Instructions for the Research Paper

COGS 4901: Honours Seminar in Cognitive Science

Jacob Beck

TOPIC

Many of the topics in cognitive science are controversial, with cognitive scientists themselves disagreeing about what the evidence shows. The point of the research paper is to force you to dig a bit deeper than you have to get a feel for the nature of such controversies and the methods that are used to address them by exploring a specific controversy and attempting to resolve it.

Since you've been studying cognitive science for several years, you've already come across a large number of controversies. For your paper, I strongly encourage you to choose a controversy that you've already encountered, but want to know more about. (Think back to your past classes. What gripped you?) Alternatively, you're also welcome to choose a topic from the following list.

- Is perception modular, or is it driven by expectations and background knowledge?
- Is there a cheater-detection module?
- Is there a domain-specific face-detection module?
- Is there a domain-specific theory-of-mind module?
- Do human beings really have picture-like representations in their heads?
- Is there a language of thought?
- Is there an innate universal grammar?
- Does the language one speaks influence how one perceives (or remembers, or reasons)?
- Can evolutionary psychology be fruitfully applied to study the human mind? Or is it doomed to trafficking in unverifiable just-so stories?
- Are there innate differences in mate preferences between the sexes that are driven by evolution?
- Is the mind massively modular?
- Can connectionist networks explain the systematicity of thought?
- Are probabilistic models of cognition superior to connectionist models?
- Can deep learning explain general intelligence?
- Are behaviorist accounts of learning (e.g. instrumental conditioning) true of animals such as rats and pigeons?
- Are there two separate visual systems? If so, how are they best characterized?
- How much plasticity attaches to the brain? Do one's genes determine which part of the brain is used for vision, audition, motor control, etc.? Or is neural organization a result of experience?
- What is (are) the neural correlate(s) of consciousness?
- Is altruism compatible with selfish genes?
- Are people altruistic (sometimes, often, never)?

- Is morality primarily the product of emotions or reason?
- Are emotions natural kinds?
- Do we have good introspective knowledge of our mental states?
- Has Molyneux's question been answered empirically?
- What does it mean to say that something is innate? Are objects (numbers, agents) represented innately?
- Is the frame problem an insurmountable obstacle for artificial intelligence?
- Is the mind confined to the brain, or does it extend out into the world?
- Is the attribution of mental states to others guided by a theory or simulation?
- Do infants attribute beliefs and other mental states to agents?
- Do chimpanzees attribute beliefs and other mental states to agents?
- Is there compelling evidence for ESP?
- Is there a human instinct for warmongering?
- Are integer concepts learned through bootstrapping?
- Is perception cognitively penetrable?
- Is attention necessary for consciousness?
- Is attention sufficient for consciousness?
- Does consciousness overflow access to consciousness?
- Does attention alter appearance?
- Why do children engage in pretend play?
- Are there different learning styles? For example, are some people "visual learners" and others "aural learners?"
- Does bilingualism bestow cognitive advantages (apart from the advantage of being able to speak more than one language)?
- Do nonhuman animals use cognitive maps?
- Is mathematical ability determined by the acuity of the approximate number system?
- What explains sound symbolism (e.g. the Bouba-Kiki effect)?
- What is the function of dreams?
- What are colors? Are they in the mind or in the world?
- What is X (where X = a specific phenomenon such as attention, consciousness, perception, intentionality, innateness, intelligence, learning, etc.)?

Many of the above topics are quite general. When you write your paper, you should narrow them down. For example, rather than considering whether language affects perception, you might write

a paper on whether the language one speaks affects the colors one perceives. All things equal, the more specific your thesis, the easier it will be to research and write effectively.

Whatever topic you choose, it should have an empirical literature associated with it and be controversial enough that there are researchers who have defended each side of the issue in print.

If you choose the last topic ("What is X?"), your paper will likely involve more philosophical analysis than if you choose many of the other topics. It will also be less straightforwardly addressed by the empirical literature. Nevertheless, you should still try to engage the empirical literature as you answer the question.

If you're feeling lost, or have questions, I can help you. Come talk to me. Choosing a good topic is half the battle.

RESEARCH

As you research your papers, you should observe the following guidelines:

- Read and cite at least one position paper that comes down on each side of the controversy you are considering. Your position papers should be general overviews or review pieces. (That means that they should summarize the existing literature, and not primarily be reporting new studies.) For example, if you write on whether language affects color perception, you should find at least two position papers: one by someone who says that it does, and one by someone who says that it doesn't. Review articles from major academic journals are best for this purpose.
- So far as possible, ground your discussion in empirical evidence (experiments, surveys, field studies, etc.). You should use logic and general arguments to structure your discussion, but shouldn't rely on them alone.
- It is often a good idea to start with a textbook to orient your search, but most of your material should derive from elsewhere.
- You want the latest evidence on your topic. The majority of your sources should thus be published within the last twenty years.
- Your sources should span at least two disciplines that are associated with cognitive science—e.g., philosophy and psychology; linguistics and neuroscience; etc.
- Quality is more important than quantity, so only use peer-reviewed sources. To be safe, stick to academic journals, books, and edited volumes written by people affiliated with universities. Along these lines:
    - Avoid articles from newspapers and non-academic magazines (*The New York Times*, *Newsweek*, *Time*, etc.).
    - Avoid citing web-based sources unless you're certain that they're peer-reviewed. You can make use of other resources from the web (e.g. *Wikipedia*), but only as a way of helping you find peer-reviewed sources.
- Consult your friendly York Library staff if you need help finding appropriate sources, or have questions about whether a given source is peer reviewed.

CITATIONS

Provide citations for all facts that aren't common knowledge (e.g., "People have brains").

Use the American Psychological Association style to format your citations and references. The basic idea is to enclose the author and date of a publication in parentheses when you refer to it. For example:

> The visual system can be explained at three different levels: computational, algorithmic, and implementational (Marr 1982).

Or:

> According to Marr (1982), the visual system can be explained at three different levels: computational, algorithmic, and implementational.

Or, if you quote from the publication instead of paraphrasing, you should include the page number:

> According to Marr (1982, p. 27), "it is the top level, the level of computational theory, which is critically important from an information-processing point of view."

Your reference section at the end of the paper should then list the details of the publications you cite in alphabetical order. (Only list references that you actually cite. So don't list a reference just because you consulted it.) For example, your entry for Marr would appear as follows:

> Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: W.H. Freeman and Company

A simplified explanation of APA style (sufficient for our purposes) is available here: http://library.williams.edu/citing/styles/apa.php. It will tell you how to format your reference section for journal articles, chapters in edited collections, etc.

IMPORTANT: To avoid plagiarism, be sure to:

- Cite any ideas that are not your own.
- Enclose any words in quotation marks that you borrow from someone else.
- Not hand in work for which you have already received credit in another class.

Anyone who is caught plagiarizing will fail the course and be reported to the University authorities. Information on York's policy regarding academic integrity can be found at: http://www.yorku.ca/secretariat/policies/.

FORMAT

Apart from using APA style for your citations and reference section, your papers should:

- Be double-spaced, in a normal sized font, with at least 1" margins on all sides.
- Begin with an abstract of 100-200 words that summarizes your topic and conclusions.
- Include few if any footnotes. Never use footnotes for citations.
- It is best to avoid quotations unless the exact wording is important—for example, because you're criticizing someone for choosing one word rather than another. If you do include a quotation, you should put it in the main flow of the text if it is less than 40 words, and set it off from the main text in an indented paragraph without quotation marks if it is less than 40 words.

- Divide your paper into sections for organization, and use two or three levels of headings to demarcate those sections as follows:

  A Capitalized Bold Heading

  A Capitalized Italicized Heading

  An italicized run-in heading. This sub-heading runs into the first sentence in the paragraph. You may or may not find it helpful to use this last type of heading.

ORGANIZATION

Start by describing the question you are considering and the various positions that you are evaluating. Be sure to explain them so that someone with no background in the subject can understand them. At the same time, make sure your characterization is fair, so that a proponent of each position would recognize and endorse it.

Now, for each position that you discuss, explain why someone might find it attractive. What are the major arguments for the position? What empirical evidence has been taken to support it? As you describe the evidence for each position, it is not enough to report the results (e.g., "the authors found that perception is influenced by language"). You need to explain the methods and logic of each study. Each study should take a lot of space to explain. It is almost always better to discuss fewer studies that are fully explained than to discuss many studies that are only partially explained. Although there are no hard and fast rules about how many studies you should discuss, an excellent paper of the expected length might discuss about three studies per position (or six total studies).

Once you have described the arguments and evidence, critically evaluate them. If the empirical evidence is based on studies, do those studies have any significant flaws? It is not enough here to point out minor weaknesses (e.g., "There were only n subjects") or non-specific possibilities ("The experimenters might have been biased"). You need to show how someone who rejects the position can explain the findings. Why might someone have predicted these findings even if they didn't accept the position?

Lastly, draw a conclusion about what the evidence and arguments on balance show. Be sure to explain why you take the evidence and arguments to lead to this conclusion. Here are your options:

- One position is right and the other is wrong.
- Both positions capture part of the truth. They explain different parts of the phenomenon, or aren't really in conflict with one another once properly stated.
- Both positions are wrong. We need to find a third way that no one has thought of yet. If this is what you decide, try to describe that third way as clearly as you can.
- We can't tell which position is right. The evidence isn't sufficient, or the positions aren't clear enough. If you choose this conclusion, try to think of an experiment that would settle the matter. (A feasible experiment is best, but even an unfeasible one would be helpful to show that you understand what's at issue.)

Approach your job as a scientist seeking the truth, not a lawyer looking to win an argument. Be fair to all sides, and see where the evidence leads you.

FURTHER WRITING ADVICE

Once you find articles and book on your topic (your "sources"), you should read them and take notes. The notes should summarize the main points in your own words. They shouldn't involve copying and pasting from your sources. Then when you start writing, you should put away your sources and rely solely on your notes. This will make sure that you put things in your own words and avoid plagiarizing your sources.

If you find it difficult to write clearly and grammatically, you may want to take a draft of your paper to the York Writing Centre: http://writing-centre.writ.laps.yorku.ca/services/one-to-one-writing-support/. In fact, this is really a great resource for anyone who is looking for extra feedback.

DEADLINES

9am September 30: a one to two-page, double-spaced preview of your paper that clearly states your topic and summarizes the main positions you will be considering. You should list at least five references, two of which should be your position papers, which should be clearly identified with an asterisk.

9am November 4: a three to five-page, double-spaced preview of your paper that clearly states your topic, summarize the main positions you will be considering, and the main arguments and pieces of evidence you will appeal to in support of your conclusion (with citations to the studies where the evidence is to be found). For each position, be sure to explain one or two studies in some detail, taking care to explain the logic and methods behind the study. Your preview should include at least two levels of headings, as described above in the section titled **FORMAT**, and should make the structure of your paper perspicuous. (You can think of it as an outline, but with complete sentences except for the headings.) Lastly, please include an annotated bibliography with at least eight references. The annotated bibliography can be single-spaced and does not count as part of your page limit. For information on how to prepare an annotated bibliography, see http://olinuris.library.cornell.edu/ref/research/skill28.htm.

9am January 13: a 4,000 to 6,000-word research paper citing at least ten sources.

9am March 2: a revision of your research paper (around 6,000 to 8,000 words) citing at least twelve sources.

A Morality Module for Machines

#Cognitive Science 4901

ABSTRACT

This paper will explore whether it is practically possible to realize artificial morality. First, I will articulate the prevailing positions on these difficulties and subsequently express why one of the established proposals, the evolutionary paradigm, ought to be considered the optimal substratum upon which to build a moral architecture. This is best conceptualized through the development of an independent moral algorithm that may be implemented in a diverse range of automata, effectively converting these machines into artificial moral agents. New innovations in deep learning may facilitate such a project in ways that have not previously been possible so as to belay many of the practical problems raised as objections to artificial, moral decision-making. I will outline a rudimentary springboard for the development of such an algorithm and subsequently discuss how, generated in the ways I will describe, this algorithm may tackle some of the preeminent roadblocks pertinent to the development of moral algorithms.

INTRODUCTION

The question of morality is unique in comparison to discussions of other forms of skill automation, which tend to focus on *when*, rather than *if*, the technology reaches its viable form. Artificial morality, on the other hand, is hampered by dissention about whether it is even *possible* to digitize this 'uniquely human' ability. Speculation of this kind is a distraction from the necessary conversations that must be had as we stand on the precipice before artificial intelligence that may soon surpass our own. Machines already exist that guide ethical decision-making, and many more maintain tremendous impacts on moral outcomes. *Any* increase in the ethical competence of such machines that results in the overall reduction of harm and increase in wellbeing is, almost by definition, a favourable outcome. As such, this paper will assume that an algorithm facilitating moral decision-making is, to some degree, advantageous. Further, since the machinery for morality already exists (even if only in humans), there must be a mechanism by which to replicate it. Thus, this paper also assumes that the synthesis of

a moral system is, at the very least, theoretically possible. With these considerations undergirding the endeavour, I shall focus this work on the barriers preventing the functional development of a working moral algorithm and present some potential solutions for consideration.

Cognitive ethicist Colin Allen has highlighted many of the main practical impediments currently hampering the development of artificial moral (ro)bots. He raises three major roadblocks: discordance in views of morality, the difficulty of incentivization to moral action, and the role and implementation of biologically realized mechanisms like emotions and empathy (Allen, Varner, & Zinser, 2000; Wallach & Allen, 2009). I will term these the problem of subjective morality, the problem of incentivization and the problem of biological parallels, respectively. Even if the theoretical possibility of developing artificial morality exists, these barriers will first need to be surmounted in order to reify it.

The Problem of Subjective Morality: Hume's observation, that one cannot get an *ought* from an *is*, is a major difficulty plaguing the development of an objective moral framework. Thus far, it appears infeasible to *empirically* derive moral principles; developers of such systems must be content with subjective moral frameworks based on inferences about morality. This, however, forces the cherry picking of foundational premises upon which to base a moral system, and there is significant disagreement about what these should look like. Artificial intelligence developers that wish to generate moral machines, therefore, must decide on a foundational set of premises about morality before they begin training. Allen describes the three approaches developers typically use in order to frame these foundational premises: virtue approaches, associative learning approaches, and evolutionary/sociobiological approaches (Allen, Varner, & Zinser, 2000).

Virtue approaches use a healthy mix of top-down and bottom-up methods to program computers to operate under deontological principles in much the same way as humans. The idea behind virtue approaches is that, if the computer is coded with moral rules that set a foundation for 'morally

good' decisions, then these rules will effect morally good outcomes. Allen notes that this approach suffers from many of the same difficulties of any deontological system, like dealing with competing virtues and establishing what 'morally good' (or any other principle, such as honesty or integrity) should mean (Allen, Varner, & Zinser, 2000). There is an added difficulty when using top-down programming methods in this way. Allen notes that the list of virtues to be programmed into a machine is likely unable to cover all of the scenarios that a machine may find itself in and, for this reason, approaches more reliant on bottom-up learning may be more appropriate (Allen, Varner, & Zinser, 2000).

Associative learning approaches are those that rely on supervised learning. Moral quandaries can be presented to the machine, which subsequently decides whether the sample is moral or not. The problem with this, as identified by Allen, is that moral responses do not lend well to binary answers. The learning of these principles is often couched in the rationale behind them that can both be identified *and* articulated (Allen, Varner, & Zinser, 2000). If, for example, a person is asked, 'is it moral to steal a loaf of bread to feed one's starving family?', the response of 'yes' or 'no' is insufficient. In fact, there would be no way to demonstrate that the responses are not a result of mere chance. It is in the justifications provided, the 'yes, *because*…' or 'no, *because…*', that demonstrate moral understanding. People may agree with either 'yes' or 'no' depending on the rationale following the '*because'*.

Associative learning approaches are a problem for coders because they require the integration of so many systems. The amount of background information required for justification is extremely high. For example, one will need to know the relative value of bread, as the response to the question may change if the question were rephrased as 'is it moral to steal a crate of sirloin steaks to feed one's starving family?'. As the system requires more and more information, the domain in which it operates becomes increasingly broad and thus, fewer resources are dedicated to answering specific questions. Therefore, these machines are unlikely to be sufficiently complex so as to pass any kind of moral Turing test.

Evolutionary/Sociobiological Approaches aim to simulate evolutionary trajectories in order to establish morality from the ground up. This will be explained in greater detail in subsequent sections, but it is worth noting some of the associated difficulties. As Allen points out, they will suffer from similar problems encountered in the associative learning approaches (Allen, Varner, & Zinser, 2000). The difficulty is less pronounced, however, due to the simplicity with which they begin (they must effectively learn *everything*). These models can learn the necessary background knowledge for a small number of scenarios, but this presents the problem of scaling them to become broader in their applicability. My hope is to explain why this problem may be alleviated by training the moral algorithm in a domain specific way and subsequently instantiating it *into* other machines such that it can operate autonomously while incorporating the training of its host machine.

The Problem of Incentivization: This ties into one of the problems that was outlined with respect to associated learning approaches: the binary nature of training in pass/fail conditions. Moral questions do not fall into categories with delineable borders but rather, fall upon a spectrum of more or less moral outputs. As Allen argues, humans have a categorization method that places actions on a spectrum: motivation to avoid punishment or seek reward correspond to being more or less desirable outcomes. Approval and disapproval are key abstractions based on this motivational system that facilitate moral learning. Humans are, by nature, social creatures and as such, degrees of approval and disapproval are often sufficiently salient to guide moral learning and override contravening goal-states. Machines have no access to social emotions, and thus, moral calculations will likely never reach the same priority levels as the goal-states for which the machine was designed.

The Problem of Biological Parallels: Emotions are universal communication systems, transcending language, culture and ideology. The presence of positively or negatively valanced emotions within others can also be experienced within oneself, bridging the gap between the actions one takes and the effects that it has on others. As mentioned above, one element of this might be approval or disapproval,

but it is not limited to these states. The capacity for empathy allows for more nuanced updates to weighted connections to facilitate social learning. Since morality is a social paradigm, it is necessary to have a universal reinforcement and feedback mechanism for learning the rules within this context. We more readily empathize with other humans than we do with rats, and we empathize with rats more readily than we do with staplers. Our moral rules tend to reflect the spectrum of this universal emotional language. If machines have no functional equivalent to emotions, it is unclear how they would both act, and be acted upon, within the confines of a human moral system that uses emotional stimuli as the feedback for learning. The design of computers renders it extremely unlikely that any biological parallel for emotions is possible, with the sole exception replacing of every organic cell in a human with a mechanized equivalent. While this may be at some point possible, it is highly unlikely that this will be the form of the most common artificial intelligence systems with whom we regularly interact.

ESTABLISHING AN ARTIFICIAL MORAL FRAMEWORK

Top-Down vs. Bottom-Up Moral Frameworks: In his book, Allen raises the question as to whether top-down or bottom-up approaches are the appropriate choice of framework (Wallach & Allen, 2009). Broadly speaking, top-down approaches are those that migrate from the larger, broader, and more general towards the smaller, narrower and more specific. Bottom-up approaches, on the other hand, tend to move in the reverse direction, starting from smaller, narrower and more specific and moving towards the increasingly large, broad and more general. With respect to the brain, bottom-up generally refers to the workings of a smaller biological mechanism that goes on to influence larger brain systems or networks. Neuroleptics, for example, contain molecules like haloperidol which exert their effects on both dopamine (typically D2) and serotonin (5-HT2) receptors in the brain. By antagonizing these receptors, most of the neurons with these receptors will fire differently, which then propagates forward, causing changes in large scale dopaminergic pathways, which in turn, causes changes in all of the neural

systems that have a causal relationship with dopaminergic projections. The change in these large-scale systems is the reason the drug then has an effect on behaviour and cognition as a whole.

Top-down mechanisms, on the other hand, capitalize on these extensively interdependent systems which feed-back to make slight alterations to increasingly smaller systems until even molecular properties can be affected. The brain has several identifiable large-scale connection networks that, while related, fulfill different roles. The default mode network is one such example and among its many functions is the facilitation of the interconnectivity between regions involved in the understanding of others (or theory of mind) (Li, Mai, & Liu, 2014). When the many brain regions associated with this network converge, and perceive a potential social threat, information is passed to the amygdalae, which in turn converge their signals on the hypothalamus. The hypothalamus then instructs a smaller system still, the pituitary, to create the adrenocorticotropic hormone molecule. In this way, even our most complex cognitive constructions can lead to changes in molecular signaling.

The 'cognitive economy' describes the neurocognitive concessions that must be made between accuracy and storage, and speed and efficiency of resource use (Rosch, 1999); top-down mechanisms of cognition are extremely valuable for speed and efficiency, but remarkably poor in accuracy and information registration. Hallucinations and delusions, erroneous beliefs and cognitive distortions are all common examples of top-down processing on insufficient or misrepresented data (Aleman, Böcker, Hijman, Haan, & Kahn, 2003; Campbell, 2001; Kahneman, 2011). Advances in brain imaging technology and studies on split-brain patients have yielded evidence as to why this may be; top-down processes appear to use heuristics to fill in knowledge gaps with inferences that may or may not be correct. Split-brain patients are those who, often due to surgery or trauma, suffer from interhemispheric disconnectivity. Some of the most exciting experiments conducted by on these patients have involved the projection of commands to the right hemisphere and the subsequent justifications confabulated by the left when asked why they are performing these actions. When, for example, patients had the

commands ‘laugh’, ‘rub’ and ‘walk’ projected onto only the left visual field, so that it would be interpreted only by the right hemisphere, they performed each of the actions without hesitation. When asked why they were doing so, the responses were ‘Haha…you guys are just too much,’ ‘An itch’ and ‘Oh I need to get a drink,’ respectively (Hirstein, 2005, p. 154). The causal reasons for performing the actions were completely divorced from their cognitively constructed reasons. This is lending increasing credibility to the social intuitionist hypothesis that top-down reasoning is simply confabulation to justify decisions rendered before reaching consciousness (Haidt, 1995). In this view, all rationalization is essentially confabulation, and rationalization is not limited to slit-brain patients. Indeed, tens of thousands of studies demonstrate the effects of unconscious stimuli on moral decision-making. These range from evidence of interoceptive signals affecting self-control (Gailliot, et al., 2007) to vacillating political perspectives in the presence of aversive stimuli (Adams, Stewart, & Blanchar, 2014). Yet, in spite of empirical evidence to the contrary, people still readily stand by what appear to be nothing more than confabulated justifications. Thus, top-down attempts at accessing principles of morality are likely misguided for an accurate picture of decision-making *in general,* and almost certainly so for each individual decision within the ethical realm.

Until recently, top-down programming methods were the only viable methods available. But new deep learning techniques have revolutionized the way in which machines can be trained, allowing bottom-up styles of data collection to yield novel outcomes in the absence of top-down systems. Moreover, modern machines have the distinct advantage of unparalleled processing and storage capacities when compared to human brains. Couple this with a lack of colossal metabolic costs and a prime candidate for bottom-up learning that can achieve accuracy levels far exceeding human competence emerges. The main criticism long levelled at bottom-up systems was the inability to use abstract constructions, but recent advancements in deep learning technologies have put such objections

to rest, rendering doubts about the superiority of bottom-up computational systems untenable. The momentous achievement of AlphaGo Zero highlights why this is so.

The game of Go requires a remarkable degree of abstract, strategic forethought, more so even than chess. In 1997, the world was awestruck when Deep Blue defeated Garry Kasparov in a six-game tournament to become the number one chess player in the world. Chess, unlike Go, has very strict rules with respect to how each piece can move in addition to a more constrained victory condition. The number of possible moves is readily calculable, albeit extremely large, but with sufficient processing power, Deep Blue was able to use a "brute-force" method: it speedily processed every possible move each turn and systematically eliminated those that would lead to less statistically viable actions than the one calculated before it. Though an impressive display of progress made in computational speed and processing power, Deep Blue was not considered intelligent; it was not able to harness the power of learned abstractions. Brute-force methods have been unsatisfactory for Go because the sheer number of available decision options and victory configurations is too great. AlphaGo instead trained with machine learning, allowing it to home in on the most effective human strategies. In 2016, it managed to beat one of the world's best professional players in 4 out of 5 games. Though impressive, it was not long before AlphaGo was dwarfed by its successor, AlphaGo Zero. In just 3 days, AlphaGo Zero managed to defeat its predecessor 100 games to 0. The major change in its learning software was conversion into a *purely* bottom-up model, that was "no longer constrained by the limits of human knowledge," (Knapton, 2017). It trained exclusively on games against itself, building a framework for 'better' and 'worse' moves through trial and error at remarkable speed. This style of machine learning works so well because it is both domain specific and the success/failure conditions are clear. Newer versions of AlphaGo have mastered other games, like chess, and the world's grandmasters have taken notice of the novel strategies it executes. Kasparov himself discussed how the new, highly aggressive styles of play are surprising to him and that studying computer strategy is making him play better chess (Kasparov, 2019).

It is also worth mentioning, to address the problem posed by Allen, that even if one were able to put aside considerations of efficacy and independence for a system of morality, there is still the problem of subjective morality to contend with. Any program developed by utilitarians will be anathema to deontologists. Religious denominations will require their own software. American cars prioritizing driver safety will face scorn in Japan, and Japanese cars prioritizing pedestrians will collect dust in America. Morality is culturally contextual, but it is also individual. The only aspect of morality common to all humans is the hardware upon which it took root. This hardware was formed by a bottom-up process over evolutionary time, which subsequently provided the flexibility for learning different, contextually relevant software updates.

The idea that humans have only moral hardware may initially seem counterintuitive because there are, at the very least, *some* moral principles we view as universal. After all, regardless of one's moral inclinations, all humans can agree that torturing children is wrong, right? What then do we make of the dozens of cultures who engaged in such practices? Moreover, what reason do we have to believe that if we placed an infant born today in such a society (assuming they made it to adulthood) they would not share the conviction that this practice was a morally acceptable? Moral convictions are based around beliefs, and beliefs need not be correct in order to make moral judgements. Suppose a belief system dictates that:

a. The tears of sacrificial children are necessary to appease the god Tlaloc
b. If Tlaloc is not appeased, then no rain will come to the crops
c. If no rain comes to the crops, every person will slowly starve to death

If all of these premises were *true*, then the practice of ripping the fingernails off of a child so as to make them cry as many tears as possible before being sacrificed is morally defensible. If you heard that a person's profession included ripping the fingernails off of children in order to make them cry, you may

speculate such an individual to be *amoral*; the person is in possession of *no* moral principles. If presented with the belief systems that they hold, you are likely instead to consider them to be *immoral*; they are in possession of *incorrect* moral principles. The bottom-up generation of principles based on premises (beliefs) is likely one of the necessary criteria for recognizing a moral system, even if the beliefs themselves are not universal.

Evolution's Solution: In order to address the issues of relevant background knowledge inherent in the problem of subjectivity, there must be a kind of selection mechanism that promotes the encoding of some information at the cost of others. Hybrid neural networks tend to use top-down systems to fulfill this role. The bottom-up elements flow based on the established framework for decision-making coded by the creator of the network. As argued above, however, the less top-down intervention necessary, the more independently effective the system becomes. Thus, it seems reasonable that a single value minmax approach is appropriate. AlphaGo Zero, for example, used this approach with criteria of efficaciousness established by wins and losses. In fact, one of nature's best selection mechanisms, natural selection, follows a similar minmax evaluation of efficaciousness, fitness, established by reproductive success or failure. The allure of natural selection-style mechanisms is that they are based purely on what *actually* works and are therefore resistant to the propensity for error inevitable in any top-down system. To quote Orgel's second law, "evolution is cleverer than you are".

Using natural selection-style mechanisms to train algorithms seems fitting for modelling human traits. After all, we evolved through this system, so it makes sense to use such criteria in an attempt to code human behaviours like morality. This is precisely the kind of work that began in the mid-1900's with genetic algorithms (GAs). GAs are not simply limited to genetics as their name might suggest, rather, they are algorithms that aim to capitalize on the principles of selection to optimize many kinds of systems from modelling ecological resource flow to developing better bidding strategies in economic

markets (Mitchell & Forrest, 1994). The most simplistic version of a GA looks something like the following (Mitchell & Forrest, 1994):

1. There is a generation of randomly selected components.
2. Each component's fitness score is evaluated.
3. Genetic modifiers apply crossover effects and various forms of mutation resulting in a new population.
4. The process is repeated from step 2.

The benefit is that it can realize rapid improvements with a number of variables based solely on efficacy. Allen discusses the beneficial outcomes of GAs but notes that attempts to generate ethical frameworks have been unsuccessful (Wallach & Allen, 2009). Complexity required is a factor, but as computing power increases, so too does this problem decline. The other major difficulty is one technology alone cannot overcome: a lack of control over the selection of variables and whether they are ethically appropriate.

Suppose, for example, an algorithm responsible for allocating organ transplants determines that statistically, one ethnicity has a mortality rate 6.7% higher than others, and as a result, denies transplants based on this variable. If a human candidly made such decisions, he would likely be considered unethical and perhaps even racist. Bias poses a problem to ethics, whether human or digital, but it seems that evolution has equipped humans with an override to purely Bayesian threat detection: the prospect of reciprocal cooperation. Many studies have now demonstrated what military officials and sports coaches have long anecdotally reported: the induction of cooperation overrides existing prejudices and draws new ingroup lines that engender favourable attitudes towards potential partners (Gaertner & Dovidio, 2000; Tajfel & Turner, 2004). A moral system that is based cooperation as its only foundational principle necessarily construes any potentially cooperative participant as a member of its

'in-group'. Equipping bottom-up genetic algorithms witch such a propensity may solve the quandary posed by bias in genetic algorithms for morality.

Reciprocal Altruism Heuristic: Moral systems are the top-down products of bottom-up mechanisms for cooperative behaviour; without any element of cooperative or reciprocal arrangement, rules become morally vacuous. Therefore, in an effort to uncover the evolutionary underpinnings of morality, in addition to tempering the effects of bias, it makes sense to begin with the evolution of cooperation. In 1980, Axelrod solicited, from interdisciplinary game theorist researchers around the world, strategies to be pitted against each other in the Prisoner's Dilemma scenario. Two versions of this tournament were held. The first was limited to 200 rounds, but the second, in an effort to better match the uncertainties in organic cooperative behaviour, left the number of rounds unknown (Axelrod & Hamilton, 1981). In both versions of the tournament, a strategy submitted by Anatol Rapoport emerged victorious: TIT FOR TAT (Axelrod & Hamilton, 1981). The strategy was fairly simple: it opened with a cooperative move, and directly matched whichever move its opponent made in the subsequent round.

Axelrod published his findings in a landmark paper, *The Evolution of Cooperation,* with evolutionary biologist William D. Hamilton, which discusses the reasons TIT FOR TAT was so successful. The calculations suggested this as the likely solution to the thorny problem of cooperation in a "selfish" system of Darwinian selection (Axelrod & Hamilton, 1981). Operating under such a strategy would provide organisms a disproportionate advantage in meeting survival and reproductive needs, provided that the organisms had a high likelihood of repeated encounters and some mechanism by which to individualize and punish defectors should they be reencountered. The better organisms fit such criteria, the more cooperative stable strategies become apparent (Axelrod & Hamilton, 1981). Subsequently, as societies become more complex, cooperation and defection behaviour become codified in both legal and moral systems (Curry, 2016).

It is important to note that, although TIT FOR TAT emerged repeatedly victorious, it is not necessarily the optimal 'nice' strategy. 'Nice' strategies are those that begin with cooperative behaviour, rather than defection. While nice strategies emerge mathematically superior, 'nicer' strategies can be more viable depending on strategies used by other members of the population. Indeed, Axelrod used a 'forgiving TIT FOR TAT' strategy, TIT FOR TAT that forgives one defection, to challenge the contenders of the first tournament. He discovered that, had this strategy been submitted, it would have won. This is because an evolutionary strategy's success will depend on the other strategies being used in its ecosystem; in the first tournament, there were fewer defection strategies and thus, the nicer strategy was superior (Dawkins, 2016). The corollary of this is that the more cooperative an overall population, the less punishment is required and thus, the more evolutionarily beneficial 'nicer' strategies become.

Reciprocity theory explains the development of cooperation and sociality of species. Cooperative strategies, in turn, cohere into moral systems. Thus, for the purposes of generating a bottom-up, machine learning algorithm for morality, it seems only logical to begin with reciprocity theory as the framework. Simple teaching scenarios can be presented to the algorithm through a game theoretical approach, and based on Axelrod's (and subsequent) research, we will find cooperative strategies to prevail throughout the training phase. This allows the algorithm to learn cooperative behaviour from the ground up with increasing scenario complexity. Eventually, much like the occurrence in humans, the machine learning process will begin to make inferences across scenarios that will act as overarching principles to guide decision-making in novel cases once the training phase is complete. Further, because it is mathematically bound to be cooperative, and it is training against itself, the emergence of a 'nicer' strategy akin to the forgiving TIT FOR TAT is likely to emerge rendering the algorithm, by default, extremely prosocial.

Reciprocity may help to mitigate concerns about conflicts between moral codes and the potential for biases, but it begs an important question. How does one bridge the gap between solving a

cooperative problem, like the Prisoner's Dilemma, and solving a philosophical moral problem, like the trolley problem? Take, for example, the scenario in which a trolley hurtles down a forked track with 5 people helplessly anchored in its path. With the flip of a switch, the trolley changes course, now barreling down on only 1 unfortunate captive. Cooperative game theory concerns some of the following considerations:

1. How great is the benefit from cooperation versus defection?
2. How likely is the individual to defect?
3. How likely, if I defect, am I to receive punishment?

To answer the first question, securing 5 potential sources of cooperation is superior to 1. If, however, the 1 person was a friend, you could reasonably expect more cooperative interactions with them than 5 strangers. The second question, for this dilemma, is less relevant; defection likelihood is virtually nil. Consider, however, the 'footpath' variation involving the heavy man on a bridge. If your defection fails to kill the man, he will almost certainly defect, possibly by trying to throw you onto the track. Finally, in the third question, though punishment is improbable, since those defected upon will likely be dead, there is still a greater chance of experiencing punishment from the group of 5; should the trolley stop after defection was made clear (refusing to pull the lever), there are 5 potential sources of punishment rather than 1. Moreover, this can be a measure of success rate. In the case of the 'footpath' version, it is the estimation of how likely you will be to succeed in killing him. If you fail, severe, presumably even fatal punishment becomes likely. These kinds of calculations are equivalent to real-world scenarios for self-driving cars. In deciding between colliding with a child or an elderly individual, cooperation calculations suggest a greater statistical likelihood of cooperative interactions involving the child, as they will live longer. Organisms that secured the most cooperative partners had the greatest levels of fitness. Iterated simulations should follow a similar trajectory. Of course, none of these calculations in humans occur consciously. Rather, they are molecular changes that breed unconscious emotional shifts. Post hoc

rationalization occurs that may bear very little resemblance the impulses that spawned it. This fact is corroborated by the discrepancy between what people claim they would do in trolley problems, and what actions are actually taken in VR trolley simulations (Francis, et al., 2017).

The development of a bottom-up moral framework helps deal with the problem of subjective morality in developing an independent moral algorithm. If the machine is generating the ability to be moral from the ground up, it is not necessary to worry about defining and codifying moral rules to be obeyed. The machine will recapitulate the evolutionary moral trajectory from simple organisms to human infants, and subsequent moral training from human infants to ethical adult agents. While this process took billions of years in humans, in machines, we will be able to watch this occur before our eyes. In this approach to machine learning, the algorithm will be training against itself and will be exposed to millions of game iterations in just days. While training itself is relatively quick, the design features necessary to generate a functional algorithm complicates matters. The purpose of the algorithm will be such that it can be instantiated into other machines with their own goal-states but remain functionally independent. This is important so that the algorithm may still maintain domain specificity. Thus, there will then need to be a way for moral decision-making to be sufficiently salient so as to compete with (and hopefully, override) these goal-states in much the same way that social pressures shape a human conscience. In other words, it is necessary to deal with the problem of incentivization.

II. DEVELOPING A MORAL ALGORITHM WITH MACHINE LEARNING TECHNIQUES

Digitizing the Conscience: Machine learning continues to improve in its efficacy as collaboration between the neuro and computer sciences yield more fruit. One of the most exciting and effective solutions emergent from such partnerships is reinforcement learning, a method of machine learning that directly mimics the way human brains learn via the dopaminergic system (Sutton & Barto, 2018). The

dopaminergic system contributes to learning through evaluation of potential reward versus actual reward, and then reconfigures synaptic weights accordingly so as to find a more accurate balance between the two variables (Glimchar, 2011). The dopaminergic system is more than just a reconciliatory mechanism for reward prediction, however; it is also generative. 'Actual reward', as measured by the network, *actually rewards* the brain allowing it to be a semi-closed system for reward-based learning. The reinforcement effect is bidirectional and multiplicative in that both positively and negatively valanced neurochemical triggers can be generated. This is based not only on actual reward but also the discrepancy between actual reward and predicted reward, with increased feedback-related negative affect as actual reward falls below the predicted (Bismark, Hajcak, Whitworth, & Allen, 2012; Schulz, Dayan, & Montague, 1997). Put simply, reinforcement learning both biological and digital do the following:

1. Assign positive value to high reward and negative value to low reward
2. Generate reward by reducing discrepancy between current state and desired state
3. Generate reward-value based predictions for available decision options
4. Calculate the difference between predicted reward and actual reward
5. Solve for the appropriate reward value and reconfigure connective weighting accordingly

The difficulty with reinforcement learning is that itcaptures the current landscape of neural connectivity, whether biological or digital and in outcomes of reward (whether positively or negatively sourced), strengthens (or weakens) connectivity to *all* connections that are active at the time of reinforcement (Sutton & Barto, 2018). While this helps mitigate some of the issues involving insufficient background knowledge, it can lead to misattribution errors and faulty premises upon which further calculations are made, which in turn, lead to flawed conclusions. Humans suffer from this problem with reinforcement learning as well. For example, suppose a person is extremely hungry but cannot leave the office. The only food vendor that is open in the building sells gluten-free products. This individual

therefore orders a gluten-free meal and after having eaten the meal, finds themselves feeling extremely refreshed and invigorated. They make the attribution between gluten-free food and feeling good, rather than any of the myriad of other variables that may have caused or contributed to that state. This weighted connection between the two variables (gluten-free food and feeling good) becomes the strongest variable in its respective networks and thus, requires a large number of contravening trials to disconfirm it.

Machine algorithms hold an advantage over humans in avoiding misattributions for two major reasons. First, computers make calculations significantly faster than humans and thus, are simply able to generate more trial scenarios in shorter spans of time. Therefore, provided the attribution is indeed false, the computer will be able to process more trials (and errors) to sufficiently weaken this connection. This is facilitated in part by the domain specificity of computerized algorithms. For example, while the human brain has more processing power than a calculator, the calculator will process 6(322 x 36) faster than a brain. This is because much of the brain's available processing power (external to survival and metabolic processes) is dedicated to inhibiting the *irrelevant* details. Calculators do not have to spend precious processing power suppressing disappointment over a breakup or questions about what to cook for dinner. This leads to the second reason for superior attribution power in machines: the problem of episodic memory. Episodic memory is what humans use to vividly remember events that occurred in their lives, drawing heavily on sensory imagery. The difficulty with episodic memory is that it is treated by the brain as a novel 'trial' and is reconsolidated. The same connections are activated, and the neurons within the network undergo minor chemical changes with each recall, which in turn causes and adjustment in the weights of these connections (Suzuki, et al., 2004). Even without a new trial and outcome taking place, networks weights are updated. Therefore, our example individual need only *remember* the instance of eating gluten free food, perhaps overemphasizing the negative state before the food and positive state after the food in order to create a large reconfiguration

of attribution. He is now *more* convinced gluten free food made a radical transformation in his affective state, despite having no new information. Computers do not use the mechanisms of episodic memory and because of this, their weights will not be updated by recapitulations of the same data.

The reinforcement outcomes in the machine must serve a greater purpose than merely expediting training times and minimizing error rates. If the reinforcement mechanism does not impel the machine to actually make ethical decisions, it fails to serve as a moral framework because competing goal-states will inevitably take precedence in machines originally developed to accomplish certain aims. The problem of incentivization must be resolved by making cooperation highly rewarding. Research done on the neuroscience of morality has demonstrated that cooperative behaviour triggers dopaminergic reward; more specifically, cooperative behaviour in game theoretical parameters of the Prisoner's Dilemma (Rilling, et al., 2002). If our dopaminergic reward system has been forged by evolution to facilitate goal-states of survival and genetic reproduction, why would choosing cooperation over defection trigger such a surge in dopaminergic activity? This harks back to Axelrod and Hamilton's findings: even goal-states that are 'selfish' in nature are more readily obtained through cooperative means. While the output of the machine's training period is simply a raw reward signal, the data sets for training implicitly teach the machine that cooperation is rewarding *because* it facilitates maximizing its own goal-states. Thus, when implemented into a machine designed for potentially conflicting goal-states, the algorithm offers the learned principle that cooperation will better facilitate individual goal-state satisfaction overall. Moreover, based on the evidence provided by Axelrod and Hamilton that this is indeed the case, the choice of cooperation in an effort to secure selfish goal-states will continue to be reinforced, provided that the machine operates within environments of cooperative agents. As a result, the module will be able to act as a moral guide based on learned principles much like a human conscience.

Reinforcement learning may provide great promise for bottom-up programming, but Allen raises a reasonable objection to bottom-up models specifically *on the grounds of* machine learning: computers can "learn the wrong thing,"(2009, p. 110). After all, humans are the product of bottom-up, reinforcement learning and we learn the wrong things almost every day. Unfortunately, it is a reality that errors will be learned even in the best of bottom-up systems: after all, even natural selection has left us with our retinas on backwards and a choking hazard in our necks. Moreover, even bottom-up AI systems have been shown to make errors and misattributions through what have come to be known as 'adversarial attacks'. These are instances in which people capitalize on the knowledge of how an algorithm works in order to lead it to inappropriate outputs. Moreover, this kind of attack is not limited to machines. Human can be similarly manipulated into solidifying inappropriate beliefs or effecting inappropriate behaviours based on knowledge of their thinking processes. Adversarial attacks come in two varieties: 'evasion attacks' and 'poisoning attacks'. Evasion attacks are those in which a person knows an algorithm's evaluation criteria and thus, uses this information to circumvent it. This is relatively simple to accomplish in fixed algorithm but much more difficult to accomplish in a machine learning system. This is because the solution to this problem is more data, which a fixed algorithm cannot make use of. A machine learning algorithm, on the other hand, is continually collecting more data and thus, the evasion issue can be solved through providing another round of accelerated training.

Poisoning attacks appear to be more what Allen has in mind when discussing his concern with machines learning the wrong thing. Poisoning attacks are those in which an individual intentionally undermines the training process by flooding the trials with samples carefully tailored to skew the weights in a certain direction. In fact, though the system is 'bottom-up', there is an element of top-down bias towards a specific outcome being injected into the training process. Therefore, rather than the

algorithm making false attributions, the misattributions can be seen as a result of *human error*.[1] While theoretically, the objection with respect to flawed learning makes good sense, the real-world examples of bottom-up, reinforcement learning machines appear to suggest this issue to be negligible. AlphaGo Zero is not the only machine that experienced significant improvements to effectiveness when unshackled from human error. Given an *appropriate* reinforcement set, machines will become far more accurate by sifting through raw data than it will by trying to imitate humans. That is not to say machines will never make errors, but rather, it is to say that as long as they make fewer errors than we do, this ought not to be a concern. After all, humans are error prone and yet we regularly trust their judgement. This ought to be extended to machines as well, especially if they prove themselves more competent than us.

Emotionality Sans Endocrinology: There is a school of thought in philosophy that supposes emotions to be a crippling force in making moral decisions and Spock-like creatures as paragons of moral arbitration. It is typically populated by utilitarians, and indeed, there is evidence to suggest these thinkers are less emotionally driven in moral decision-making. Harvard scholar Joshua Greene, one of the world's leading experts on the neuroscience of morality, has found a link between philosophical morality perspectives and activation of anticorrelated regions involving reasoning and emotional response. People framing the trolley problem as utilitarians show greater activation of the "rational, cognitive" dorsolateral prefrontal cortex (dlPFC) (Greene, Somerville, Nystrom, Darley, & Cohen, 2001). As the difficulty of the moral dilemmas increases, so too does the activation of emotional centers at the cost of dlPFC activation; as such utilitarian decision-making drops off (Greene, Nystrom, Engell, Darley, & Cohen, 2004). Emotional decision-making is at some point, inevitable and necessary for morality to prevent moral perversions. No matter how one may anachronistically rationalize the beneficial scientific outcomes generated by the

[1] 'Human Error' in this case is not to be construed as 'unintentional' but rather, an error in the cause and effect determination mechanism caused by a human.

mountain of cadavers proffered by Josef Mengele, no moral agent recognizable to humankind would suggest recreating this scenario. In considering such situations, it appears more plausible that emotions act as the bottom-up, driving force for principle-generation. When no current principle reconciles our "gut-feelings", we rely more on bottom-up, emotional decision-making that involves empathy.

There is a question worth considering at this point: what benefit to information processing are emotions providing? In his work with lesioned patients, Antonio Damasio discovered that emotional affect is necessary for decision-making. Patients who suffer trauma to emotional centers like the ventromedial prefrontal cortex (vmPFC) or amygdala, for example, fail to make appropriate decisions (or any decisions at all) (Damasio, 2004). Indeed, virtually all disorders of decision-making, such as abulia, present with flattened affect and apathy. This makes perfect sense, for without any mechanism by which to make some data points more salient than others, one could continue processing information *ad infinitum* without ever determining if any of it is relevant to the task at hand. Emotions facilitate the generation of abstractions, collections of salient and related variables, that can act as heuristics, making decision-making possible. Moreover, the *expression* of emotions allows salience to be *conveyed* in addition to being processed. Evolution has provided emotional or endocrinological mechanisms to take expansive sets of inputs, filter the irrelevant information and transmit this information quickly throughout the nervous system so that it can be readily recapitulated and utilized; this is, in effect, the exact role of neurotransmitters.

New deep learning techniques called 'autoencoders' have been developed to accomplish specifically such a role. The basic idea of a standard autoencoder is that it takes a set of input nodes and tries to convert the configuration into a sufficiently low-resolution version of the information. It seeks to filter out information from the inputs through an 'encoder' to generate increasingly low-resolution versions so that this information might pass through a bottleneck consisting of a significantly reduced number of nodes than the original input set. From the compression state of the bottleneck, the

information passes through a ‘decoder’ which aims to recreate the information lost in the encoder phase and submit it to an output. The better the output expresses the input, the more effective the autoencoder. The benefit of this process is that it allows for the generation of compressed versions of a data set that capture only the most salient information necessary for use.

For example, the glucocorticoid hormones involved in the stress response, offer a relatively simple illustration of the analogous role in biological organisms. Suppose a person is walking through the forest and hears a loud growling coming from directly behind them. Virtually all of the input information being received in that moment become impediments to the goal of decision-making. Were this individual to spend precious seconds noticing the colour contrast in the leaves or the coolness of the breeze, they likely would not have lived to propagate as many offspring. Evolution has selected for the ability to filter out stimuli from a large set of inputs and compress the data to the most relevant, salient features. A low-resolution version of the inputs, cortisol, can then move quickly and effectively throughout the body to convey the pertinent message of ‘danger’ and upon decoding, contextual information provides cues as to how movement ought to occur and what steps are to be taken. In reality, many molecules and their location of secretion and action are what generate the sensation of emotion and encompass all of the analogous components of an autoencoder, complete with input system (activated neural network), encoders (pre-synaptic neuron locations), bottleneck (molecules released), decoder (post-synaptic neuron locations) and output set. Indeed, newer studies show that this configuration is more plausible than fixed categories of emotion (Skerry & Saxe, 2015; Dubois & Adolphs, 2015), making bottom-up, heuristic learning the more plausible mechanism for emotional development. Moreover, autoencoders have already been used to generate representations of molecules based on their predicted effects (Gómez-Bombarelli, et al., 2018), suggesting that neurotransmitters and their outcomes could be accounted for by machines. This is of particular relevance in addressing the problem of biological parallels.

Uploading Empathy: Empathic machines, at first glance, appear wildly unintuitive. After all, machines do not have biological substrata, so how could they ever *truly* embody human states and understand our perspective? While this might be true, it is equally true to say that no *human* could ever truly embody the state of another human, because for all their biological similarities, experience of events will likely differ significantly. If, however, empathy is the ability to vicariously experience the emotional state of another, it is possible that machines may end up superior to humans in this respect, provided that an appropriate surrogate for emotional experience (like the one listed above) is present. This boils down to the different methods of learning in machines versus humans.

Empathy is often framed as pure, adulterated emotional response when in reality, there are both cognitive and emotional components that are equally necessary. In fact, when the pain is emotional rather than physiological, there is more activation of the aforementioned reasoning center, the dlPFC (Sapolsky, 2017). In our own brain, bottom-up processes regularly feed information to emotional centers but when empathizing, we must do a top-down reconstruction of the abstract variables of another in order to piece together the emotional picture. This is a cognitively heavy task and as such, not only will there be a great deal of signal loss in attempting to replicate the emotions of others, there is a problem of anticorrelation between the dlPFC and the more emotionally receptive vmPFC. Emotional activation must therefore be weakened in order to spare the cognitive resources of unravelling the emotional state of another (Sapolsky, 2017). As humans practice empathy, they are always experiencing some degree of emotional signal loss of the other person.

Unsupervised machine learning, on the other hand, is done with the algorithm being trained against itself. In the Prisoner's Dilemma scenarios, it is both player 1 and player 2. Therefore, because it does not need to waste computing power on deciphering external experience in emotional contexts, its resources are readily available for use in 'emotional' networks. Thus, it would be trained on millions of observations in a particular context, each time learning that its opponent's emotional status is equally

valuable to its own because they are one in the same. When it does eventually finish training against itself and move on to people, this principle is still being used to calculate the decision to be made when other parties are involved. The weight given to the values and emotional states of others, could, for this reason, be stronger than that given by the average human.

CONCLUSION

The model outlined above, while rudimentary, might suggest fertile ground in the artificial instantiation of moral and ethical frameworks for intelligent machines. Such a paradigm is meant to offer solutions available to us right now that could stand to improve the overall wellbeing of humans. The sooner such a project begins realization, the sooner the training processes can begin expanding and the more effective such algorithms can become. Moreover, beginning the training process now can help prevent humans from being blindsided by the inevitable discovery of intelligent machines that have been programmed in the absence of a moral framework. Inoculation against such an outcome may prevent what could otherwise be a disaster for both our species and our planet.

Bibliography

Adams, T. G., Stewart, P. A., & Blanchar, J. C. (2014). Disgust and the Politics of Sex: Exposure to a Disgusting Odorant Increases Politically Conservative Views on Sex and Decreases Support for Gay Marriage. *PLoS ONE*, e95572. Retrieved from https://doi.org/10.1371/journal.pone.0095572

Aleman, A., Böcker, K. B., Hijman, R., Haan, E. H., & Kahn, R. S. (2003). Cognitive basis of hallucinations in schizophrenia: role of top-down information processing. *Schizophrenia Research, 64*(2-3), 175-185.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(3), 251-261.

Axelrod, R., & Hamilton, W. D. (1981). The Evolution of Cooperation. *Science, 211*(4489), 1390-1396.

Bismark, A. W., Hajcak, G., Whitworth, N. M., & Allen, J. J. (2012). The role of outcome expectations in the generation of the feedback-related negativity. *Psychophysiology, 50*(2), 125-133.

Campbell, J. (2001). Rationality, Meaning and the Analysis of Delusion. *Philosophy, Psychiatry, & Physiology, 8*(2-3), 89-100.

Curry, O. S. (2016). Morality as Cooperation: A Problem Centered Approach. In T. K. Shackleford, & R. D. Hansen, *The Evolution of Morality* (pp. 22-51). Springer.

Damasio, A. (2004). *Decartes' Error: Emotion, Reason, and the Human Brain.* New York: Penguin House.

Dawkins, R. (2016). *The Selfish Gene* (40th Anniversary ed.). Oxford: Oxford University Press.

Dubois, J., & Adolphs, R. (2015). Neuropsychology: How Many Emotions Are There? *Current Biology, 25*(15), 669-672.

Francis, K. B., Terbeck, S., Briazu, R. A., Haines, A., Gummerum, M., Ganis, G., & Howard, I. S. (2017). Simulating Moral Actions: An Investigation of Personal Force in Virtual Moral Dilemmas. *Scientific Reports*, 13954.

Gaertner, S. L., & Dovidio, J. F. (2000). *Reducing Intergroup Bias.* New York: Routledge.

Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., & Brewer, L. E. (2007). Self-Control Relies on Glucose as a Limited Energy Source: Willpower Is More Than a Metaphor. *Journal of Personality and Social Psychology, 92*(2), 325-336.

Glimchar, P. W. (2011). Understanding dopamine reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings in the National Academy of Sciences*, 15647-15654.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., . . . Aspuru-Guzik, A. (2018). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science, 4*(2), 268-276.

Greene, J. D., Somerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgement. *Science, 2105*(293), 2105-2108.

Greene, J., Nystrom, L., Engell, A., Darley, J., & Cohen, J. (2004). The neural bases of cognitive conflict and control in moral judgement. *Neuron, 44*(2), 389-400.

Haidt, J. (1995). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgement. *Psychological Review*(108), 814-834.

Hirstein, W. (2005). *Brain Fiction: Self-Deception and the Riddle of Confabulation.* Masschusetts: The MIT Press.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Anchor Canada.

Kasparov, G. (2019, February 19). Making Sense: The Putin Question (with Garry Kasparov). (S. Harris, Interviewer)

Knapton, S. (2017, October 18). AlphaGo Zero: Google DeepMind supercomputer learns 3000 years of human knowledge in 40 days". *The Telegraph*.

Li, W., Mai, X., & Liu, C. (2014). The default mode network and social understanding of others: what do brain connectivity studies tell us. *Front. Hum. Neurosci., 24*. doi:https://doi.org/10.3389/fnhum.2014.00074

Mitchell, M., & Forrest, S. (1994). Genetic algorithms and artificial life. *Artificial Life, 1*(3), 267-289.

Popper, K. (1959). *The Logic of Scientific Discovery* (2 ed.). London: Routledge.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A Neural Basis for Social Cooperation. *Neuron, 35*(2), 395-405.

Rosch, E. (1999). Principles of Categorization. In E. Margolis, & S. Laurence (Eds.), *Concepts: Core Readings* (pp. 189-207). Cambridge, Massachusetts: The MIT Press.

Sapolsky, R. M. (2017). *Behave: The Biology of Humans at Our Best and Worst.* New York: Penguin Books.

Schulz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science, 275*(5306), 1593-1599.

Skerry, A., & Saxe, R. (2015). Neural representations of emotion are organized around abstract event features. *Current Biology, 25*(15), 1945-1954.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (Second ed.). Cambridge, Massachusetts: The MIT Press.

Suzuki, A., Josselyn, S. A., Frankland, P. W., Masushige, S., Silva, A. J., & Kida, S. (2004). Memory Reconsolidation and Extinction Have Distinct Temporal and Biochemical Signatures. *The Journal of Neuroscience, 24*(20), 4787-4795.

Tajfel, H., & Turner, J. (2004). An Integrative Theory of Intergroup Conflict. In M. J. Hatch, & M. Schultz, *Organizational Identity* (pp. 56-64). Oxford: Oxford University Press.

Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong.* New York: Oxford University Press.