circular preferences, dutch books
  - But we already know there are ways to avoid getting dutch-booked without a utility function: just notice if you are facing a dutch book and decline to play. Bookies don’t always get to look inside at your circuitry to know the full range of bets you’re willing to take.
  - Etc.

- Ultimately this strikes me as significantly less compelling than Jaynes' argument for how to construct a robot. (And this we know fails if the robot has to worry about having its memory messed with, e.g., sleeping beauty problems).

Bostrom's defense (Superintelligence)

- XXX

Orthogonality thesis

Bostrom's defense (2012)

- Should be the default assumption. Then we rebut contrary intuitions based on errors:
    - Mind space is vast, humans just a tiny part. Even aliens should have sorts human-like (or at least, biological) goals. But AI could want anything
    - Humean defense
    - DIstinguish between mere intelligence and the "normatively thick" version of rationality.

Armstrong's defense (2013)

- Rebutting who?: "A. Kornai, for instance, considers it as a worthwhile starting point for investigations into AI morality (Kornai, 2013). He bases his argument on A. Gewith's approach in his book, Reason and Morality (Gewirth, 1978) (the book's argument can be found in a summarised form in one of E. M. Adams's papers (Adams, 1980)) in which it is argued that all agents must follow a "Principle of Generic Consistency" that causes them to behave in accordance with all other agent's generic rights to freedom and well-being. Others have argued that certain specific moralities are attractors in the space of moral systems, towards which any AI will tend if they start off with certain mild constraints (Waser, 2008)."
- Arguments for orthogonality:
    - Intelligent humans seem to have widely divergent moralities (good/bad/weird)
        - But haven't they been argued to be more "moral" according to common sense morality? (Data and interpretation highly disputed, no doubt.)
    - Humean Defence: Beliefs are motivationally inert (is-ought, etc.)
    - AIXI: If you can label goals, you can just define your superintelligence in an uncomputable way like AIXI to pursue those goals with its infinite computing

power. Thus, there has so be something about what makes AIXI unrealistic[78]: e.g., that it's *incomputable*.
    - Expected utility maximizers -- can't they exist?
    - Humans exist, their goals are pretty diverse, and we expect to be able to create AI that have any goals humans can have. Furthermore, we expect to be able to modify an AI even more than a human; could probably warp their goals pretty good.
    - We can put guns to (some) people's head and use their existing terminal goal (don't die) to get them to pursue almost only any instrumental goal. They don't get dummer. Instead of a gun, it could just be a lot of money.
    - Seems like there could be anti-agents: whatever one agent wants, it would want to do the opposite (perhaps out of vengeance or spite).
    - Oracle AI + a chooser: Seems like you could build an AI that just was able to answer questions about the world really accurately, such what the effects of certain actions would be. Then you could have a chooser (an evil human or whatever) take the action that pursued whatever goal.
    - Or trick them: Humans will follow just about any action for a day if you promise a sufficient reward the next day. If you consider this as an input-output algorithm, you can run it, use the output, and then turn it off.
- Me:
    - It seems like any rebuttal to this is going to be of this form: *real* general intelligence, which human have a bit of, is necessary for doing seriously intelligent things with feasible amounts of computing power in the real world. (Maybe this is something like "having a true theory of mind where you really *get* that other sentient people exist".) This is merely a component of the human brain is, and of course its optimization power can be co-opted with an amoral add-on that just pursues immoral goals. But you'd be constantly "fighting" this with some sort of coercion.
    - Which hypothesis is more plausible for explaining why there's a lot of overlap between the goals of different humans: that they have a shared evolutionary and cultural history (genes, etc.), or *merely* that they are all have a measure of general intelligence.
    -

[78] For criticism due to its dependence on choice of universal Turing machine, see Jan Leike and Marcus Hutter, "Bad Universal Priors and Notions of Optimality" [PDF], *Proceedings of the 28th Conference on Learning Theory* (2015). For discussion of Cartesianism and uncomputability, see Less Wrong here and here and Arbital here and here.

The Totalitarian Threat

The deep question, however, is whether this short duration was inherent or accidental. If the short lifespan of totalitarianism is inherent, it probably does not count as a "global catastrophic risk" at all. On the other hand, if the rapid demise of totalitarianism was a lucky accident, if future totalitarians could learn from history to indefinitely prolong their rule, then totalitarianism is one of the most important global catastrophic risks to stop before it starts.

The main obstacle to answering this question is the small number of observations. Indeed, the collapse of the Soviet bloc was so interconnected that it basically counts as only one data point. However, most of the historical evidence supports the view that totalitarianism could have been much more durable than it was.

This is clearest in the case of Nazi Germany. Only crushing military defeat forced the Nazis from power. Once Hitler became dictator, there was no serious internal opposition to his rule. If he had simply pursued a less aggressive foreign policy, there is every reason to think he would have remained dictator for life. One might argue that grassroots pressure forced Hitler to bite off more than he could militarily chew, but in fact the pressure went the other way. His generals in particular favored a less aggressive posture. (Bullock 1993: 393-4, 568-574, 582)

The history of the Soviet Union and Maoist China confirms this analysis. They were far less expansionist than Nazi Germany, and their most tyrannical leaders – Stalin and Mao - ruled until their deaths. But at the same time, the demise of Stalin and Mao reveals the stumbling block that the Nazis would have eventually faced too: succession. How can a totalitarian regime ensure that each generation of leaders remains stridently totalitarian? Both Stalin and Mao fumbled here, and perhaps Hitler would have done the same.

…

It is tempting for Westerners to argue that the Soviet Union and Maoist China changed course because their systems proved unworkable, but this is fundamentally incorrect. These systems were most stable when their performance was worst. Communist rule was very secure when Stalin and Mao were starving millions to death. Conditions were comparatively good when reforms began. Totalitarianism ended not because totalitarian policies were unaffordable, but because new leaders were unwilling to keep paying the price in lives and wealth.

Probably the most important reason why a change in leaders often led totalitarian regimes to moderate their policies is that they existed side-by-side with non-totalitarian regimes. It was obvious by comparison that people in the non-totalitarian world were richer and happier. Totalitarian regimes limited contact with foreigners, but news of the disparities inevitably leaked in. Even more corrosively, party elites were especially likely to see the outside world first-hand. As a result, officials at the highest levels lost faith in their own system.

This problem could have been largely solved by cutting off contact with the non-totalitarian world, becoming "hermit kingdoms" like North Korea or Albania. But the hermit strategy has a major drawback. Totalitarian regimes have trouble growing and learning as it is; if they cannot borrow ideas from the rest of the world, progress slows to a crawl. But if other societies are growing and learning and yours is not, you will lose the race for political, economic, and military dominance. You may even fall so far behind that foreign nations gain the ability to remove you from power at little risk to themselves.

Thus, a totalitarian regime that tried to preserve itself by turning inwards could probably increase its life expectancy. For a few generations the pool of potential successors would be less corrupted by alien ideas. But in the long-run the non-totalitarian neighbors of a hermit kingdom would overwhelm it.

The totalitarian dilemma, then, is that succession is the key to longevity. But as long as totalitarian states co-exist with non-totalitarian ones, they have to expose potential successors to demoralizing outside influences to avoid falling dangerously behind their rivals.

To understand this dilemma, however, is also to understand its solution: Totalitarianism would be much more stable if there were no non-totalitarian world. The worse-case scenario for human freedom would be a global totalitarian state. Without an outside world for comparison, totalitarian elites would have no direct evidence that any better way of life was on the menu. It would no longer be possible to borrow new ideas from the non-totalitarian world, but it would also no longer be necessary. The global government could economically and scientifically stagnate without falling behind. Indeed, stagnation could easily increase stability. The rule of thumb "Avoid all change" is easier to correctly apply than the rule "Avoid all change that makes the regime less likely to stay in power."

…

Technologically, the great danger is anything that helps solve the problem of succession.

…

Improved surveillance technology like the telescreen would clearly make it easier to root out dissent, but is unlikely to make totalitarianism last longer. Even without telescreens, totalitarian regimes were extremely stable as long as their leaders remained committed totalitarians. Indeed, one of the main lessons of the post-Stalin era was that a nation can be kept in fear by jailing a few thousand dissidents per year.

Better surveillance would do little to expose the real threat to totalitarian regimes: closet skeptics within the party. However, other technological advances might solve this problem. In Orwell's 1984, one of the few scientific questions still being researched is "how to discover, against his

will, what another human being is thinking." (1983: 159) Advances in brain research and related fields have the potential to do just this. Brain scans, for example, might one day be used to screen closet skeptics out of the party. Alternately, the new and improved psychiatric drugs of the future might increase docility without noticeably reducing productivity.

Behavioral genetics could yield similar fruit. Instead of searching for skeptical thoughts, a totalitarian regime might use genetic testing to defend itself. Political orientation is already known to have a significant genetic component. (Pinker 2002: 283-305) A "moderate" totalitarian regime could exclude citizens with a genetic predisposition for critical thinking and individualism from the party. A more ambitious solution – and totalitarian regimes are nothing if not ambitious – would be genetic engineering. The most primitive version would be sterilization and murder of carriers of "anti-party" genes, but you could get the same effect from selective abortion. A technologically advanced totalitarian regime could take over the whole process of reproduction, breeding loyal citizens of the future in test tubes and raising them in state-run "orphanages." This would not have to go on for long before the odds of closet skeptics rising to the top of their system and taking over would be extremely small.

A very different route to totalitarian stability is extending the lifespan of the leader so that the problem of succession rarely if ever comes up. Both Stalin and Mao ruled for decades until their deaths, facing no serious internal threats to their power. If life extension technology had been advanced enough to keep them in peak condition forever, it is reasonable to believe that they would still be in power today.

…

At the same time, it should be acknowledged that some of these technologies might lead totalitarianism to be less violent than it was historically. Suppose psychiatric drugs or genetic engineering created a docile, homogeneous population. Totalitarian ambitions could then be realized without extreme brutality, because people would want to do what their government asked – a possibility explored at length in the dystopian novel Brave New World. (Huxley 1996)

Corrigibility

Wei Dai

“The main motivation here is (as I understand it) that learning corrigibility may be easier and more tolerant of errors than learning values. So for example, whereas an AI that learns slightly wrong values may be motivated to manipulate H into accepting those wrong values or prevent itself from being turned off, intuitively it seems like it would take bigger errors in learning corrigibility for those things to happen. (This may well be a mirage; when people look more deeply into Paul's idea of corrigibility maybe we'll realize that learning it is actually as hard and

error-sensitive as learning values. Sort of like how AI alignment through value learning perhaps didn't seem that hard at first glance.) Again see Paul's posts linked above for his own views on this."
https://www.lesswrong.com/posts/ZyyMPXY27TTxKsR5X/problems-with-amplification-distillation#hLqTDmMrGmfHb4BvJ