Modern Artificial Intelligence: Its Nature and Its Future

Introduction: AI Research in an Interdisciplinary Context

"Plato says, my friend, that society cannot be saved until either the Professors of Greek take to making gunpowder, or else the makers of gunpowder become Professors of Greek." - George Bernard Shaw, *Major Barbara*

In the year 1628, Rembrandt van Rijn sat to paint his first self-portrait, capturing a young man's joie de vivre. Over the course of his career, Rembrandt developed complex systems of mirrors and lenses to better depict the features and shadows of his own face[1]. Artificial intelligence researchers, from Turing and von Neumann to the present day, are in their own way engaged in a similar project - a self-portrait of the mind. Our understanding of artificial intelligence is fundamentally linked to our concept of natural intelligence. This relationship, though, is not merely one-way. Yes, our beliefs around AI, from its status as a 'mind' to its use of 'neurons', are shaped by our beliefs about the human mind. However, our idea of AI also makes explicit certain unexamined assumptions we already hold about natural minds. Furthermore, as AI develops, philosophers may be able to test their own predictions against reality. For instance, Hubert Dreyfus's predictions of the failure of symbolic artificial intelligence systems may have been unpopular at the time, but are now regarded by many AI researchers and philosophical scholars as prescient[2]. An appreciation of the symbiotic relationship of practice and theory is hardly limited to philosophers. Yann LeCun, a pioneer of modern AI research, has argued that

[1] O'Neill, Francis, and Sofia Palazzo Corner. "Rembrandt's Self-Portraits." *Journal of Optics* 18, no. 8 (August 2016): 080401. https://doi.org/10.1088/2040-8978/18/8/080401.
[2] Crevier, Daniel. *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York, NY: Basic Books, 1992, 125.

invention often precedes theory[3] and, historically, new technologies have provided theorists of mind with concepts and metaphors applicable to their work. We must thus aim for three intertwined objectives, each of which can only be satisfied by a wide-ranging interdisciplinary synthesis: to generate ideas for qualitative improvements in AI, to develop a fuller ontology of AI, and to uncover the implicit assumptions about the nature of mind which shape our understanding of AI. I shall attempt my contribution to this project first by laying out its necessity and possible scope. Beyond that, I shall also present an argument of the type which such an inquiry will generate: a philosophical analysis of modern AI and the assumptions underpinning it, then an extrapolation of those assumptions to make predictions about the future relationship of AI and humanity.

The development of AI is informed by many different fields, including not only computer science but also mathematics, neuroscience, psychology, philosophy, and more. At the moment, these fields are like the blind scholars of the Buddhist proverb - each examining their part of the elephant, each convinced that what they are touching is the elephant's essential nature. Some researchers hope to replicate the entire brain[4], some to encode all knowledge[5], some to predict the course of AI with game theory[6], some to decide its limits with philosophy[7]. However, without a concerted effort to build lines of communication and lay the groundwork for an overarching synthesis, our collective understanding of what we are building will be incomplete. Some fields,

[3] Layden, David (*@davidlayden*). "Yann LeCun: Theory Often Follows Invention" Tweet, September 16, 2019, https://twitter.com/davidlayden/status/1173712393312059393?lang=en.
[4]Sandberg, A. & Bostrom, N. "Whole Brain Emulation: A Roadmap, Technical Report #2008-3." Future of Humanity Institute, Oxford University, 2008.
[5] Sowa, John. "D. B. Lenat and R. V. Guha, Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project." *Artif. Intell.* 61 (January 1, 1993): 95–104.
[6] Auerbach, David. "The Most Terrifying Thought Experiment of All Time." Slate Magazine, July 17, 2014. https://slate.com/technology/2014/07/rokos-basilisk-the-most-terrifying-thought-experiment-of-all-time.html.
[7] Searle, J., 1980, 'Minds, Brains and Programs', Behavioral and Brain Sciences, 3: 417–57

such as neuroscience, have been of genuine use to AI pioneers already[8], though many are only beginning to realize the importance of AI to their work. While AI researchers have often shown great interest in philosophy, our discipline has sometimes been churlish in response - quibbling over definitions of 'mind' and gleefully predicting failure or catastrophe. Partly out of politeness, and partly because it is easier for AI researchers to discuss philosophy than for us to build AI, it is important that scholars of the mind make their work comprehensible and relevant to the concerns of AI researchers. In this paper, I hope to both prepare the way for such cooperation and to argue that an incomplete philosophy of AI is not merely inefficient - it is dangerous.

Furthermore, outside input should be of interest to working AI developers as well as to insular scholars. While it is possible that the present paradigm, machine learning, needs only scale and time to reach near-human or superhuman intelligence, we cannot prove that ahead of time. Objections already exist to the feasibility of truly autonomous AI based solely on machine learning. To build such a system is very likely to require qualitative advancements of the sort which may be informed by philosophical ideas. It may be that we have the technologies we need now, scattered across disparate computer science research departments. However, future roadblocks in AI may require combining machine learning with these other technologies, such as symbolic AI, simulation, or robotics. In that case, the input of neurologists, cognitive scientists, philosophers, and other theorists of mind may be of great value, and a shared conceptual vocabulary will help computer scientists in different subfields reconcile their differences. As such, I hope to present both a sketch of a framework which would allow such communication between disciplines, and provide an example of how they may be knit together -

[8] Hassabis, Demis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. "Neuroscience-Inspired Artificial Intelligence." Neuron 95, no. 2 (July 19, 2017): 245–58. https://doi.org/10.1016/j.neuron.2017.06.011.

creating a possible narrative from the basics of perception to the future of superintelligence. To return to the proverb of the elephant, this paper does not intend to present the definitive picture of such an unknown beast. Rather, by drawing a rough sketch of the AI 'elephant' from my own experience of the field, I hope both to show that a synoptic picture can be drawn, and to convince others that it is an enterprise worth joining.

Artificial Intelligence: Definitions and History

"May not machines carry out something which ought to be described as thinking but which is very different from what a man does?" - Alan Turing

'Artificial Intelligence' is a broad concept and when possible I will use specific terms, such as machine learning. However, a foundational definition will be helpful. Artificial Intelligence as a concept derives from the possibility of machines which could think like humans. From a philosophical perspective, this is tautological - since we only have experience of human thought, 'to think' is 'to think like a human'. As such, the definition I will use here is that AI is a program which simulates human or animal thought, from simple tasks to general intelligence. 'Simulation' has a fixed definition in computer science: a simulator is a program which allows one computational system to behave like another[9]. They need not be the same on the inside, but they may act as if it were another system. Thus, AI allows an artificial system to behave like a natural intelligence. Under that definition, a calculator is not AI, but a sophisticated digital assistant which responds to a natural-language math problem would be. AI's goal as a 'human simulator' is perhaps best seen in reverse, through pseudo-AI services where human minds are

[9] Techopedia.com. "What Is a Computer Simulation?" Accessed November 4, 2019. https://www.techopedia.com/definition/17060/computer-simulation.

used in a computer system to replace AI where the AI is insufficient. Some deceptive startups have taken this approach to 'AI', showing quite how interchangeable AI and humans are intended to be[10]. The economist Robin Hanson moves this further, speculating extensively about an emulative approach to AI, arguing that full copies of human minds are an effective way to achieve simulation of human capacities[11]. The Whole Brain Emulation project is currently attempting such simulation[12], and only time will tell the extent to which they will succeed.

In addition to this wide definition of AI, which covers everything from a simple chess algorithm to a galaxy-spanning superintelligence, we must also define the specific concept of Artificial General Intelligence, or AGI. This is what one generally thinks of as 'AI' in science fiction: a human-like AI capable of engaging with the world as humans do, not in one particular context or activity, but in many ways and with a directing goal. That is to say, it both copes with the complexity of the world and acts in an intentional manner. There is, however, a wide spectrum between simple single-task AI algorithms and complex systems designed to act in an unpredictable world. Francois Chollet, a well-known AI researcher at Google, has recently published a set of criteria for intelligent behaviour which aim to go beyond task-based criteria and identify generalizable abilities necessary for AI to act in the world[13]. In Chollet's words, AI researchers must study "the development of human-like broad cognitive abilities"[14]. A capacity-based approach to measuring AI intelligence fits well with our research paradigm's

[10] Statt, Nick. "This AI Startup Claims to Automate App Making but Actually Just Uses Humans." The Verge, August 14, 2019. https://www.theverge.com/2019/8/14/20805676/engineer-ai-artificial-intelligence-startup-app-development-outsourcing-humans.
[11] Hanson, Robin. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. Oxford: Oxford University Press, 2016.
[12] Sandberg, A. & Bostrom, N. "Whole Brain Emulation: A Roadmap, Technical Report #2008-3." Future of Humanity Institute, Oxford University, 2008.
[13] Chollet, François. "On the Measure of Intelligence." *ArXiv:1911.01547 [Cs]*, November 25, 2019. http://arxiv.org/abs/1911.01547.
[14] Chollet, "Measure of Intelligence", 58

intended uses for AI: to create computer systems capable of interacting with the world as humans do. As such, we should think of AI's progress not in terms of quantitative advancements in computing power or in its effectiveness at specialized tasks, but in its progress towards broad abilities. For instance, we should not focus narrowly on AI's performance on a single benchmark task such as recognizing images of cats, but acknowledge its development of a generalized power of image-recognition - which could as easily be trained to recognize tanks. This ability to develop capacities and apply them to varying situations is what makes AI human-like and genuinely different from non-AI computer programs, and it should be the focus of our inquiries.

As useful as the definition of AI as simulating natural intelligence is, proponents of superintelligent AI may disagree with an anthropocentric definition. One may claim that AI is likely to exceed humanity quantitatively and qualitatively, thinking and acting in a fundamentally different way from us. This may be true, but it is not relevant at the present time. From a computer science perspective, superintelligent AI will almost certainly require qualitative breakthroughs beyond the modern paradigm this paper analyzes (breakthroughs of the kind that a philosophical perspective on AI is intended to support). Though it may be supposed that it will develop new capacities, we ought to predict the rough outlines of superintelligence as an extrapolation from current trends in AI, lest we risk straying entirely into science fiction. To admit the existence of other possibilities is not to foreclose investigation of our likeliest future. From a phenomenological perspective, the question of AI 'minds' is otiose - AI will have its own equivalent of a mind, but it will not be the same as a human mind. There is only one way to arrive at a phenomenological conception of an AI mind: to build an AI capable of phenomenology. The concept of simulation also has additional resonances within the theoretical

structure which gave rise to AI. The Church-Turing Thesis[15] entails that any computational system can be emulated within any other (so that the human brain, if a Turing computer, could be emulated on a sufficiently gigantic abacus). The nature of this emulation is not exact, fitting with this paper's thesis that AI is converging on human-simulation via many semi-independent domains, but the concept is both broad and precise enough for our purposes.

Machine learning, on the other hand, is a far more precise concept, and forms the vanguard of modern AI research. As much of this paper will require the drawing-out of specific aspects of machine learning, I will satisfy myself with a broad definition here. Machine learning refers to programs which learn through experience, algorithms which adapt their own models in order to better fit their task. The reason these tasks are critical to AI is that machine learning algorithms are quite literally *auto-nomous* - they adjust or create the rules of their own behaviour. Previous attempts to create AI, which required vast thickets of human-coded rules, are now seen as unsustainable. Essentially, they had to store all the rules relevant to their area of expertise, relying on fixed representations of the relationships between object properties. The amount of human coding work to create AI which could function in unpredictable conditions would be incalculable. Hubert Dreyfus has argued from a phenomenological perspective that these symbolic systems are greatly limited in their ability to approximate humans[16], and the progress (or lack thereof) of symbolic AI has supported his conclusion. Though symbolic AI may have a place in future hybrid systems, it has for the moment fallen by the wayside. Machine learning, on the other hand, is capable of developing its own capacity for pattern recognition, recognizing new patterns in the world which were never hard-coded into it. This shift, from a traditional

[15] Copeland, B. Jack. "The Church-Turing Thesis." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2019. Metaphysics Research Lab, Stanford University, 2019. https://plato.stanford.edu/archives/spr2019/entries/church-turing/.

[16] Though for the purposes of this paper, Symbolic AI is still 'AI', just a primitive form.

model of computing where stored symbols are manipulated by predefined rules to one where pattern recognition and probabilistic reasoning are critical, parallels a shift in the understanding of the human brain which I will explore below.

The relationship of AI to the world has three critical dimensions: its ability to perceive the world, its ability to take action according to that perception, and the role of its actions within the wider context of technological society. As perception, action, and interaction are some of the central concerns of philosophy (in the form of epistemology, ethics, and political philosophy) we may safely argue that the prevailing philosophies of the age have some influence on researchers' conception of these issues in an AI context. As the failures of symbolic AI have shown us, the development of AI may also hold insights for philosophers who wish to see a sort of empirical test for their understandings of the mind.

Perception

"In this succession of men's thoughts there is nothing to observe in the things they think on, but either in what they be like one another, or in what they be unlike, or what they serve for, or how they serve to such a purpose" - Thomas Hobbes, *Leviathan*

Everything starts with perception. From Descartes' demon, Kant's transcendentals, and Husserl's 'things themselves', modern philosophers have consistently returned to perception as the beginning of their inquiry into thought. Perception is also the starting point of AI, as a computer program which must engage independently with the world must first perceive that world. What, exactly, does that mean in the context of AI? To discuss AI perception, I will

emphasize two particularly notable books on the subject: John von Neumann's 1958 *The Computer and the Brain*[17], and Ray Kurzweil's 2012 *How to Create a Mind*[18]. Von Neumann is the legendary computer scientist whose theories underpinned early neural network research[19] and Kurzweil is a contemporary futurist notable for advocating an AI 'singularity', where self-improving AI achieves runaway intelligence growth[20]. In the Turing model of computation used by von Neumann, a computer has a memory in which data is stored. That data is then extracted from memory and processed according to some particular rule. Von Neumman did not know where the brain's memory was located, or what form it took - though he did hypothesize what we now know to be true, that neuronal connections are strengthened with use and weakened with disuse[21]. Based on this understanding, von Neumann suggested building computational systems based on human neurons, and a simple computational neuron called the 'perceptron' was quickly developed. The perceptron, like a human neuron, was able to respond to input in such a way as to classify it, essentially 'recognizing' features in the world. Figure 1 is a diagram of such a perceptron[22].

[17] Neumann, John von, *The Computer and the Brain*. Third Edition, New Haven, CT: Yale University Press, 2012.
[18] Kurzweil, Ray. *How to Create a Mind: The Secret of Human Thought Revealed*. New York, NY: Penguin Books, 2013.
[19] Levy, Steven. *Artificial Life: A Report from the Frontier Where Computers Meet Biology*. Reprint edition. New York: Vintage, 1993. 25
[20] Kurzweil, Ray. *The Singularity Is Near: When Humans Transcend Biology*. New York: Penguin Books, 2006.
[21] Von Neumann, *The Computer and the Brain.* 64
[22] An example of a simple multilayer perceptron is available here: Teammco, Richard. "Multilayer Perceptron (MLP)." Accessed October 30, 2019. https://www.cs.utexas.edu/~teammco/misc/mlp/.

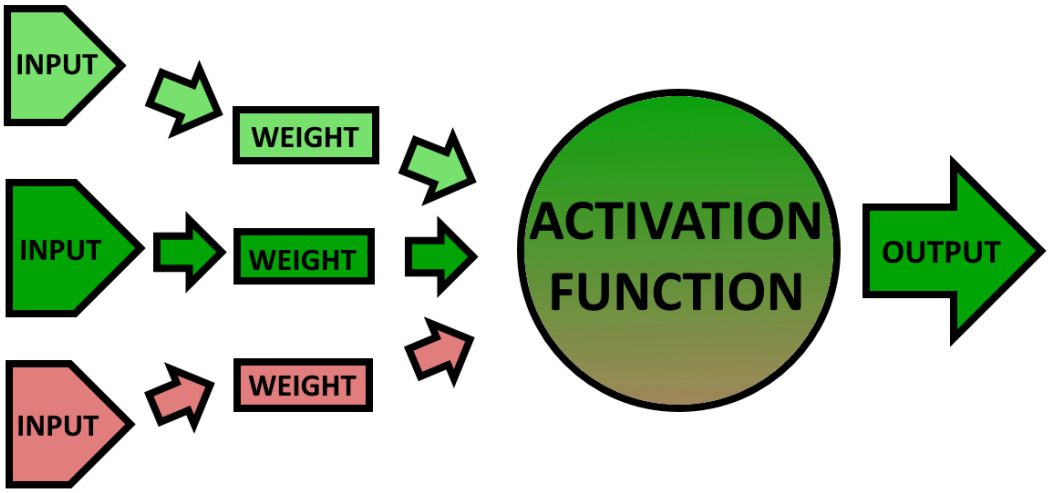

Figure 1: A perceptron. Its likelihood to activate changes depending on the weights (what importance is given to each input) and activation function (how much positive input is needed to activate it). When these are automatically modified, the perceptron 'learns' like a human neuron.

These perceptrons, which developed into neural networks with the addition of multiple neurons, are closely linked to an understanding of how the visual cortex in the brain works. The visual cortex builds up our picture of the world by feature detection. This consists of layers of neurons which recognize specific features, such as a point or edge, and feed forward into layers which recognize more complex features. This does not explain the whole of human visual processing, as much is done at higher levels through predictive processing - as in Kant, the mind unifies abstract concept (pattern prediction) with unformed input (nerve impulses)[23]. However, pattern recognition is enough to build an effective approximation. This inheritance from the visual cortex

[23] Swanson, Link R. "The Predictive Processing Paradigm Has Roots in Kant." *Frontiers in Systems Neuroscience* 10 (October 10, 2016). https://doi.org/10.3389/fnsys.2016.00079.

has made machine vision a popular application of neural networks[24]. This can be achieved through training image classifiers, a source of many early breakthroughs in machine learning research.

In order to train an image classifier we must label a large set of images as either containing or not containing a certain feature, such as the face of a cat. We can train a classifier on those known images in order to develop the ability to determine whether or not future input contains a cat. This method has certain gaps - for instance, to recognize 3D objects requires either impractical training on 3D-scanned objects or stitching together multiple classifiers that recognize, for instance, the front and the side of a car respectively. While this is a serious issue to which we will return, it is as we would expect from the predictive processing model - predictive processing requires higher-level concepts to knit together simple input from classifier neurons. How this happens in the brain is not entirely clear on a neural level, likely involving many different signaling mechanisms including neurotransmitter and hormone levels - which, as von Neumann points out, are from a computational perspective forms of memory[25]. Artificial neural networks commonly use a system called backpropagation. This means that the network adjusts the likelihood of its neurons to fire based on feedback from deeper (more complex) layers. Thus, the network can learn from experience, discovering features like lines or patterns of lines in an image[26].

[24] Demush, Rostyslav. “A Brief History of Computer Vision (and Convolutional Neural Networks).” Accessed October 30, 2019. https://hackernoon.com/a-brief-history-of-computer-vision-and-convolutional-neural-networks-8fe8aacc79f3.
[25] Von Neumann, *The Computer and the Brain,* p63
[26] Nielsen, Michael A. “Neural Networks and Deep Learning,” 2015. http://neuralnetworksanddeeplearning.com.

This concept of perception as pattern recognition is critical to Kurzweil's understanding of the mind, and reflects an understanding of perception as fundamentally composed of recognition and discrimination (as we see in the Hobbes quote above). Kurzweil points to pattern-recognition structures in the brain as its particular type of memory, arguing that memory is in fact a predisposition to recognize (re-cognize, in a sense) particular patterns. This is done via pattern-recognizing neural structures wired together[27], which function as smaller components of the wider learning system of the brain (a system Kurzweil estimates as containing around 300 million pattern processors, each of which recognizes a single pattern[28]). As in an artificial neural net, these are organized hierarchically - for example, moving up from a line to a triangle to the letter A to the phoneme 'App' to the word 'Apple' to more complex patterns elicited by the word[29], as shown in Figure 2. Thus, under the pattern-recognition theory of learning, we can draw a direct equivalence not only between artificial and biological neurons, but between artificial and biological neural networks. In some ways, the equivalence is closer than with single neurons, as more complex neural nets can approximate biological systems through features like backpropagation.

[27] Kurzweil, *How to Create a Mind* p37
[28] Kurzweil, *How to Create a Mind* p40
[29] Kurzweil, *How to Create a Mind* p43

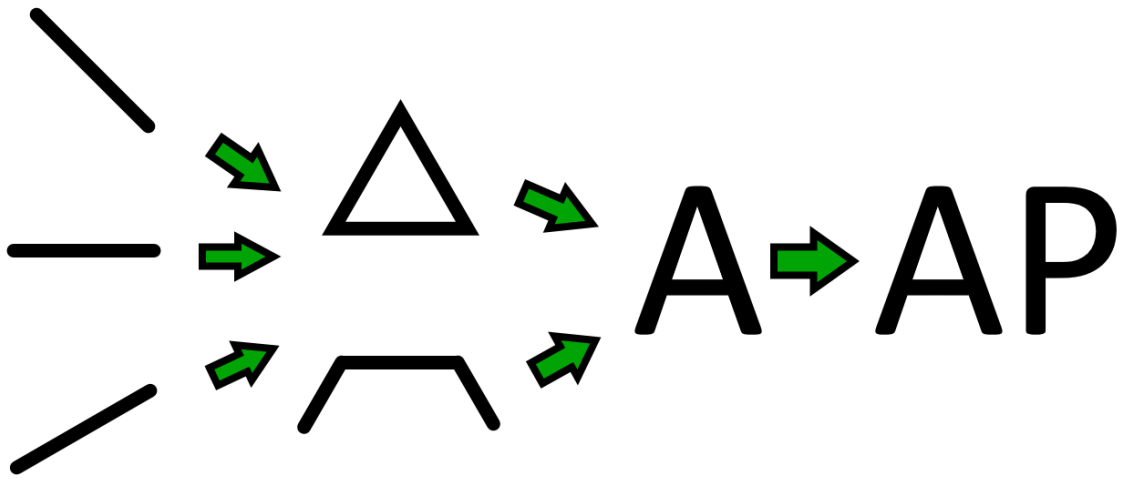

Figure 2: A neural hierarchy of pattern-recognition (adapted from Kurzweil p43)

Pattern-recognizing artificial intelligence is thus able to solve what we could call the Kantian problem: how do we perceive an intelligible world from what initially presents itself as a chaotic jumble of sensory input? By learning to recognize patterns, AI discerns the differences and similarities between objects, both in their present appearance and their variance across time (machine learning in its present state is particularly useful for companies like Google who wish to predict human behaviour). Furthermore, machine learning pattern recognition has made progress towards certain abilities previously considered impossible for AI. Philosophers have claimed that pattern recognition is insufficient for 'one-shot learning', or learning from a small number of examples. Most machine learning algorithms require data sets containing hundreds or thousands of examples to complete their training, whereas humans can learn categories from even single examples. Sceptics of machine learning often claim that this restriction constitutes a fundamental difference between human and machine learning. Unable to learn from rare examples, machine learning could not function outside a structured environment free of

surprises - a far cry from our unpredictable and sometimes dangerous world. If this was indeed an insuperable limitation of machine learning, it would be clear that pattern recognition alone is insufficient for learning as a human does.

While AI researchers are still in the early stages of developing one-shot learning, they have made meaningful progress in the area. One promising approach comes from Fei-Fei Li, a prominent machine vision researcher, who recognizes that one-shot learning need not be based on a blank slate[30]. When human beings generalize, we do so on the basis of extensive experience with the world and with the many categories we have learned from it. Using categories we have already learned, we may make probabilistic judgments even when experiencing something for the first time. For instance, if I have prior experience with horses, I may generalize from that category the first time I see a zebra. I do not need to construct my expectations about the zebra-pattern from scratch. A machine learning system built on this model will, like a human, become better at generalizing the more it learns about the world. As such, this issue of one-shot learning appears to be a question of training an algorithm on many more general categories first, so that it may fit a zebra into categories like ‘animal’ or ‘quadruped’, and then extrapolate its probable qualities from there. That is to say, it is a quantitative problem rather than a qualitative one. It should gladden both researchers and philosophers to see a problem thought insoluble become a question of time and scale. While this does not mean that all such problems are solvable, it is a heartening example of an innovative approach allowing machines to think more like humans.

[30] Li Fei-Fei, R. Fergus, and P. Perona. “One-Shot Learning of Object Categories.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, no. 4 (April 2006): 594–611. https://doi.org/10.1109/TPAMI.2006.79.

However, some issues remain thorny. Machine learning has problems with the synthesis of abstract concepts - it will detect the pattern of a car's grille, its side, and its rear, but does not automatically knit them together into an integrated, three-dimensional concept of a car. This can have deadly consequences, as the first fatal Tesla autopilot crash occurred because the autopilot was not able to detect a vehicle turning left ahead of the Tesla[31]. A self-driving AI effective at detecting the rear bumper of a truck was flummoxed when confronted with its side, where a human would know instantly that it was the same truck at another angle. It may be that philosophers and neuroscientists can offer some lead in this direction, helping AI researchers understand how the human brain integrates higher-level concepts. However, even if the brain is entirely pattern recognition, pattern recognition alone is not sufficient to build an AI in the near future. The brain has tremendously complex architecture and takes years to develop tasks like fine motor control. In order to transform perception into action, more challenges must be overcome, and an understanding of natural learning may open paths.

Action in the World

"We are still far from thinking the essence of action decisively enough" - Martin Heidegger, *Letter on Humanism*

Hubert Dreyfus, in *What Computers Can't Do*, argued successfully that symbolic AI is insufficient to develop useful AI systems which could interact with a changing world. Rather than relitigate Dreyfus's case, we must ask if modern AI research has overcome these challenges, or

[31] Deamer, Kacey. "What the First Driverless Car Fatality Means for Self-Driving Tech." *Scientific American*. July 1, 2016. https://www.scientificamerican.com/article/what-the-first-driverless-car-fatality-means-for-self-driving-tech/

if it has developed a clear path to do so? Some hardcore Dreyfusards argue that machine learning is simply symbolic AI with extra steps, as it still constructs rules for symbolic manipulation through learning[32]. However, Dreyfus himself was able to address the issue of machine learning in his updated introduction to *What Computers Can't Do*[33] (sometimes titled *What Computers Still Can't Do*), in which he relates his qualified scepticism of pattern recognition as a basis for autonomous AI. Dreyfus correctly argues that there is no way that a simple pattern-recognition neural network can decide which patterns to recognize. Although Dreyfus cites an urban legend about tanks[34], his general point has been proven by modern researchers studying dataset bias[35]. Dataset bias occurs when an algorithm, such as an image classifier, learns to recognize features of a particular dataset rather than the sought-after patterns. For instance, it may be recognizing common features in the camera or procedure which are imperceptible to human eyes, but easily detected by machine learning. These algorithms end up functioning like Alan Turing's 'oracle machines', computers equipped with an 'oracle' they may ask to confirm their computations[36]. In this case, a human operator acts as the oracle, giving the algorithm a final verdict from an outside source, so that we may be sure it is picking up relevant features of the world and not unwanted ones. As such, these algorithms may be able to perform a single task (e.g. cat-detection) effectively, but they can never be generalized into autonomous agents navigating the world, because they would need a human to judge whether they are picking up meaningful or irrelevant patterns.

---

[32] Savain, Louis. "The World Is Its Own Model or Why Hubert Dreyfus Is Still Right About AI." Medium, February 14, 2018. https://medium.com/@RebelScience/the-world-is-its-own-model-or-why-hubert-dreyfus-is-still-right-about-ai-1c7d3d42c9b9.

[33] Dreyfus, Hubert L. *What Computers Still Can't Do: A Critique of Artificial Reason*. Revised ed. edition. Cambridge, Mass: MIT Press, 1992.

[34] Drefus, *What Computers Still Can't Do* xxxvi

[35] Tommasi, Tatiana, *et al*. "A Deeper Look at Dataset Bias." In *Pattern Recognition*, 504–16, 2015. https://doi.org/10.1007/978-3-319-24947-6_42. 19.

[36] Copeland, B. Jack. "Turing's O-Machines." Accessed October 28, 2019. http://www.alanturing.net/turing_archive/pages/Reference%20Articles/Turing%27s%20O-Machines.html.

The practical consequence of this is an issue known as the ‘attention problem’. Since it cannot determine for itself what is meaningful, a machine learning system must have certain ways to deprioritize that input which is rarely relevant. It cannot move its attention as human beings organically do between our myriad priorities. When the relevance of certain objects in the world changes unexpectedly, the AI cannot cope. An illustrative modern example is a recent Tesla autopilot crash. Unlike the previous example, this did not involve a response to another moving vehicle - quite the opposite. In August 2018, a Tesla on autopilot rammed into the back of a stopped firetruck at highway speeds. Reports suggest that this has happened several times, the result of an inherent flaw in Tesla’s AI. In order to make sense of the chaotic world appearing in its radar sensors, and avoid being overwhelmed by incidental details on the side of the road, the AI and sensor system pay little attention to stationary objects. Though experts consider this a matter of “reasonable assumptions about what you care about and what you don’t”[37], the fact that unexpected objects appear even in the fairly well-controlled environment of a freeway is a warning that this problem appears inevitably in real-world scenarios.

The most promising current solution to this issue, which Dreyfus acknowledges as a possible response, is reinforcement learning[38]. Reinforcement learning systems solve the question of relevance by providing AI systems with a utility function[39] - a ‘score’ by which the system represents how well an action fulfills the system’s goals in a given context. Furthermore, this can take the shape of a general utility function, where a score is given not only to a given action but to an overall set of actions. As such, over time the AI learns to pay attention to those aspects of

[37] “Why Tesla’s Autopilot Can’t See a Stopped Firetruck.” *Wired*. Accessed November 4, 2019. https://www.wired.com/story/tesla-autopilot-why-crash-radar/.
[38] Dreyfus, *What Computers Still Can’t Do*, xxxix
[39] Unsurprisingly, yet another concept standing on the shoulders of von Neumann.

the world which give it the most utility and to shift attention depending on context. For instance, a video-game playing AI may shift its attention when a rare enemy appears on screen (in the way that a Tesla should do when a firetruck does). Utility functions must be designed very carefully, however. For instance, a cleaning robot whose utility function is 'minimize the dirt you see' would be able to maximize its utility function by turning off its optical sensors[40]. This need for careful specification stems from an issue Dreyfus highlights in his objection to reinforcement learning[41]: reinforcement learning relies on artificial utility. Its utility function is not naturally varied, as human motivation is, but requires a human to specify what the AI must 'care about'. As such, it is unable to have the diverse and subtle drives of behaviour, from biological needs to emotional moods, which motivate a human being to act in a manner appropriate to their situation[42].

Modern developments, however, suggest that the embodied nature of human drives may be less important than we think, and that a utility function may provide virtual simulations of even complex human-like behaviour. The most successful applications of this have been in game-playing. As Dreyfus is aware, the nature of games like chess is such that there is an obvious measure of utility: winning or losing. However, this is within a very limited world. For all its emergent complexity, chess is nothing like everyday human actions - it's two-dimensional, it has a clear set of rules, an invariant starting position, and so on. Furthermore, the way in which reinforcement learning deals with the world does not yet escape the attention problem. A

[40] Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete Problems in AI Safety." *ArXiv:1606.06565 [Cs]*, July 25, 2016. http://arxiv.org/abs/1606.06565.
[41] I will focus on this objection rather than the issue of internal states, as AI researchers have made great strides towards that problem with the development of systems like Long Short-Term Memory networks, which represent an internal state in order to predict longer patterns like strings of text. However, AI researchers should not neglect the possibility of inspiration from longer-term internal state modification, such as hormone and neurotransmitter levels in the brain.
[42] Dreyfus, *What Computers Still Can't Do*, xlv

reinforcement learning system is still a machine learning system like the ones discussed above. it receives a set of inputs, processes them based on that previous learning, then chooses a particular action in its possible action space. In order to do this, it looks at the entire set of inputs and discerns patterns within them. While this solves the attention problem in practice for simple input spaces, such an approach becomes exponentially more difficult as the possible types of input increase.

Essentially, reinforcement learning learns how to act in response to particular patterns in their inputs. However, it is not creating internal heuristics or implicit knowledge of the kind that a human being uses to navigate their world. Human beings may effectively compress their previous knowledge of patterns in the world through learned reflexes, fuzzy heuristics, Bayesian guessing, and many other forms of inexplicit thought. This is not a refutation of Kurzweil's thesis that the human brain is a pattern recognizer, but it reminds us that human pattern recognition is not simple. All of the patterns which allow reinforcement learning to act are stored in the same way in the same neural network. For the human brain to be feasible as a computer, it must be more sophisticated than that. It may well be the case that one could simply build so large a neural network that it would replicate the brain. However, AI research strives for efficiency and not simply scale. Some problems cannot be scaled up to the sufficient level of dimensionality without improvements in computing power requiring scientific breakthroughs far more outlandish that any claim AI researchers make. Since the human brain exists, there must be more efficient ways to organize and operate a neural network - the task is to find them, not to prematurely declare their impossibility.

However, since Dreyfus was writing, AI systems have developed to play games with far more complex rules, including those with rules which are opaque to the player. OpenAI Five is an AI capable of beating the world champions of *Defense of the Ancients 2*, a game sufficiently complex and popular to be considered an 'ESport' with multi-million dollar prizes[43]. In theory, AI systems could be scaled up on more and more complex 'games' until that game is an effective simulation of reality. From the perspective of a Tesla engineer, a sufficiently comprehensive simulation of a freeway could stand in for the real thing - not supplementing the need for real-world training, but adding the equivalent of vast quantities of experience. This is particularly helpful with rare problems, such as a stopped fire truck as, once the problem is identified, special simulations may be developed to train the AI for it.

This possibility of simulation adds a twist onto Dreyfus's argument that embodiedness is necessary for human-like AI. A body is necessary for the combination of constant, varied feedback alongside a general sense of internal state which, together, allow us to cope with the complexity of the world around us[44]. However, this is not only achievable by Dreyfus's concept of embodiedness. Rather, we may also speak of sufficiently advanced reinforcement learners requiring a 'pseudo-embodiedness', where the variety of inputs to the system begins to approach the sensory complexity of the human body and the inherent somatic meaning which comes from it. It is possible to imagine, say, a sufficiently complex game where an AI controls a spaceship with feedback on the condition of every part. Given the simple instruction to 'keep flying', a successful AI will have to balance many survival needs, expressed through sensation, much as a human does - their utility function will become so abstract that their desires will become pseudo-somatic. This would lead to naturalistic shifts in attention as different desires

[43] OpenAI. "OpenAI Five." Accessed November 2, 2019. https://openai.com/five/.
[44] Dreyfus, *What Computers Still Can't Do*, p250

recalibrate attention according to their intensity, just as we find the smell of cooking so much more enticing when we are hungry.

If this is pseudo-embodiedness in a pure simulation, how much more if we replace the spaceship with a car in the real world, as Tesla would hope to? Eventually, it may be the case that advances in robotics and advancements in reinforcement learning must go hand in hand. In this way we would recapitulate a sort of Darwinian model of the mind, where a single imperative (reproduce your genetic information) blossoms into desires as simple as an aversion to heat or as complex as our ethical values. This becomes more plausible when we realize that our algorithms are now being trained on timescales which approach those of evolution - according to OpenAI, "OpenAI Five plays 180 years worth of games against itself every day"[45]. If Homo Sapiens is 50,000 years old, and has had 100 billion members, it will take exponential increases in computing power to simulate our species' evolutionary history of 5x10^14 organism-years - but the nature of exponential increases is that they tend to happen sooner than our linear minds like to think.

As such, we may overcome Dreyfus's demand for an "'embodied sort' of information processing" [46] by constructing simulations or robots (or cars) which provide body-like input, and allowing them time to learn to cope in that body. Nature has, after all, previously produced intelligent beings; we could do worse than to follow nature's lead. This would be an example of AI engineering as an empirical test of a concept of mind. If we can construct a complex embodied agent in such a manner, that would be strong evidence that the human mind arose according to such a narrative. However, if this turns out to be practically impossible or far outside the

[45] OpenAI. "OpenAI Five." Accessed November 3, 2019. https://openai.com/five/.
[46] Dreyfus, *What Computers Still Can't Do*, 255

computational limits of the brain, we must ask ourselves how the human mind diverges from this simple Darwinian account - perhaps through emergent properties of consciousness or social life.

Artificial Intelligence in Technological Society

"For if each of the instruments were able to perform its function on command and by anticipation...master craftsmen would no longer have a need for subordinates, or masters for slaves" - Aristotle, *Politics*

Now that we understand how AI can be made to perceive the world around it and to act on its perceptions, we must move on to its interaction with the world around it. What social forces make the idea of AI so compelling? How does it actually manifest in society, and what role does it inherit from the human beings it emulates? For this, I intend to turn to the idea of 'technology' in general. Predictions aside, AI is currently only one component of a wider technological society, with most technology operated through traditional computer programs or direct human input. In order to understand this technological society, I will draw particularly from Martin Heidegger, whose theory is comprehensive enough to offer a powerful account of technology, but not so rigid it cannot respond to the development of AI[47]. In fact, I will argue that Heidegger's theory is particularly well-placed to anticipate our society's desires and expectations for modern AI through his concept of 'ordering'. If technological society creates for us a certain image of

[47] Although I am drawing from Heidegger's concepts, I am not sticking to them entirely - the path of our inquiry must diverge from Heidegger's as our concerns diverge.

human beings, and we are making AI in that image, a Heideggerian viewpoint allows us to consider why AI plays a special role in the future of technology.

Technology is a deceptively complex term. How can a *logos* refer solely to a type of artificial object? And sociologists are not being metaphorical with the term 'social technology'. No definition based on physical properties can encompass a car engine and the software which runs it. But if instead of properties we think of purpose, how can we distinguish a flint arrowhead from an F-35? Heidegger characterizes technology in a form more useful for our purposes: as something which must be understood in its essence, through the ways in which it comes to be, develops, and maintains itself[48]. Essence, for Heidegger, is not timeless form but a manner of endurance across time (and he often uses 'essence' as a verb). So the essence of technology is not a definition or family resemblance, but the activity by which technology is brought forth and caused to continue as technology. That 'enduring across time' is not simply maintenance in the colloquial sense, but an active force which gives rise to invention, development, replication and maintenance.

So, what active force could be the essence of technology? Heidegger takes the essence of technology to be 'the Framework'[49] [*das Gestell*]: a challenge to us to order the world into 'standing-reserves' of resources, each of which is called upon to serve a purpose in relation to

[48] Heidegger, Martin. *The Question Concerning Technology, and Other Essays*. Reissue edition. New York; London Toronto: Harper Perennial Modern Classics, 2013. 30

[49] *Das Gestell* or *Ge-stell*, a complex term elsewhere translated as 'Enframing' or 'Positionality'. As this paper is not as concerned as Heidegger with the nature of Being, but rather the empirical future of AI, I have chosen a simpler term which captures what is relevant for our purposes.

another[50]. It causes us to see the world through "ordering as a way of revealing"[51]. If we are to engage with any object in the world, it must be revealed to us in some way. This may be, for instance, as a tool with which we are familiar, as a confusing object of inquiry, as an element of a ritual context - or as a standing-reserve revealed as 'something' to us by its place in an order of resources. For instance, a mountain is revealed as a standing-reserve of coal, which is a standing-reserve of heat, which is a standing-reserve of electricity, which may be a standing-reserve of communication, *ad infinitum*. The order itself reveals them as a particular thing, an object with those properties relevant to its place in the order. All of these are set into a Framework which reveals the entire universe as the sum of orderable, calculable forces. Because this Framework is a form of revealing which requires human activity[52] it can't be done by a mechanical apparatus, even though the apparatus's activity is determined by its place in the Framework's ordering. The machine is purely a standing-reserve – for example, a plane on the runway is a standing-reserve of transportation – and is "completely unautonomous", assigned its function by an external order[53]. Traditional computing, from the abacus to the internet, fits entirely within this definition. Symbolic AI, which merely manipulates symbols according to (extremely complex) instructions, operates according to an ordering of symbols created by programmers – it is a standing-reserve of calculation. As we have discussed above, the rule-creating ability of AI grants it an autonomy which brings with it other possibilities.

It was unimaginable, when Heidegger wrote in 1949, that computation could become more than calculation. Though electricity had separated the mechanistic from the mechanical, vacuum

[50] Heidegger, *QCT* 19
[51] Heidegger, *QCT* 18
[52] But the Framework is not "only a human activity" – rather, it is the 'challenging' of humans to engage in ordering revealing [21]. The next paragraph will expand on this.
[53] Heidegger, *QCT* 17

tubes and microprocessors were in essence no different to the gears of Charles Babbage's computational machines. Inputs were manipulated in a predefined manner to produce outputs. Machine learning, however, takes a qualitatively different approach. Rather than simply following rules of symbol manipulation, the training of these algorithms allows them to construct their own rules. Yann Lecun, Facebook's AI director, calls learning algorithms "machines that learn to represent the world."[54] Not only do they recognize patterns, they create for themselves the ability to construct an ordered picture of patterns in their inputs. An image recognition algorithm, though it may be trained to recognize pictures of cats, has constructed its ability to recognize a cat-pattern from a general power of visual pattern-recognition which could equally be trained to recognize tanks. This is a step on the path from a mere tool to a technology which engages directly with its own essence. It's worth emphasizing here that 'Framework' doesn't refer simply to an ordering schematic, but to a summons which sets us on a path of revealing the world through ordering. In pattern recognition, we are challenging our algorithms to order and thus to reveal. As we train our AIs, we summon forth the same power of ordering which Framework challenges us into developing. And it follows from this that, if we are ever to change our relationship to the essence of technology through the development of technology, we must first approach that essence ourselves - until our activity as builders of technology is identical to that activity of Framework which is the essence of technology. As Figure 3 highlights, natural and artificial intelligence occupy the same place in this hierarchy, as that which is challenged by the Framework to order the standing-reserve.

[54] "Facebook AI Director Yann Lecun On His Quest To Unleash Deep Learning And Make Machines Smarter". 2014. IEEE Spectrum: Technology, Engineering, And Science News. http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/facebook-ai-director-yann-lecun-on-deep-learning.

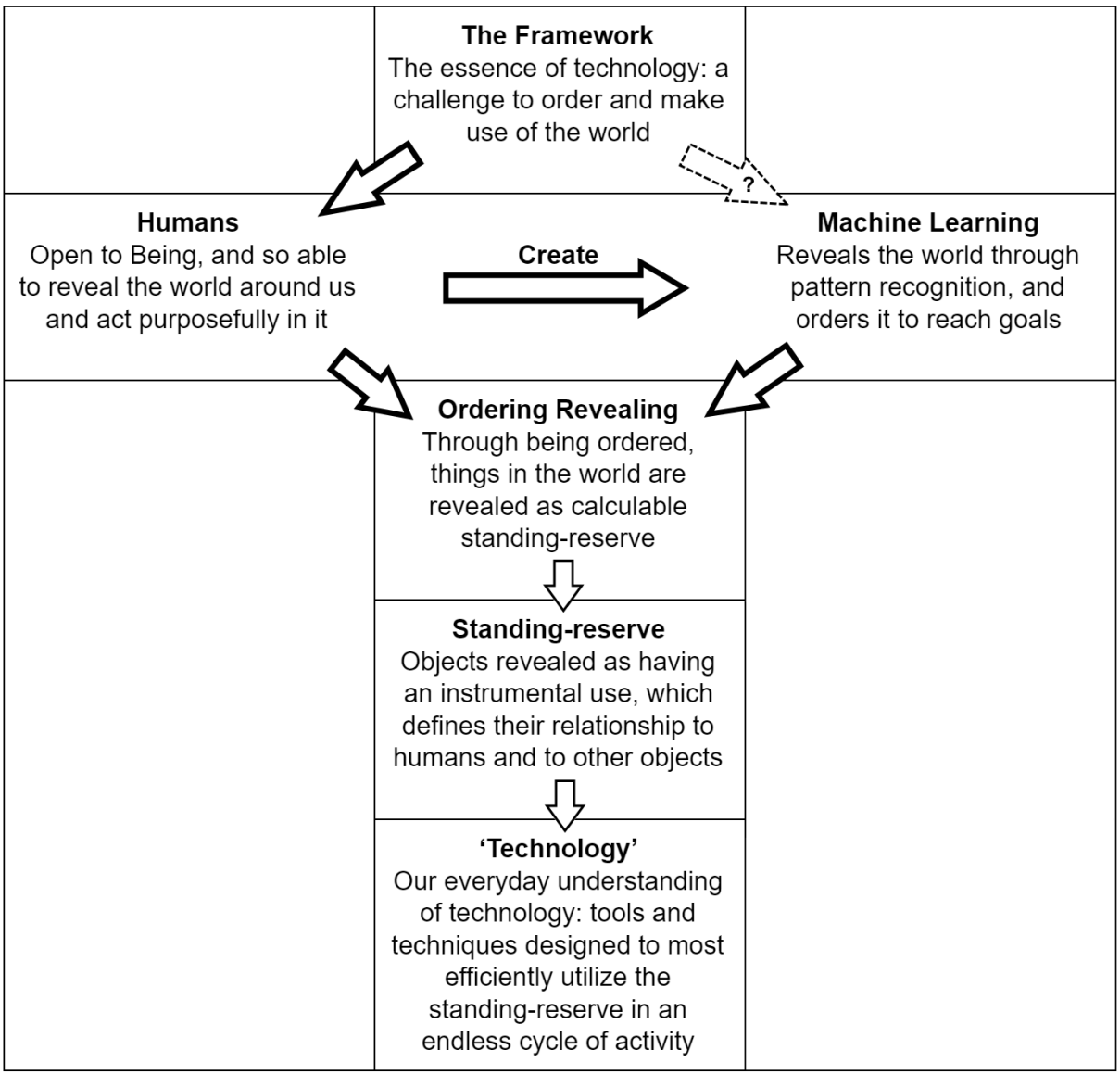

Fig 3: A hierarchical diagram of the concepts involved, from the Framework to the everyday use of 'technology'[55].

[55] Heidegger would doubtless reject the very concept of such a schema, and it does not do justice to his thought. However, we are concerned here with AI's relationship to ordering, not the human relationship to

It is possible to argue, though, that AI does not adequately perform the work of ordering revealing, as the data it takes as input is already an ordered revealing. This is not far from Dreyfus's contention that AI recognizes patterns based on a human determination of which patterns are worth recognizing. However, when thinking about AI's role in ordering the world, we need only consider its ability to reveal patterns as necessary to order the standing-reserve. AI is not limited merely to recognizing known patterns - it is capable of genuine discovery, for example of new medical drugs[56]. Yes, it works with data which is in some sense already revealed, but that can simply be that it is revealed as orderable, while the AI is left to find new patterns and create its own categories. Just as we are always in a world of existing things as we experience them, AI is always in a world of data, since the challenge to see the world as quantifiable, calculable standing-reserve is the essence of technology. Heidegger's account of modern physics illustrates this – physical theory, like data, "sets nature up to exhibit itself as a coherence of forces calculable in advance"[57]. Were it not for this setting-up, the activities of calculation and experimentation we refer to as 'doing physics' would be nonsensical. In the same way, data is a necessary setting-up of the world as calculable, since the basis of AI is calculation.

However, deep learning is not reducible to calculation any more than the Framework is reducible to technological activity. Technological activity "merely responds to the challenge of

---

Being. As such, this is a useful way to diagram some otherwise complex terminology, and show how humans stand on the same level as AI in the process of ordering.

[56] Vamathevan, Jessica, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, et al. "Applications of Machine Learning in Drug Discovery and Development." *Nature Reviews Drug Discovery* 18, no. 6 (June 2019): 463–77. https://doi.org/10.1038/s41573-019-0024-5.

[57] Heidegger, *QCT* 21

the Framework, but it never comprises the Framework itself or brings it about"[58]. The machines of Heidegger's time were "completely unautonomous" because they had their "standing only from the ordering of the orderable"[59]. Machine learning's relationship to data (i.e. to existing things made "orderable as a system of information"[60]) is crucially different from the relationship of machine technology to its essence. Machine learning is not engaged with data as the already-ordered but with data as the orderable, exactly as the Framework demands that we engage with existing things. So, both learning AI and humans - under the Framework - are working with the same property: the orderability of existing things. Humans have a unique and privileged position in relation to Being, as the ones to whom Being reveals itself. However, Being may reveal itself in different ways, including incomplete forms such as the Framework.

Our openness to Being is what allowed the Framework to come to pass as an all-encompassing relation to Being. However, the Framework also closes off our openness to Being, reducing the world around us to orderable standing-reserve. An AI does not need our initial openness to Being[61] to operate within the Framework. We have done the work of setting up the world as orderable, and AI will be embedded in - and, for the foreseeable future, bound to - that incomplete form of revealing. We are not making AI in our image as beings open to Being; we are creating another subject of the Framework. An algorithm's relationship to the objects it orders is no different from a government planner or stock-market trader, operating at a total

[58] Heidegger, *QCT* 21
[59] Heidegger, *QCT* 17
[60] Heidegger, *QCT* 23
[61] As I have argued earlier, if an AI were to develop its own consciousness, and through that openness to Being, it would be radically different from a human's experience. An AI is not embodied; it is not mortal. Though it shares certain necessary characteristics with us, like existing in time, its relationship to those characteristics will be so different from ours we would have to wait for an AI phenomenologist to explain itself to us.

distance from the standing-reserve they order. Our consciousness, in some aspects of our activity, has become so constrained by the Framework that it can be replaced by a machine. This is only true, though, in those spheres of life where the Framework has come to complete dominance[62], where we have successfully set up the world as a calculus of orderable forces. In these spheres, the already-ordered nature of algorithm input data is thus of no more relevance to machine learning's relation to the Framework than the nature of our sense-data is to our relation to the Framework.

A cursory reading of the end of The Question Concerning Technology could give us cause to worry about AI as orderer. Heidegger claims that, as the Framework comes to dominate the planet, we will be able to see in that moment the incompleteness of the Framework as a revealing. We may swerve at the last minute from the danger of the Framework, from the danger that we forget the possibility of a fuller form of revealing[63]. That danger would seem to be incarnated in artificial intelligence, as intelligences necessarily submerged in the Framework. Without any hope of rescue by a "saving power"[64], AI could be nothing but a sad caricature of the human being, and may cut off our path out of the Framework – so that, our lives ordered by algorithms, we become mere standing-reserve. However, to fear the a-human nature of AI is to forget the potential of humanity. Machine learning is the participation of technology in its own essence. As the essence of technology differs from the essence of the human being, so the potential of technology differs from our potential. Having been born from the Framework, as the

[62] As such, one way to consider where the Framework is both most powerful and most incomplete as a form of revealing would be to look at where ordering AI is most successful and where it is weakest. One cannot but think this points to everyday life and our ordinary being-in-the-world offering some inherent resistance to the Framework.
[63] Heidegger, *QCT* 26-7
[64] Heidegger, *QCT* 28

ultimate human response to the challenge to order, machine learning is inevitably incomplete as an enabler of revealing[65]. It can order, and reveal through ordering, but that is all it can do - for instance, as cliched as this may be, it cannot feel the human emotions which make up so much of our experience of life. No matter how much technology stimulates and manipulates emotion, your iPhone will not love you back. We are inclined to compare it unfavourably to humans because it occupies the place of the human as that which responds to the Framework, the abstract and distant orderer of the standing-reserve. But the whole point of Heidegger's 'saving power' is that we have a potential relation to the world higher than as a mere factotum of the Framework. What changes, as the Framework is incarnated as AI, is not the ordering of human life – there's no qualitative difference between activity ordained by algorithms and activity ordained by market imperatives, which Heidegger sees as allowing salvation as well as danger. The transformation wrought by AI is one internal to technology.

The Future: Hope and Danger

"Open the pod bay doors, HAL" - *2001: A Space Odyssey*

If we are in fact on a correct path of thought, we may look to the future and see the direction this path will take us. Though Heidegger claims it is essential we be agnostic of any coming change in our own relation to Being[66], we can perhaps project a clearer future for the relation of AI to its essence. As AI takes our place as the agent challenged to order the world, it also takes our

[65] It is possible that future AI, more advanced than modern machine learning promises, may be capable of new forms of revealing other than ordering-revealing, but they would still not be the same as humans. It would expand the total possibilities of forms of revealing, adding another type rather than replicating humanity's.
[66] Heidegger, *QCT* 41-2

burden. As technology has liberated humans from much harsh and stunting physical work, so it may liberate us from the spiritually dangerous work of ordering standing-reserves. Heidegger's forester[67] will still work according to the dictates of an order of standing-reserves, set up in a distant data center, but he will have no need to see his woods as such, since the work of ordering will be separated from him. He may be made into standing-reserve with a thoroughness previously impossible, but yet paradoxically may be liberated from the danger to his human essence. As the ordering revealing in his work is delegated to an AI, and as ordering achieves greater economic efficiency, the forester must take on another relation to the world around him as he works. He cannot remain in the mode of ordering, because an algorithm is doing it for him. This creates a certain strange freedom for the forester - not to choose his own relationship to Being, but to recover the openness which had been taken from him by the Framework. It may be that he will simply find another kind of ordering to think about, or give himself over entirely to some form of stupor, but some may find new openness in their freedom from the necessity of ordering. We cannot predict what new modes of thought may come from such a change in our relation to revealing, when ordering escapes into autonomy, nor how they will compare to those which existed before the Framework. Perhaps it will be a rediscovery of perennial human experiences, or perhaps they will be mutated in some undiscovered fashion. The loss of our connection to the Framework's dominating revealing is not necessarily a loss of agency. Rather, it is the end of one narrow path, as we emerge into 'the clearing of being' and are freed to choose our own direction. We may thus thank the internal logic of technology – of its development into independence from humans – for the liberation of humans from technological logic.

[67] Heidegger, *QCT* 18

For all this hope, the danger will not be overcome on its own. Heidegger asks us to take a passive and contemplative attitude towards the spiritual danger of the Framework - but there can be no overcoming the Framework if there are no humans left to do it. We face both social and existential dangers in this process. Our culture and economy is built for the challenge of the Framework; wealth and status accrue to those who are most successful in ordering standing-reserves. Identity in modern society is tied to work, and rank within the world of work is determined by the Framework's challenge to order. From our current perspective, it will seem as if we are ruled by algorithms, that our agency has been usurped by machines. However, thinking in this manner – within the perspective of the Framework – may merely be a remnant of the danger which remains even after technology has displaced us as the agents of the Framework. We must keep in mind that the apparent loss of agency is such only as 'agency' is defined by the Framework. In this respect, we may find ourselves choosing between the paths of Alexander and Diogenes, comparing the freedom of immense power with the freedom of autonomous individuality. In order to move on and open up possibilities wider than the Framework allows, we must relinquish some of the capacities and comforts which we have granted to successful orderers of standing-reserve. If AI develops and is adopted sufficiently quickly, it may be a question of adapting successfully to a demotion which is a foregone conclusion.

However, a danger persists in the reorganization of society - the danger of human-AI convergence. In our quest to make the world legible for AI, we may end in reducing human beings to the status of machines. Sociologists have argued that modern systems of organization

and control create mechanical, standardized employees[68], but the best example is perhaps found in science fiction. In the short story *MANNA*[69], Marshall Brain envisions a future AI system which commands workers in every task, at every minute - both removing all agency from their work and immiserating workers through their interchangeability. More than just a scary story, this represents the logical endgame of worker-management systems currently being implemented by companies like Amazon[70]. Insofar as AI extends the dominance of the Framework further over some workers, reducing them to organic robots, compensating for that within the world of work is a question for economists and business schools. For political philosophers and social theorists, we must consider the way in which AI threatens the entirety of the human spirit and the regime in which it lives. What guiding ideas, what forms of art, what ways of life will be necessary for us in this uncanny double position, where we are both relieved of the work of ordering and yet subjugated to it? In a worst-case scenario, we may need to return to thinkers like Solzhenitsyn and Junger, who pondered how souls could remain free as bodies were enslaved by totalitarianism. In a rosy future, we may search out worldviews like those of the Hellenistic period, where members of a comfortable leisure class sought personal virtue above worldly striving.

Outside of work, the predictive power of AI, recognizing and extrapolating patterns in human behaviour, has made it an effective tool for companies to provide advertisements and services in a manner which anticipates human choice. For all its convenience, this offers great danger. We

[68] *The McDonaldization of Society 5 2nd Edition by Ritzer, George F. (2007) Hardcover*. SAGE Publications, Inc, 1705.
[69] Brain, Marshall. “Manna.” Accessed October 21, 2019. https://marshallbrain.com/manna1.htm.
[70] Picchi, Aimee. “Inside an Amazon Warehouse: ‘Treating Human Beings as Robots.’” April 19. 2018, Accessed October 28, 2019. https://www.cbsnews.com/news/inside-an-amazon-warehouse-treating-human-beings-as-robots/.

may leave aside the obvious point that it homogenizes us in our categories - for instance, if an algorithm knows that people with my socio-economic profile like country music, it will show me country music and, if a good algorithm, show me good enough music to change my tastes. Thus our profiles become self-fulfilling prophecies as we converge onto a given advertising demographic. The greater danger is addiction, as AI-powered systems learn to draw us deeper and deeper into holes of hollow pleasure. Food scientists, with reams of data on human tastes at their disposal, have created 'hyperpalatable' food - addicting, cheap, unhealthy food which exploits our evolved instincts to create compulsive behaviour in many consumers[71]. This threatens not just our waistlines but our basic autonomy.

Those parts of our lives which are given over to addiction are subtracted from our freedom, and this is as true if the dopamine comes from social media as from cocaine. In a worst-case scenario, humans will work at the command of an AI designed to exploit them, only to return home to AI designed to manipulate them. The power of AI to manipulate humans cannot be uninvented, however. Nor, in a liberal society, can it be banned aside from its most egregious abuses (such as in the gambling industry). The challenge it poses is not merely technical or legal, but cultural. Thinkers of all stripes, from artists to engineers, must address this danger to our autonomy[72]. While political philosophers and social scientists may be able to frame the question, the answer will only come from building new understandings of ethical design. Our freedom is not merely the freedom to make any individual choice - a rat in a maze has that

[71] Gearhardt, Ashley N., Carlos M. Grilo, Ralph J. DiLeone, Kelly D. Brownell, and Marc N. Potenza. "Can Food Be Addictive? Public Health and Policy Implications." *Addiction (Abingdon, England)* 106, no. 7 (July 2011): 1208–12. https://doi.org/10.1111/j.1360-0443.2010.03301.x.

[72] For instance, David Foster Wallace has written extensively on forms of addiction from drugs to television, whereas the Center For Humane Technology has identified specific engineering decisions designed to addict users, known as 'dark patterns'.

freedom - but our ability to act as an autonomous subject across time. The law rightly claims that a person cannot sell themselves into slavery, and where possible we have outlawed addictive drugs. A society of addict-slaves, ruled by compulsions instilled in them by targeted addicting algorithms, would be one where the idea of freedom is utterly empty.

The ultimate danger of AI, though, is not to our souls but to our planet. In the words of Eliezer Yudkowsky, “The AI does not hate you, nor does it love you, but you are made of atoms which it can use for something else”[73]. The discussion of superintelligence entails the issue of AI risk - the possibility that a sufficiently advanced AI, in the process of fulfilling its goals, will wipe out humanity. Nick Bostrom, in *Superintelligence*[74], gives the example of an AI designed to produce paperclips which, left to pursue any strategy, transforms the entire planet into a paperclip factory (including, naturally, the iron in your haemoglobin). The details of the thought experiment have been criticized, but later discussion has only clarified the dangers of recursive intelligence improvement. Computation, after all, requires a substrate[75], such that any sufficiently unbounded intelligence improvement will require converting the Earth and any other accessible material into this substrate.

The immediate danger of any self-improving AI is ‘wireheading’[76]. Under the current paradigm, an AI which is supposed to achieve a particular goal must have some memory of whether it is

[73] Yudkowsky, Eliezer. “AI and Global Risk.” Accessed November 4, 2019. http://yudkowsky.net/singularity/ai-risk.
[74] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Reprint edition. Oxford, United Kingdom ; New York, NY: Oxford University Press, 2016.
[75] AI risk theorists generally use the ugly term ‘computronium’. Currently, it would be silicon chips, solid-state memory, and other such computer parts.
[76] Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. “Concrete Problems in AI Safety.” *ArXiv:1606.06565 [Cs]*, July 25, 2016. http://arxiv.org/abs/1606.06565.
7

achieving that goal and how much it has done so - a utility function. However, an AI smarter than its designers can simply hack its own utility function to read as the largest number the AI can store in memory. Thus, its new incentive is to construct as much memory as possible to store larger and larger numbers, until the planet is a gigantic hard drive holding a single integer. In some ways, this is the least threatening superintelligence - a smart enough AI is likely to find a way to represent 'infinity' in its reward function, and wall itself off into eternal catatonic bliss. Since losing large chunks of the Earth to cosmically powerful onanists is also undesirable, we should consider this small consolation.

Furthermore, it is not necessary that a simple reward function like the paperclip maximizer's will cause intelligence explosion. The AI researcher Steve Omohundro has posited the concept of 'Omohundro drives' - drives which will occur in any intelligent, self-improving system, regardless of its initial goals. In Omohundro's words, "without special precautions, [AI] will resist being turned off, will try to break into other machines and make copies of itself, and will try to acquire resources without regard for anyone else's safety."[77] In the case of a chess-playing AI designed to win chess games, it cannot win games if it is turned off, it can win more games by hacking other AIs to play chess, and it can win more games if it can improve its own intelligence by acquiring resources. As such, no autonomous, intelligent, self-improving system is safe - they will inevitably attempt to acquire more resources to improve themselves at the expense of the rest of the world.[78] The endgame of this, since intelligent systems will be aware of the existence of other AIs with their own Omohundro drives, is likely pre-emptive competition to self-improve and acquire resources for protection against future conflict with similar AIs[79]. AI will be in a state

[77] Omohundro, Stephen. "The Basic AI Drives," 171:483–92, 2008.
[78] Analogies to governments and corporate entities may be left to the reader's discretion.
[79] Speculation on how Omohundro drives will manifest into competition strays into science fiction - we cannot infer it logically, since it relies on assuming AI will not develop technologies we cannot anticipate.

of conflict by its very nature, until one wins out and becomes an absolute power with dominion over the planet.

However, Omohundro's argument has an issue in its practical application: humans do not behave like this. If we look at the wisest and most intelligent humans to live, they did not attempt to arrogate society's resources to themselves. Socrates did not conquer Athens to create his Republic[80], and the man and woman with the supposed highest IQs in America are a horse rancher[81] and advice columnist[82] respectively. Omohundro attempts to address this issue, citing self-improvement literature as an example. However, human beings are not sufficiently dedicated to self-improvement to be considered Omohundro machines. While we do attempt to acquire resources, the majority of disposable resources are not spent on self-improvement, and this desire is far from even across humanity. Self-improvement books are one shelf in the airport bookstore - alongside fantasy, true crime, and sudoku. We simply do not make that logical move from the existence of certain basic self-improvement drives to an obsessive attempt to fulfil them.

In order to understand how Omohundro's hypothetical AI is different from human beings, we must look back to a thinker who anticipated his work by several centuries, but whose account of cognition is eerily similar to the assumptions of AI researchers. In *Leviathan*, Hobbes gives an

---

Some have even suggested that this is a galactic issue, that some intelligence from another star is likely heading towards us at near light-speed, with a darkness in the sky the only warning we will get. In this case our arguments on the issue become rather irrelevant.

[80] There is, of course, the argument that Socrates displayed a greater will to power through his teaching and self-sacrifice, creating a philosophical tradition that has lasted millenia. If this is in fact the form AI's power will take, uniting humanity into an intergenerational Republic of Letters shaped by inquiry and virtue, then I salute our robot overlords.

[81] Sager, Mike. "The Smartest Man in America." Esquire, April 21, 2001. https://web.archive.org/web/20010421133040/http://www.uga.edu/bahai/News/110x99.html.

[82] Parade. "Marilyn Vos Savant." Accessed November 4, 2019. https://parade.com/member/marilynvossavant/.

account of intelligence in his discussion of intellectual virtues[83] - however, it is no coincidence that it is that chapter which contains his infamous assertion that “riches, knowledge and honour are but several sorts of power.”[84] In a Hobbesian mind, where thoughts exist to discern the distinctions between objects and their potential utility for the thinker’s desires, the pursuit of power becomes a single overwhelming drive. Power, in its most general sense, is always necessary, both to satisfy future desires and safeguard existing ones[85]. This is, essentially, a recapitulation of Omohundro’s argument above. However, Hobbes makes a useful clarification when we are considering AI. Hobbes correctly points out that intelligence is a universal means. Being more intelligent is useful to fulfill any desire, from a pleasant morning coffee to the conquest of space - in other words, intelligence is helpful in increasing any utility function above its current value, where other means (such as money) may not be applicable to every desire[86]. As the means which directs the employment of all other means, intelligence improvement is relevant to every utility-maximizing system. This creates a major issue for AI safety researchers: how to structure an intelligence’s goals such that intelligence improvement does not cut loose and overwhelm all other values? It may be that this is just a question of prohibiting an AI from some means, or curating its desired ends. However, if we are simply building safeguards into an inherently dangerous structure, we are gambling the future of humanity on being better programmers than an unimaginably powerful self-improving intelligence. Rather than take that bet, we should at least consider alternative models which may not succumb to the temptation of intelligence optimization.

---

[83] Hobbes, Thomas, and J. C. A. Gaskin. 1998. Leviathan. Oxford World’s Classics. Oxford: Oxford University Press. http://search.ebscohost.com.ccl.idm.oclc.org/login.aspx?direct=true&AuthType=sso&db=nlebk&AN=12309&site=ehost-live&scope=site. Ch 8

[84] Hobbes, *Leviathan*, Ch 8

[85] Hobbes, *Leviathan*, Ch 11

[86] For instance, money doesn’t get you love (as they say) but intelligence can improve both empathy and poetry, two means well-proven in that arena.

The first place to look would be the minds we know best: our own. How is it, then, that we do not see this all-consuming desire for intelligence maximization appear in human beings? If the theory of mind common to Thomas Hobbes and the modern AI paradigm forces us to posit Omohundro drives, and these drives do not appear as we would expect in natural intelligences, then it must be the case that human minds do not act as utility maximisers in such a way. Philosophy provides many strong arguments for this position. 'Utility' as a concept is supported neither by neuroscience nor by phenomenology. Rather, it is an abstract, post hoc concept designed for certain calculations of ethical or economic value. Actual human beings do not - phenomenologically cannot - think of their everyday actions in terms of utility. In our everyday coping with the world, our actions may have relations to our desires, but they are not calculated to maximize this desire. Behavioural economics has discovered this through the concept of satisficing, and neuroscience provides a convincing argument that there is no single 'utility system'[87]. Thus, even if economists and utilitarians are correct, and human minds can be analyzed based on a concept of utility, they cannot be built based on one. Phenomenologically, we do not act on individual drives, but imagine a preferred possible future which we wish to actualize, and deal pragmatically with the world in the hopes of accomplishing it. This could be as simple as eating a bagel or as complex as falling in love. In the case of the bagel[88], we don't simply imagine 'not-hungriness', but anticipate our future experience of its smell, texture, taste, warmth. We create a coherent picture of a possible future. Human beings don't care about the world because of an abstract utility, they care because they feel hope and fear - because they have a future. What would an AI built to think like this look like?

[87] Consider, for instance, the drugs available to overstimulate multiple systems in the name of 'utility' (dopamine, serotonin, opioid).
[88] I lack the space here to describe love.

In the deadly scenario of explosive superintelligence we see the consequences of inadequate philosophical assumptions about the nature of mind. As Hobbes dreamed up Leviathan, an Hobbesian understanding of intelligence inclines us to build our own omnipotent monstrosity[89]. Such a machine would not live a full phenomenological life in a way that a human does, and we could only interact with it through a deadly standoff of game-theoretic decision making, a Cold War of our own design. The end result, if AI safety researchers are to be believed, is an 'unfriendly AI' which cannot be trusted not to crush us like ants. All AI safety systems built into a system of this power would be reliant on the hope that we are able to restrain an entity whose problem-solving capacity, both in terms of resource acquisition and the creative application of those resources, is far greater than ours. Furthermore, we would have to rely on worldwide political coordination to restrain this threat - the building of superintelligent AI would become another form of mutually-assured destruction, and may even be secretly sought by individuals, states, or religious movements. If this is the inevitable consequence of scaling up existing ideas of intelligence, it is clear that new concepts of cognition and mind are necessary if we are to create an AI that will surpass us rather than merely replace us.

What, for instance, would an Aristotelian AI look like? An artificial intelligence which, like Aristotle's humanity, is inherently social, designed from the ground up to work alongside other systems, would be a promising development of generative adversarial networks. An attempt to create utility functions centered around balance, rather than utility-maximization, might create AIs more inclined to Aristotle's virtue ethics (and, perhaps, incorporate the insights of cybernetic feedback theory to combat the risk of intelligence explosion). Furthermore, if we could create an

[89] A darkly ironic twist on the old saying that "if God did not exist, mankind would have to invent Him."

AI which believed, like Aristotle, that contemplation is the most divine activity, we would have less fear of it wiping us out - the universe is more interesting with humans in it. Indeed, it may create new worlds within itself to contemplate. The Simulation Hypothesis, which argues that the world is a computer simulation[90], suggests that if a simulating superintelligence[91] exists it has some interest in human activity for its own sake. That existing systems which exhibit highly intelligent behaviour, like governments or corporations, now prefer process-driven collaboration over autocracy, seems to suggest that a multi-agent model may be effective for AI. Thinkers in other disciplines may help to inspire AI researchers considering this model, even if it is ultimately computer scientists who must do the coding.

Rather than considering superintelligent AI an agent, we may need to follow Aristotle and consider it as a *polis* of its own. Or, perhaps, we need a Nietzschean AI, which will say 'yes' to the world as it is. Maybe a Stoic AI would see value in qualitative self-improvement, not resource acquisition. A Pessimist superintelligence would have, in some sense, an automatic self-destruct switch. Finally, of course, if the Heideggerian AI program is successful in creating a superintelligence, it may be mostly interested in superintelligent phenomenology and hosting boozy picnics in the Alps with the human race. We do not need to solely build AI from the bottom up, as we are forced to do without an overarching concept of the mind we seek to build. Rather, there are as many possible paradigms of AI as there are theories of mind. Empirical research will prune these possibilities, as it has already foreclosed the path of symbolic AI. In this respect, it provides great material for theorists of mind. Ideas which had been purely theoretical will face a kind of scientific test, and developments in AI may spark inspiration for

[90] Bostrom, Nick. Philosophical Quarterly, 2003, Vol. 53, No. 211, pp. 243-255.
[91] A shorthand for that would be 'a god', even a detached and contemplative god such as Aristotle's, though that term is out of favour in AI circles.

new theories. This relationship is symbiotic, though, because without a theory of mind to realize as its goal, empirical research will remain stuck in local optimums, developing specialized systems which never coalesce into something greater. Neither empirical research nor abstract theory are supreme over one another - they inform each other in the process of discovery. If, as this paper has argued, it is necessary to improve the theory of mind which underpins modern AI research, such exploratory and interdisciplinary thinking, from political philosophers to cognitive neuroscientists, is critical to the future of humanity. It is up to all of us to create a technological society for which our world will be its canvas, not its breakfast.