Lawrence University
Lux

Lawrence University Honors Projects

5-31-2017

Building Morality: A New Strategy for Creating Human-Like Moral Psychology in Artificial General Intelligence

Christopher Barr
christopher.l.barr@lawrence.edu

Follow this and additional works at: https://lux.lawrence.edu/luhp

 Part of the Philosophy Commons

© Copyright is owned by the author of this document.

Recommended Citation
Barr, Christopher, "Building Morality: A New Strategy for Creating Human-Like Moral Psychology in Artificial General Intelligence" (2017). *Lawrence University Honors Projects*. 112.
https://lux.lawrence.edu/luhp/112

This Honors Project is brought to you for free and open access by Lux. It has been accepted for inclusion in Lawrence University Honors Projects by an authorized administrator of Lux. For more information, please contact colette.brautigam@lawrence.edu.

Building Morality

A New Strategy for Creating Human-Like Moral Psychology in Artificial General Intelligence

Christopher Barr

Lawrence University | christopher.l.barr@lawrence.edu

Honors Thesis

Advisor: Mark Phelan

Date: 1 May 2017

IHRTLUHC

Acknowledgements

There are a great many people I should thank for their assistance and support. Without their help, I would have never completed this paper.

First off, I want to thank my parents, Jeanne and Mark Mosher, for everything they've done to help me along. I would never have gotten to where I am today without them. In the more immediate moment, I'm extremely grateful for the time and effort they put into editing this paper. It would be far worse without their assistance; plus, had they not helped, I would not have had the opportunity to poke fun at my mom for the best typo I've seen this year: "humrans".

My thanks to my philosophy professors, particularly Mark Phelan, John Dreher, and Thomas Ryckman, for formally introducing me to philosophy and showing me its uses. Turns out that it's not so dry and boring after all! In this vein, I'd also like to thank professors Ingrid Albrecht, for her illuminating class on Existentialism, and Chloe Armstrong, for mixing my two favorite topics: science fiction and philosophy.

I'd like to thank Jasper Olsen for their hours of line editing (and, occasionally, bringing me food). I can't imagine the typos that would be here were it not for their help.

Many thanks to Thomas Kilmer for the many fascinating discussions about the nature of morality and the kinds of errors people can make in their judgments.

And, finally, thanks to Jaeden Atwater, Will Criste, Alex Thorp, Alex Lessenger, and Sari Hoffman-Dachelet for their varied contributions.

Table of Contents

Introduction

Intelligence, in a broad sense, can be thought of as the ability of an agent to optimize towards a goal. We see this everywhere, to varying degrees: your dog is intelligent enough to figure out when you generally get home and to listen for your car; you can read this paper, drive your car, talk, and make your own food; Google's search algorithms usually give the most relevant results for a given input. These examples demonstrate the different kinds of intelligence: *general* and *narrow*. A narrow intelligence is good at one thing, but it does that one thing very well. Google's search engine, for instance, can't hold a conversation, but it is certainly quite good at obtaining relevant search results. A general intelligence can at least attempt many different things, and more powerful general intelligences are good at many things. This is the kind of intelligence that humans have.

We may also classify intelligence as being *weak* or *powerful*[1]. This is a statement of efficiency. A weak intelligence, whether narrow or general, requires many resources to compute a certain thing; a comparatively powerful intelligence requires fewer resources to make the same computation. This is the difference, for example, between a two-year-old child and an engineer. They are both general intelligences, but the two-year-old is a weak one and the engineer is powerful. Though the two-year-old can attempt many things, it cannot do them quickly or well, and it takes a long time to comprehend the task at hand. The engineer, on the other hand, can attempt more things (making them a more-general intelligence) and do them better, faster, and understand them more quickly. Though there are many examples of general intelligence, such as dolphins, elephants, and great apes, so far as we know humans are the most powerful general intelligences in existence.

[1] Notably, in discussions of efficiency of artificial intelligence, one must avoid using "strong": in the AI research community, "strong AI" is one of the terms used to describe conscious machines.

It is, however, increasingly likely that this will change: humanity is very likely to create a powerful artificial general intelligence (AGI) before the year 2100 (Muehlhauser & Salamon, 2012). Such an intelligence might, for a short time, be about as intelligent as a human. However, there is widespread agreement among experts that such a creation would become recursively self-improving shortly after its creation, thus undergoing an "intelligence explosion" (Muehlhauser & Salamon, 2012). Such an intelligence explosion would very likely lead to a *foom*[2]: an event that, one way or another, destroys the value of money and previous investments, restructures the status quo, and dramatically transforms life as we know it. A foom consists primarily in an *artificial intelligence explosion*[3], and manifests as one of three scenarios (one good and two bad)[4] which I will discuss. Going foom is the result of the AGI's ability to learn and to edit its own source code; these two abilities together allow it to learn how to learn more efficiently. Whereas a human can, for example, merely learn that they have a certain bias and try to remain cognizant of it, an AGI could write the bias out of its code. This, combined with general intelligence and the advantages that a computer brings to cognition[5], means that beyond a certain point in AGI development[6], it is more-or-less impossible to prevent a foom, let alone stop one which is in progress. Therefore, humanity really only has one chance to create a friendly AI[7]. If we mess up, we die.

---

[2] This term from prominent AI theorist Eliezer Yudkowsky.

[3] Notably, all fooms are artificial intelligence explosions, but not all artificial intelligence explosions are necessarily fooms.

[4] While theoretically each type of foom could come in varying degrees of severity at first, I will set this notion aside. This is because, as I will discuss, a foom more-or-less cannot be stopped; thus, a bad foom will *eventually* kill everyone.

[5] Including, but not limited to: processing speed; parallel processing, aka *actual* multitasking; not needing to eat or sleep; perfect memory; and a data-gathering ability limited only by the presence of the internet.

[6] Precisely where this point lies is irrelevant to my argument, as it is highly probable that the critical point lies toward the lower end of possible power of intelligences. Once an AGI reaches this point, it is probably able to kill us off.

[7] The term "friendly AI", also from Yudkowsky, is perhaps somewhat misleading. An AGI need not actually be friendly: we don't need to get along. Rather, it merely needs to have the same goals we do so that it doesn't destroy us. A friendly AI works for the betterment of humanity.

The massive existential risk posed by the bad fooms presents us with the Control Problem: How do we create a friendly artificial intelligence?

In this paper, I present a new answer to the Control Problem. Previous answers are Asimovian, or rule-based. They seek to control AGI through a set of pre-programmed rules, like Isaac Asimov's Three Laws of Robotics. In such systems, the most basic guide of behavior is this rule set. Such systems focus less on morality than they do on following orders and attempting to simulate common sense. In contrast, my proposed system is property-based. Rather than giving a set of rules, it gives the AGI a set of properties which are built into it in such a way as to necessitate their use in its decision-making processes. These properties allow any sufficiently intelligent agent to learn human-like morality. It is important to note, however, that I am a philosopher, not an engineer. As such, my solution to the Control Problem is purely theoretical in nature.

I will argue that previous answers, which are Asimovian in nature, are inadequate, as any such rule system is either uselessly broad or can be circumvented by even a moderate human intelligence. Rather, what we need is to create an AGI which possesses a human-like moral psychology. Importantly, human moral psychology is not Asimovian and thus cannot be worked around or lead to suboptimal scenarios (such as the bad fooms) without other substantial system failures beforehand. I will then present an account of the initial state of the human moral faculty, building upon views put forward by Mikhail.

Fooms

To make clear the necessity of solving the Control Problem, I will consider the apparent inevitability of AGI and three general types of fooms.

Moore's Law has held for decades, and though it seems that it has finally failed, this does not by any means suggest that we have reached the maximum possible computing power (Simonite, 2016). Google suggested that it can achieve "quantum supremacy" in 2017; Google has both developed a quantum computer and believes that, within the year, it will be able to perform certain kinds of calculations that no other computer can (Simonite, 2017). Though they have yet to formally release technical data regarding their 6-qubit chip, such a success would pave the way for large-scale quantum computing. It seems not unreasonable to assume that a large-scale quantum computer would have more than enough processing power to support an AGI with human-power intelligence; it is certainly a reasonable assumption that Google and other companies will continue pursuing quantum computing (among other kinds[8]), eventually reaching the requisite power for supporting human-level AGI. This progression can't be stopped.

Artificial General Intelligence provides immense first-mover incentives[9]; If AGI research were outlawed in one country, all the researchers would simply move to another. Given that such research has already been undertaken around the globe, if an AGI can be created, it probably will be. With this in mind, we should consider the possible types of fooms.

The worst-case foom is the *Terminator Scenario* (Parsons, 2015). In this case, humanity creates an AGI that is, for whatever reason, malevolent. Though this foom may begin as a mere dystopia, a malevolent AGI will eventually be able to kill humanity off; furthermore, humanity probably dies quite quickly in this scenario, as the AGI actively dislikes us. To briefly illustrate

[8] Quantum computing is not the only means by which we might realize a powerful AGI. A sufficiently advanced neural network, for example, could also become a suitably powerful AGI.
[9] Benefits to being the first person to possess a particular thing. A good comparison is nuclear weapons: the first to have them has a substantial advantage in terms of military power. The same is true for AGI.

the methods by which it might annihilate us, consider the following possibilities. A sufficiently advanced AGI could solve the problems surrounding molecular nanotechnology (though presently impossible in practice, there is significant evidence to indicate its theoretical possibility (National Research Council, 2006; Foresight Nanotech Institute, 2007), thereby allowing the AGI to consume much of the planet and repurpose the material, humanity included. Alternately, it might create a bioweapon and kill us all with an unstoppable plague. A final example is the method used by Skynet in the original *Terminator* movie: an AGI could turn our own nuclear weapons against us. However this foom played out, humanity would be doomed.

The second scenario is that of the *Paperclip Maximizer*, which was first described by Bostrom (2003). Broadly put, this scenario is the result of an AGI which moves towards one goal with too few constraints on its actions. This lack of proper constraint leads the AGI to destroy humanity as a byproduct of its actions, because it was never told to preserve us. Consider, for example, an AGI which is given one goal: maximize the number of paperclips it possesses[10]. At first, it might make do with whatever materials and machines are available in its immediate vicinity, assuming it comes to life in some server farm. It will quickly graduate to fully-automated factories, while at the same time working to make itself more intelligent, because greater intelligence will help it generate more paperclips. At some point, humanity will be run over by the AGI in its quest for ever more paperclips in much the same way that we would pave over an ant colony when building a new road. In this case "the AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else" (Yudkowsky, 2008). Clippy, it seems, has the last laugh.

The final scenario is the only positive one. This is the *Friendly AI Scenario*, in which we create an AGI and our creation is cooperative. Here, because of the enormous power at hand,

[10] Its goal need not be something so mundane, as Bostrom notes; consider, for example, an investment-focused AGI that maximizes its controller's stock portfolio by investing in defense futures and starting a major war.

unfathomable advances are made in every possible field, and immense benefits for humanity are easily imagined. In this scenario the end result is less clear than the others, because there is no definite end-state for humanity; whereas in the bad fooms humanity is killed off, the upper bounds of the good fooms seem only to be limited by imagination and physical possibility.

This, I think, shows why we need to solve the Control Problem. Even if the good foom progresses quite slowly, it seems highly preferable to the extinction of the human species.

The Problems with Asimovian Rule Systems

An Asimovian system is rule-based: that is, the rules of the system are the most-basic guide of the agent's behavior, and every action is measured against whether or not it follows the rules. These rules determine such things as how the agent assigns values, what their[11] ultimate goals are, and what limitations are placed on them in attaining those goals. As Asimov conceived of such a system, the rules could not be modified, though this is not necessarily the case for all such systems.

The Three Laws of Robotics

Isaac Asimov's Three Laws of Robotics[12], or the Three Laws, are the most famous solution to the Control Problem, and one of the oldest. The Laws, as he first proposed them in the collection of short stories *I, Robot*, are (Asimov, 1950):

1. A robot may not harm a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Though many have already written on the flaws of the Three Laws[13], including Asimov himself, I will analyze four problems with the Laws, eventually showing that following the Laws leads either to decision paralysis or unintended and sub-optimal consequences. Using these

[11] Asimov supposed that any robot which contained the Three Laws was sentient. I see no reason to drop this supposition; therefore, I will use "they" as the singular, genderless pronoun for robots in this section. It seems inappropriate to call a sentient being "it."
[12] In discussing Asimovian systems I will use the terms "robot" and "AGI" interchangeably.
[13] See, for example, Anderson, S.L. AI & Soc (2008) 22: 477.

objections, I will show why Asimovian systems in general are ineffective as the basis for machine ethics, at least if we want AGI to have human-like morality.

The Problem of Physical Distance

The first problem I will consider is that of physical distance which, though closely tied to the Frame Problem[14], seems different enough to merit its own section. Precisely put, the Laws treat physical distance as nonexistent, effectively making a robot's obligation to a human five feet away and one across the planet the same. This problem leads to decision paralysis, as the Three Laws give the robot no way to differentiate between these obligations.

For a concrete example of this problem, consider a robot working in a strip mine. If the robot does their job the mine will expand, creating an ever-growing hole in the ground. However, strip mining is incredibly destructive, as (among other things) it eliminates wildlife habitat and generates harmful runoff[15]. This will harm the environment in a broad way which is bad for many people, though in ways which are hard to precisely value, as well as more-directly harming people immediately downstream from the mine, as they will drink polluted water or use it to irrigate their crops. On the other hand, if they do not do their job, the mine will produce less, which hurts their employers[16] economically. Which of these should it value more? The Three Laws give a robot no way to figure out how to weigh these variables, and any choice creates harm. With no system to guide them, this quickly leads to decision paralysis.

---

[14] I will discuss the frame problem in the upcoming section.

[15] See http://www.greenpeace.org/international/en/campaigns/climate-change/coal/Coal-mining-impacts/ for a broader discussion of the effects of strip mining.

[16] There is an interesting discussion to be had regarding whether we would rather use the term "employers" or "owners" here. It is highly plausible that sentience is required for human-like morality, though non-sentient creatures can and do seem to have their own concepts of justice: monkeys, for example, appear to have a concept of fairness. Furthermore, since Asimov's work (and mine, here) both concern sentient machines, we could easily assume our mine-worker robot is sentient. If this were the case, could it be owned? That is, is is morally permissible to own a sentient creature even when the creature is non-biological? My intuition is that it isn't, but this issue must be set aside. A proper discussion would wander too far afield.

This would be a problem if it only presented itself with larger-scale issues, as it might with the strip-mining robot, but it seems fairly obvious that it would result from *any* action undertaken by a robot. A robot could easily spend all of their time computing whether it was acceptable to merely leave their charging station because of the sheer number of calculations involved in determining if and how they might hurt someone and which of those harms should be valued most. Not only is this problematic in and of itself, but it seems indicative of a greater problem: not only do the Three Laws fail to address the Frame Problem, they exacerbate it.

The Frame Problem

The Frame Problem asks how a mind with many beliefs about the world updates those beliefs with respect to new information gained or actions taken such that those beliefs remain "roughly faithful to the world" (Dennett, 1984). Though there are solutions for the Frame Problem within various systems of logic (Shanahan, 2003; Lifschitz, 2015), epistemology has yet to reach a suitable solution. This presents a problem for the creation of AGI, as a general solution is probably necessary for the construction of a general intelligence. If we are to find a general solution to the frame problem, there are two aspects of belief-system updates ("system updates", or "updates") which need to be understood: how the mind knows which beliefs to check with each update and when to update its beliefs. Under the Three Laws, a system update not only includes checking each belief for relevance to the present situation, but against the Laws themselves in order to ensure compliance. In this section I will show that the Three Laws make any system using them unable to solve the frame problem.

The Three Laws give no guide for knowing which beliefs to check with each system update; this is problematic, because it is impossible that a mind with any substantial (or useful) number of beliefs which checked each of its beliefs with each system check could ever come to have even remotely accurate beliefs about the world. Imagine that the average person has

2,000 beliefs (a low estimate, I think, given how many things the average person is able to talk about and do) and that it takes the mind 1/100 of a second to check each belief for relevance. If this were the case, it would take 20 seconds to check the entire belief system every time a belief needed to be updated. That is absurd. A mind which worked this way would never get anything done.

This is particularly problematic for implementation of the Three Laws because, when writing a rule, not mentioning a thing or set of things implicitly expresses indifference over that thing or set. Since the Three Laws make no mention of the Frame Problem, they express indifference to it. Thus, instead of helping the robot know which types of beliefs they do not need to check with each update, it appears to mandate an entire system check with each action. Checks would likely require even more time than merely checking the belief system, because checking for conflicts between the Laws can require deeply philosophical thought. A robot would only be allowed to act if they knew what it meant to harm someone, if some harms are acceptable if they lead to future goods, if some harms are worse than others, and so on. Clearly, in addition to providing no guidance on which facts or beliefs to check with each new input, the Three Laws add an additional burden of substantial and constant checks of the Laws. A mere belief-system check could easily turn a robot into an expensive mannequin.

In humans, it seems reasonable to assume that many system updates are, while not infrequent, also not constant: though some beliefs must be updated constantly, such as those about the relative location of each of one's fingers, most others, such as those about one's age or the position of Jupiter relative to the Sun, need not be updated often. Unfortunately, under the Three Laws, a robot would be obligated to check every belief during each update to ensure that no beliefs would lead to conflicts with the Laws. Therefore, absent the addition of Laws 4, 5, 6, etc. to govern system updates, a robot would have to check every belief after each action or new piece of information input to ensure that every belief was accurate and that none of them conflicted with the Three Laws. Even with such an addition, it seems unlikely that a robot would

be able to avoid checking all of their beliefs often, because additional laws could not override the first three without fundamentally undermining their purpose.

To illustrate the problem, let's return to our strip-mining robot. Imagine, for a moment, that they have somehow gotten past their decision paralysis and have gotten to work. The robot has scooped up some earth with a mining machine. What moral beliefs do they update? If the Three Laws are the only guide, it seems that the robot must update hundreds of their beliefs and they must check each of these for relevance. After all, moving the machine full of earth *could* hurt someone (and, by some measures, probably does), and the robot needs to know if and how if they are to proceed. Perhaps everything is fine, so the robot adjusts the machine's bucket another foot. Time for another system check! The Three Laws give no instruction as to when the beliefs should be checked, so the robot should check often to be sure it does not violate them; after every action seems appropriate. Which beliefs to check? The Three Laws give no guidance, so they must check all of them, as the robot has no idea which may be affected by what actions. Clearly our robot once again becomes frozen in contemplation and system updates.

The Necessity of Common Sense

The third problem I will consider with Asimovian systems in general is that they require common sense[17] to interpret. In effect, Asimovian systems require a "Do-What-I-Mean" (or DWIM) program in order to be useful and, if one has developed a DWIM program, they have presumably already solved the Control Problem and the need for Asimovian rule systems falls

[17] For my purposes, common sense is a learned, though not explicitly taught, understanding of what is reasonable or appropriate which is shared by most people. Most people can be reasonably expected to possess common sense. It may sometimes be cultural: most people who speak a particular language to a high degree of fluency will understand puns in that language, for example, while those who are just learning that language won't. I will suppose that, in a very general way, the ability to reason pragmatically contributes to one's possession of common sense. My thanks to Jasper Olsen, Will Criste, and Sari Hoffman-Dachelet for their help in clarifying this definition.

away. Put differently, the ability to perform pragmatic reasoning is necessary, though not sufficient, for moral reasoning. Interpreting an Asimovian ruleset the way we would want it to be interpreted requires both pragmatic reasoning and something else that allows one to perform moral reasoning[18], and if a system has those things, the rules aren't needed.

To illustrate the necessity of common sense, consider the following joke:

> One day, a computer programmer is about to leave the house to run some errands. As he's leaving, his wife says: "While you're at the store, get a gallon of milk. If they have eggs, get a dozen."
>
> The programmer never goes home.[19]

For the sake of argument, allow me to kill the joke. The programmer's wife has given him no directions after "if they have eggs, get a dozen". Therefore, if he behaves like a computer program, if the store has eggs, he gets a dozen eggs (or a dozen gallons of milk, given the unclear referent of the phrase "get a dozen"), but after that he doesn't know what to do. Thus, the programmer could either have:

1. Gone to the store, gotten a gallon of milk and a dozen eggs, and remained in place by the eggs.
2. Gone to the store, gotten one gallon of milk, checked to see if there were eggs, confirmed that there were eggs, and thus gotten a dozen gallons of milk because the referent of "get a dozen" *could be* the milk (even though we understand it not to be) and remained in place by the milk.

Assuming the programmer was instead an AGI which understood the request to go to the store and get milk and eggs but lacked common sense, it becomes entirely reasonable to expect this joke to play out. Understanding that one returns home after going to the store or that the wife means for her husband to get a dozen eggs if the store has them requires that one

[18] As I will later argue, my theory composes this "something else".
[19] Alternately; "The programmer goes to the store and buys a dozen gallons of milk. He never goes home."

possess a common-sense understanding of English and the concepts involved in the sentences.

Computers have no such implicit knowledge. A computer does exactly what you tell it to do and no more. Given the wife's request, it seems likely that a non-DWIM AGI would go to the store, pick up but not purchase milk and eggs, and then just sit there doing nothing rather than doing what she wanted it to do. It might be tempting to try and give the AGI more rules to solve this problem; however, as I will show in the next section, doing so leads to an infinite rule explosion.

It's apparent that understanding language as it is used in everyday life requires an enormous amount of knowledge and inferential capacity. Because of this, the Three Laws (and Asimovian systems generally) create many problems. Consider the First Law's prohibition against harming humans. A whole slew of questions arise from this, almost none of which have even generally-accepted answers. For example, what is it to be human? What is harm? Are some harms acceptable if they lead to gains later? If they are, what kinds of gains, and why? What is it to do harm? To what lengths must one go to avoid harming another? If it is impossible to prevent at least one human from being harmed, which human should you harm to spare other humans? Entire books could be (and have been) written about each of these questions, and the philosophical community has not reached a broad consensus about any of them. If we were to hope for any system to understand what is meant by the First Law, we would need it not only to have the moral intuitions of the average human, but to have made significant progress in ethics.

This problem gets worse with the inclusion of the Second and Third Laws. With the second, the robot must, among other things, understand what an order is, what it means to obey an order, what it means to contradict, and what each order it is given means, as well as how to obey such orders. Once we add in the Third Law, the robot must also understand what it means to exist, what it means to protect their own existence, how one protects oneself, and how one

might do so without contradicting the First or Second Laws. For all three Laws the robot must have a self-concept and knowledge of cause and effect, among other things.

It seems apparent that, in order to follow the Three Laws, we would first need to solve many major philosophical problems; otherwise, we would be tasked with somehow creating an intelligence which did what we meant it to do *even when we didn't really know what that was*. Though not technically impossible, it seems ridiculously improbable that we could ever create a DWIM AGI which also used the Three Laws as its moral guidance system and knew even better than we did what we wanted it to do.

There is, however, a larger problem for Asimovian systems when it comes to common sense. Imagine that there existed a perfect Asimovian system: a finite and manageable set of rules (a set composed of a googol, or 10^{100}, of rules is finite, but obviously not usefully finite) which, if all followed, lead to human-like morality. This system covers all possible cases. Of what kind of rules might such a system be composed? Since one of the problems with the Three Laws is that they are too broad and open to interpretation, a perfect Asimovian system would probably consist of relatively simple laws which together yield complexity. Such laws might concern what ethics is, what moral agents are, what has moral standing, acceptable kinds of actions, types of scenarios, and how to assign values.

Such a system *still* requires a DWIM intelligence in order to operate, because in order to run an Asimovian system one needs to be able to understand what is meant, rather than what is said. This system would have the definitions necessary to know what each of the words in the rules meant and what they meant *in a very literal sense* when put into the sentences that composed each rule, but it requires common sense in order to interpret the rules and definitions in a productive way. If a computer can productively interpret an Asimovian ruleset, it must be DWIM, and if it is DWIM, it has no need of such a ruleset.

In this final section, I will turn to the main problem with the above hypothetical perfect Asimovian ruleset: creating such a set with a finite number of rules seems impossible.

The Problem of Infinite Rules

The final and most difficult problem to overcome for Asimovian systems is that creating such a system seems to lead to a sort of rule explosion: the creation of infinite rules in an attempt to cover all possible scenarios. In this section, I will draw upon the blog post *Bringing Precision to the AI Safety Discussion* and corresponding paper, *Concrete Problems in AI Safety*, which were published by Google in 2016 (Olah, 2016; Amodei et al., 2016)[20]. These put forward what the authors refer to as the "five key problems" in AGI safety. I agree with the authors' general stance toward these problems, and as such, I will use these problems to illustrate, in a broader sense, the futility and risk involved in trying to create an Asimovian system. As put forward in the blog post, these problems are:

- **Avoiding Negative Side Effects**: How can we ensure that an AI system will not disturb its environment in negative ways while pursuing its goals, e.g. a cleaning robot knocking over a vase because it can clean faster by doing so?
- **Avoiding Reward Hacking**: How can we avoid gaming of the reward function? For example, we don't want this cleaning robot simply covering over messes with materials it can't see through.
- **Scalable Oversight**: How can we efficiently ensure that a given AI system respects aspects of the objective that are too expensive to be frequently evaluated during training? For example, if an AI system gets human feedback as it performs a task, it needs to use that feedback efficiently because asking too often would be annoying.
- **Safe Exploration**: How do we ensure that an AI system doesn't make exploratory moves with very negative repercussions? For example, maybe a cleaning robot should experiment with mopping strategies, but clearly it shouldn't try putting a wet mop in an electrical outlet.
- **Robustness to Distributional Shift**: How do we ensure that an AI system recognizes, and behaves robustly, when it's in an environment very different from its training environment? For example, heuristics learned for a factory workfloor may not be safe enough for an office.

[20] The blog post can be found at https://research.googleblog.com/2016/06/bringing-precision-to-ai-safety.html.

Though each of these problems is important in its own right, I need only consider the first one in great detail: Avoiding Negative Side Effects. I will attempt to solve this problem by creating a ruleset; eventually I will show that solving any one of these problems requires an infinite ruleset and that addressing all five simultaneously does not make the rule system finite, let alone usefully finite. I will also use the hypothetical provided by Amodei et al. which, while generally lacking moral content, is more than suitable as a demonstrative device.

Imagine a cleaning robot. It[21] has been tasked with moving a box from one side of the room to the other. Let's also assume (and it's a massive assumption, as I've already shown) that the robot is a moderately-DWIM program. It knows what we want it to do, within reason, and does that. However, as with a child who is being difficult, the robot sometimes gets things wrong (though the robot may not do it intentionally, as we will see). Even with this restriction in place[22], the rule system quickly grows out of control. Absent any restrictions, the robot could pick it up, push it along, or throw it. Any of these might well complete the task of "moving the box", but only one is remotely close to helpful: picking up the box and carrying it. We might well suppose that this is because whatever is in the box is fragile, and we don't want to damage whatever is in the box. So, let's create a rule: *do not damage task items.*

Our robot does not harm its cargo now (or any tools it uses for its tasks, assuming we properly define "task items"), but it does not care about anything else yet. If there is a vase in the way, our robot will not mind if it knocks the vase over because the vase is not a task item. By ignoring the vase in our rules, we have implicitly told our robot that everything else is irrelevant, so we need another rule: *disturb the surrounding environment as little as possible.* Unfortunately, as Amodei et al. note, a restriction of this type leads to a different problem: the robot will resist any and all change to the surrounding environment. To fix this, another rule:

---

[21] Returning to "it" as the pronoun of choice because this robot need not be sentient for this argument.
[22] Being DWIM seems to be a restriction in this case because otherwise we must create a set of rules for simulating common sense. Unfortunately, infinity minus a finite number is still infinity.

allow humans to change the surrounding environment. Of course, this has obvious problems: the robot now will not stop any change, even harmful change, from happening around it, so long as that change is caused by a human. Another rule, then: *resist unreasonable change to the environment*. What is “unreasonable”, here? Even many people, who know what we mean (though not always completely DWIM, we might say that other people are know-what-I-mean) would disagree on what qualifies as “unreasonable”. Several laws are then required to define unreasonable change. There is yet a further problem, however: how should the robot resist the change? It might become violent: we did not specify peaceful resistance. Furthermore, what qualifies as the environment? This rule runs directly into the Frame Problem, even within DWIM systems.

Clearly the ruleset is massive, and trying to solve it alongside another of the problems merely introduces more rules. Consider the second problem: an AGI should avoid reward hacking. If we tell our robot to clean any messes it can see, then it might come to realize that if it cannot see any messes, it is doing fine (in a certain sense) and therefore close its eyes or cover the messes up (Amodei et al., 2016). We first have to tell it what a mess is: rules would be needed to help it differentiate between a folder full of papers, a lost cell phone, and a pile of trash. We then have to somehow specify that it must actively seek out but not create messes so that it does not merely stare at the nearest wall.

The robot might have to move things in order to do its job. Does this constitute unreasonable change to the environment, per the rules from the first problem? Perhaps, but perhaps not. Additional rules are probably needed to clarify, as “unreasonable” probably differs for each object and scenario: the trash goes to the nearest trash can, the folder probably stays as close to where it originated (or is moved and then returned), and the phone is probably taken to the lost-and-found or equivalent place. Do we create a rule for each of these? Even for just phones this seems problematic: the robot would need to have rules for knowing when a phone

was lost, and whose phone belonged to whom, where to take lost phones, and so on and so forth.

Several of our previous rules create a potential problem for our robot now: if it is to allow humans to change the environment, not to allow unreasonable change to the environment, and is to clean any messes it sees, it can be expected to resist any task (even one undertaken by a human) which causes a mess by necessity unless it checks with the human beforehand to see if the task qualifies as reasonable. If it is to know this, however, it must have a rule for knowing when to check in with a human, and for what information, and with whom it should check in.

This, I think, is enough to demonstrate that constructing an Asimovian system results in an infinite ruleset; however, I should show why this is both an impossibility to create and to operate under. Fortunately, this is an easy case to make. Making rules takes time. No matter how infinitesimal the amount of time required to create a single rule, the total amount of required time is infinite. Similarly, no matter how little room it takes to store each rule, it will take infinite space. We have neither infinite space to work with nor infinite time to spare. It seems, therefore, that a project which requires both is, if not impossible, then undesirable.

Mikhail's Theory of Universal Moral Grammar

I have shown that rules-based systems cannot yield sufficiently human-like morality. So, if rules are insufficient, we need something else. I contend that this something else is John Mikhail's theory of Universal Moral Grammar (UMG), which he describes in *Elements of Moral Cognition*. Mikhail argues for and expands upon Rawls' linguistic analogy from *A Theory of Justice*, providing a longer and more formal explanation and defense of the theory than had previously been forwarded (Mikhail, 2011). In this section, I will briefly review Mikhail's work and argue that a Mikhailian system is a suitable basis for human moral psychology; then, building on this theory, I will put forward my own version of what Mikhail calls an *explanatorily accurate* theory of UMG.

The Linguistic Analogy

MIkhail's theory of UMG is built upon Rawls' (1971) Linguistic Analogy, which he contends Rawls embraces in discussing the "sense of justice" (p 46). The Linguistic Analogy makes a comparison between the early work of Chomsky, particularly his theory of Universal Grammar, and moral theory. Chomsky argued that humans are genetically programmed to be able to learn language; Rawls thought that humans were genetically programmed to be able to learn morality. Put differently, the thought is that, though a sense of morality is not innate, the necessary mental capacities for learning it are.

Chomsky studied language by pursuing the answers to three questions (Mikhail, 2011):

1) Chomsky's Questions:

(a) What constitutes knowledge of language?

(b) How is knowledge of language acquired?

(c) How is knowledge of language put to use?

The answer to 1(a) is a theory of *linguistic competence*, or *generative grammar*: "a theory of the steady state of the mind... of a person who 'knows'... a particular natural language like English, Hebrew, Arabic, or Japanese" (Mikhail, 2011, p. 14). Question 1(b) is answered by *Universal Grammar* (UG), a theory which describes the initial state[23] of the language-acquiring ability of the mind and a description of how the properties postulated by UG "interact with experience to yield knowledge of a particular language" (Mikhail, 2011, p. 14). Question 1(c) is answered by a theory of *linguistic performance*: "a theory of how knowledge of language enters into the actual expression and interpretation of language specimens, as well as into interpersonal communication and other actual uses of language" (Chomsky, 1965, p. 4).

Mikhail argues that, when rephrased for the topic, these questions can help us organize a theory of moral cognition:

2) Mikhail's Questions:

(a) What constitutes moral knowledge?

(b) How is moral knowledge acquired?

(c) How is moral knowledge put to use?

These questions are answered much like their parent questions. Question 2(a) is answered through a *generative moral grammar* or theory of *moral competence*: "a theory of the steady state or acquired state of the mind...of a person who possesses a system of moral knowledge…" (Mikhail, 2011, p. 15). Such an answer is *descriptively adequate*; as such, question 2(a) can also be thought of as the problem of *descriptive adequacy*. Questions 2(b) can be thought of as the problem of *explanatory accuracy*. The answer is provided by *Universal Moral Grammar* (UMG). This is a theory of the initial state of the "moral faculty", which is assumed to be a "distinct subsystem" of the mind; this is accompanied by an account of "how

[23] The terms "initial state" and "steady state" are technical terms from linguistics. To paraphrase Chomsky's explanation, the initial state is a genetically-determined state of the mind which is common to the human species with "at most minor variations apart from pathology." The mind is then trained through experience, thereby achieving a steady state at a fairly fixed age; this state then changes only in minor ways (Chomsky, 1980, p. 187-188).

the properties UMG postulates interact with experience to yield a mature system of moral knowledge" (Mikhail, 2011). This answer is *explanatorily accurate*. A theory of *moral performance* answers 2(c), or the problem of *observational accuracy*. This *observationally accurate* theory describes how moral knowledge is used in "the representation and evaluation of human acts...and other forms of actual behavior" (Mikhail, 2011).

Mikhail's questions, 2(a-c), are (as he notes) empirical; furthermore, science understands little about the structures or processes that might answer these questions. We must, therefore, give fairly tentative answers. Even so, he supposes that the answers would have the following general form:

> 2(a) The normal individual possesses a *moral grammar*: "a complex and largely unconscious system of moral rules, concepts, and principles that generates and relates mental representations of various types."
>
> 2(b) This grammar is learned/acquired as an effect of a genetic program and moderate environmental inputs. The solution describes the initial state of the mind provided by the genetic program and the manner in which this interacts with the environmental inputs to produce the moral grammar.
>
> 2(c) The solution to this question has two parts: the *production problem* and the *perception problem*. The production problem is answered by an account of how individuals apply their moral knowledge to their daily lives[24]. The perception problem, on the other hand, would be answered by an account of how individuals parse actions and circumstances such that their moral grammar can assign it a structural description that specifies its properties. (Mikhail, 2011)

In describing these answers, Mikhail describes a universal moral grammar; I will explain his theory of UMG momentarily. However, in order to understand how the answers to the

[24] Mikhail holds that it may well be the case that such an account may be beyond the ability of science (or human intelligence) to produce.

linguistic and moral questions are formulated and understand the constituent parts of a linguistic or moral grammar, several distinctions must be drawn.

The first of these is the distinction between two theoretical types of grammar: an *observationally accurate* grammar and a *descriptively accurate* one. An *observationally accurate grammar* system (linguistic or moral) lists every possible sentence or situation possible in the system. For linguistics, this describes every possible well-formed sentence in the relevant language; in ethics, this enumerates every possible situation and its deontic status (forbidden, permitted, obligatory, etc.). Such a list is almost certainly infinite in either case. In contrast, a *descriptively accurate grammar* is the set of principles and/or rules which, when combined with knowledge of the circumstances and our beliefs, lead to well-formed sentences (for linguistics) or judgments (for ethics). Only a descriptively accurate grammar is interesting, as an observationally accurate grammar would imply that each individual possessed the infinite list. That is obviously absurd: the mind does not have infinite space. Furthermore, under an observationally accurate grammar, the process of checking for grammaticality would be a mere comparison to each item of the list.

The second distinction is the Competence-Performance Distinction. To properly understand this distinction, one should know Rawls' concept of a *considered judgment*. As Rawls says:

> These are judgments in which our moral capacities are most likely to be displayed without distortion...we can discard those judgments made with hesitation...in which we have little confidence…[and] those given when we are upset or frightened, or when we stand to gain one way or the other...Considered judgments are simply those rendered under conditions favorable to the exercise of the sense of justice, and therefore in circumstances where the more common excuses and explanations for making a mistake do not obtain. (Rawls, 1971)

This passage draws a distinction between *moral performance*, or the judgments one actually makes, and *moral competence*, or the cognitive system underlying those judgments. This is similar to a distinction Chomsky draws in *Aspects of a Theory of Syntax* (Chomsky, 1965). Furthermore, Rawls argues that "competence, not performance, [is] the moral philosopher's proper object of inquiry" (Mikhail, 2011).

The Competence-Performance Distinction is particularly relevant because "moral judgment is a flexible, context-dependent process, which cannot be accurately described by simple consequentialist or deontological principles, and which is clearly subject to framing effects and other familiar manipulations" (Mikhail, 2011). On Mikhail's view this does not indicate that pursuing considered judgments is a pointless task; his position is that these facts reinforce the need to adopt the position that moral judgments reflect an ideal core human competence and we merely make mistakes in applying this competence because of psychological limitations, performance errors, or other exogenous factors (Mikhail, 2011).

Less formally, this is the difference between 2(a) and 2(c), or what someone knows versus what they do. It is particularly useful to the Linguistic Analogy to clarify different parts of the research program. Hereafter I will use Mikhail's definitions for *moral competence* and *moral performance*, where the former refers to an individual's moral knowledge and the latter to how that knowledge is used (Mikhail, 2011).

The final distinction to attend to is between *operative* and *express* principles. *Operative principles* are those which an individual actually follows in making decisions: these are the principles of their *moral competence*. It is assumed that the average person is unaware of these principles, cannot become aware of them through introspection, or that their statements about them are necessarily accurate (Mikhail, 2011). It is assumed that operative principles are, in fact, below the level of conscious (or even potential conscious) thought. *Express principles*, on the other hand, are those which a person "verbalizes in the attempt to describe, explain, or justify [their] moral judgments" (Mikhhail, 2011, p. 20).

With these distinctions in place, I can now explain MIkhail's argument for UMG; that is, his argument that his answers to 2(a-c) constitute the correct answers.

The Argument for Moral Grammar

Mikhail's *argument for universal moral grammar* is abductive, or an argument to the best explanation. It rests on one key observation: that we are able to make a "potentially infinite number and variety of judgments" about the moral content and deontic status of "acts, agents, institutional arrangements" in novel situations (Mikhail, 2011, p. 46). Notably, these situations are different from the finite number to which the individual has been exposed previously. Given that the mind has a finite storage capacity, it follows that rather than storing the representation of each situation in the mind, the mind must have some system, like a set of principles, properties, or rules, which allows it to come to these judgments (Mikhail, 2011). This system, he contends, is a *descriptively* and *explanatorily accurate* account of UMG.

There is moderate empirical support for the theory of UMG from several fields, including comparative linguistics, cognitive neuroscience, developmental psychology, and legal anthropology (Mikhail, 2011). For example, every natural language has "words or phrases to express basic deontic concepts, such as *must not*, *may*, or *must*" (Mikhail, 2011, p. 104). These words' "natural domain of application is the voluntary acts and omissions of moral agents" (Mikhail, 2011, p. 104). This implies that such concepts are universal across cultures. Further, some functional imaging studies have lead researchers to conclude that a "fairly consistent network of brain regions is involved in moral cognition", though these findings are "both preliminary and controversial" (Mikhail, 2011, 105). If successfully replicated, such findings would provide significant evidence to support the idea that there are sections of the brain devoted to moral cognition, which is a central posit of UMG. Developmental psychologists have found that the intuitive judgments of young children are remarkably complex, even showing

"many characteristics of a well-developed legal code" (Mikhail, 2011, 104). Three-to-four-year-old children distinguish between moral violations and violations of social convention as well as using intent to distinguish between actions with the same outcome; four-to-five-year-old children use "a proportionality principle to determine the correct level of punishment for principles and accessories" (Mikhail, 2011, 104). Given that these children had no special tutoring in ethics, these facts are significant evidence in favor of an innate ability to perform moral reasoning. Finally, murder, rape, and similar acts of aggression are universally (or nearly universally) prohibited; the legal distinctions based on cause, intent, and voluntary behavior are similarly common (Mikhail, 2011).

Available empirical evidence suggests an innate, universally present moral capacity such as the one postulated by UMG. The most popular alternate thesis is empiricist. This theory suggests that all of human moral thought is learned; however, this faces significant problems. Firstly, it fails to explain the universality in linguistics and legal thought and of certain moral concepts. Were moral knowledge exclusively learned through experience, it would be reasonable to assume that moral systems would vary greatly based on geographic location. Furthermore, the empiricist has trouble explaining how young children with no special training in ethics nevertheless produce such complex judgments, especially given their extremely limited experience. Finally, the empiricist position fails to account for the apparently systemicized nature of moral cognition in the brain. Though it does allow for it, the fact that UMG predicted these results counts against the empiricist.

An Explanatorily Accurate Theory of Universal Moral Grammar

The project of moral philosophy seems to be to discover the *operative principles* by which humans make their moral judgments. Unfortunately, I think we have been attacking the problem from the wrong end until now. Moral philosophers seem to have tried to discover the *operative principles* by introspecting from the *express principles*. An individual might say that they acted to maximize social utility or because their action followed the Categorical Imperative, but as we have already seen, operative principles are below the level of consciousness and *cannot* be introspected. It seems potentially futile to pursue *operative principles* in this way, especially given that we do not entirely know what kinds of mistakes or manipulations affect our moral judgments. Though analyzing individuals' *moral performance* is interesting, I do not think it the most productive course.

This is problematic for the creation of AGI, because if we want to create a friendly AI (FAI), and we do not know the principles by which we make ethical judgments, we cannot make an AGI with human-like moral psychology and thus cannot make FAI.

Therefore, I think we should start our search for the *operative principles* of human morality by searching for an *explanatorily accurate* theory of UMG. If there is some initial state of the mind which allows us to learn morality, it seems that this would be fairly uniform across all people and thus far easier to discover than the fairly complicated *express* or *operative* principles by which we make judgments. It also seems that it would be far easier to determine operative principles by applying known manipulations to the judgments they yield than by trying to backtrack through the manipulations.

I think that there are five key properties of human minds which form half of an *explanatorily accurate* theory of UMG as described by Mikhail. As he would put it, I provide half

of the answer to 2(b). I describe, in a theoretical way, the initial state of the moral faculty. The primary difference between my theory and the Asimovian systems which I attacked earlier is that in an Asimovian system the most basic guide of behavior is a ruleset, whereas in my system the most basic guide is a set of properties. These properties may yield rules, but the rules are malleable, unlike those in an Asimovian system. This, I think, helps account both for the variance in *moral performance* across many people as well as allowing for the infinite flexibility required of an *explanatorily accurate* moral grammar. With properties creating malleable rules, the system can adapt to new situations without the need for an ever-increasing and ever-more complicated ruleset. These properties are necessary and sufficient to create a framework upon which a suitably intelligent system can build human-like moral grammar. They are, in no particular order, as follows:

1. An understanding of agency which is self-applied
2. The recognition that other things which appear to be *minds-like-me* are *minds-like-me*
3. Having beliefs and desires
4. Appropriately attuned[25] empathy and sympathy
5. The ability to learn from experience and theorize about yet-to-be-experienced experiences

In the following sections I will explain two things. Firstly, I will give as precise a definition as I can of what each property is and what it allows an agent to do; secondly, I will show why this property is necessary for the creation of moral grammar.

[25] My thanks to Mark Phelan for his comments about tuning empathy and sympathy.

The First Property

Any system which is to make moral judgments must have a conception of agents and conceive of itself as being an agent[26]. Essentially, if it can tell that a suitably mature human is a moral agent and a tree is not, it has this property. This is necessary for the creation of moral grammar because any system which is to make moral calculations needs to understand what agents are. If it does not, it cannot understand who or what is performing actions, determine intent, or assign blame. Without these concepts, moral judgments cannot be made.

The Second Property

The second property is the recognition that entities which appear to be minds like mine are in fact *minds-like-me*. A *mind-like-me* is a mind of similar structure to my mind. This property is what allows us to identify other agents and things with moral standing[27]. It is necessary to emphasize that this property makes no assumptions about the content of other minds. To do so would cause catastrophic problems if one mind supposed that because it was in a certain state, every other mind like it was in that state too. Consider, for example, if one mind was suicidal and supposed that all other minds were this way as well. With a few other assumptions, such a mind could conclude that it should kill all other *minds-like-me*. This is obviously bad.

The supposition and recognition of *minds-like-me* is critical to the creation of human-like moral grammar because of the necessity of agents to moral judgments. This ties in very closely with the First Property because not only must an agent recognize itself as such, it must be able to recognize other minds in the world. If it cannot recognize other minds in the world, though it may make moral judgments, it cannot apply them. Effectively, a mind without the Second

---

[26] There are many complicated philosophical questions to answer about what it is to be an agent. Unfortunately, I do not have space for them here.

[27] I think it might well be this property that causes humans to adopt the Intentional Stance; sadly, there is not enough space to discuss this here.

Property is solipsistic by necessity and has little reason to be moral. In AGI, this could lead to a *paperclip maximizer*-like creation: not because it did not care about humans, but because it could not recognize humans as the things which it was told to care about.

The Third Property

The Third Property is the possession of beliefs and desires, including meta-desires. Since beliefs and desires are the subject of much discussion in the philosophical community, it seems prudent to attempt to clarify the sense in which I use these terms. An agent has a belief when there is something which it takes to be true or takes to be the case. Beliefs need not be consciously reflected upon; in fact, it seems probable that the vast majority of them are held unconsciously. Roughly speaking, an agent desires *p* when it is disposed to take whatever actions it believes are necessary to bring about *p*. We might posit many reasons for the agent desiring *p*, such as *p* being pleasurable or thought of as having inherent or as having instrumental value.

Beliefs and desires are necessary for the creation of human-like moral grammar for several reasons. Beliefs are necessary for the First and Second Properties, as otherwise a system cannot have a conception of agents or minds. Furthermore, beliefs allow the system to interact with the world in a productive way, as they give it a way to represent the world internally. Without such representations, a system cannot understand the world it inhabits and thus cannot act. Desires are necessary because they give reasons. More specifically, because of the Second and Fourth Properties, desires give the agent reasons to act morally.

Though in some instances desires could give an agent reasons to act immorally, I do not think this is necessarily a problem for my argument so long as those reasons are not the most motivating reasons for the agent and those reasons do not motivate the majority of people to act immorally. To consider a concrete (but highly simplified) example, I might desire to rob a corner

store because I want the money, but since I also believe that stealing is immoral and that one should not threaten lethal force except to protect oneself or others from immediate harm, the desire to rob the corner store does not give me the most motivating reasons. As such, I do not act upon them.

I take it to be fairly self-evident that, excepting extreme cases, having desires necessitates also desiring to survive: desired states of affairs cannot obtain if one does not exist, with the exception of desired states of affairs wherein one does not exist.

The Fourth Property

The Fourth Property consists in the appropriately tuned ability to empathize and sympathize. This wording is potentially misleading. Empathy, here, is an automatic reaction. It is the experience of an emotion because another agent is feeling that emotion. So, for example, I am being empathetic if I feel sad when I see that my friend is sad. Sympathy, on the other hand, is used intentionally. It is the ability to imagine what others feel by imagining oneself in another's situation. So, to sympathize, I might imagine that if my friend got a chocolate bar, they would be happy because I would be happy under similar circumstances.

This property is necessary for the creation of human-like moral grammar because it seems to yield the *operative principles* which contain prohibitions against acts of violence like murder and rape. Consider an intelligent system which has the First, Second, and Third Properties. Without the Fourth, it would recognize the existence of *minds-like-me* and their status as agents but would have no reason to care about them outside mere instrumental use. There is a specific kind of person which thinks this way:

a psychopath[28]. We do not want to create psychopaths. Without the Fourth Property, the best AGI we could hope to create would be a *paperclip maximizer*.

There is one critical flaw, I think, with human-realized empathy and sympathy: they do not properly scale. To illustrate this point, imagine that we can quantify feelings. When one person is hurt in our immediate vicinity, we feel a single unit of sadness. Call this one *sad*. If two people are hurt, we feel about two *sads*. However, if 1,000 people are hurt, we do not feel 1,000 *sads*. This does not make sense. Obviously there is an evolutionary advantage to empathy and sympathy that do not linearly scale: at some point, we might imagine that this would be crippling. However, it seems absurd that the 347th or the 851st person to be hurt makes us less sad than the first. Morally, there should be no difference.

Furthermore, human-realized empathy and sympathy fall victim to the Problem of Physical Distance. One person being hurt in my immediate vicinity has immediate, visceral pull. On the other hand, knowing that one person on the other side of the planet has been hurt in a similar fashion does not affect me in the same way at all. Again, there is a good evolutionary reason for this: we can only easily help those in our immediate vicinity, and expending the energy to help those far away would be a great drain on resources. However, it is not apparent that there is a significant moral difference between a person right next to me and one across the globe.

It might be the case that human-realized empathy and sympathy do, in fact, scale linearly, and that what I describe above is merely reflective of our *moral performance* rather than our *moral competence*. However, I am unsure of whether or not this is the case. Whatever the case, in implementing this property in AGI I think it would be critical to build in linearly-scaling empathy and sympathy. Not only would this yield more

[28] Amusingly, as Mark Phelan remarked, we might well conceive of psychopaths as effectively being human *paperclip maximizers*.

accurate *operative principles*, but I think properly-scaling empathy and sympathy would yield far more motivation to make the world a better place.

The Fifth Property

The Fifth Property is the ability to learn from experience and to extrapolate from available data to theorize about experiences which the agent has yet to experience.

This is particularly important for the implementation of the Third and Fourth Properties. For the Third, it allows the system to update its beliefs as well as yielding a better understanding of how to satisfy its desires; this Property includes a solution to the Frame Problem. With respect to the Fourth Property, the ability to learn from experience allows an agent to more accurately empathize and sympathize with other agents, especially with the ability to theorize about new experiences. It is this property which allows human moral grammar to adapt to infinite scenarios.

With these properties in mind, only one other thing is necessary for a complete *explanatorily accurate* theory of UMG: an account of how these properties interact with the inputs from the world to create a fully-realized moral grammar. I do not have such an account. However, I think it would be prudent to turn to developmental and child psychology for clues: if there is any current analogue to an untrained AGI with only these properties, it seems like it would be a very young child.

Conclusion

It is apparent that, despite all we do not know about human moral psychology, it cannot be based upon rules. I have shown that, because of this, Asimovian systems fail to describe human *moral competence* and that we should instead utilize a property-based *explanatorily accurate* theory of MIkhailian UMG like the one I described. Only with this framework can we ensure the creation of a a morally-competent AGI.

Works Cited

Amodei, D., Olah, C., Steinhart, J., Christiano, P., Schulman, J., Mane, D. (2016) Concrete Problems in AI Safety. *Google Research Blog.* From https://arxiv.org/pdf/1606.06565v1.pdf

Anderson, S. L. (2008). Asimov's "three laws of robotics" and machine metaethics. *AI & Society*, 22, 477-493.

Asimov, I. (1950). *I, Robot*. New York: Gnome Press

Bostrom, N. (2003). *Ethical Issues in Advanced Artificial Intelligence*. From http://www.nickbostrom.com/ethics/ai.html

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge: MIT Press.

Chomsky, N. (1980). *Rules and Representations*. The Hague: Mouton.

Dennett, D. (1984). Cognitive Wheels: The Frame Problem of AI. In Hookway, C. (Ed), *Minds, Machines and Evolution*. Cambridge: Cambridge University Press.

Foresight Nanotech Institute. (2007). *Productive Nanosystems: A Technology Roadmap*. From http://www.foresight.org/roadmaps/Nanotech_Roadmap_2007_main.pdf

Greenpeace International. (2016). About coal mining impacts. From http://www.greenpeace.org/international/en/campaigns/climate-change/coal/Coal-mining-impacts/

Lifschitz, V. (2015). *The Dramatic True Story of the Frame Default.* Journal of Philosophical Logic, 44: 163–196.

Mikhail, J. (2011). *Elements of Moral Cognition*. New York: Cambridge University Press

Muehlhauser, L., & Salamon, A. (2012). Intelligence Explosion: Evidence and Import. In A. H. Eden, J. H. Moor, J. H. Soraker, & E. Steinhart (eds.), *Singularity Hypothesis: A Scientific and Philosophical Assessment* (15-41). New York: Springer.

National Research Council. (2006). *A Matter of Size: Triennial Review of the National Nanotechnology Initiative*. Washington, DC: The National Academies Press. doi:https://doi.org/10.17226/11752.

Olah, C. (2016). *Bringing Precision to the AI Safety Discussion*. Google Research Blog. From https://research.googleblog.com/2016/06/bringing-precision-to-ai-safety.html

Simonite, T. (2016). Moore's Law is Dead. Now What? *MIT Technology Review*. From https://www.technologyreview.com/s/601441/moores-law-is-dead-now-what/

Simonite, T. (2017). Google's New Chip is a Stepping Stone to Quantum Computing Supremacy. *MIT Technology Review*. From https://www.technologyreview.com/s/604242/googles-new-chip-is-a-stepping-stone-to-quantum-computing-supremacy/

Shanahan, M. (2003). *The Frame Problem.* In the Macmillan Encyclopedia of Cognitive Science. L. Nadel (ed.), Macmillan, pp. 144–150.

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In N. Bostrom and M. Cirkovic (Eds.), *Global Catastrophic Risks* (308-345). New York, Oxford University Press.