Theopolis Monk: Envisioning a Future of A.I. Public Service

Scott H. Hawley

Department of Chemistry & Physics, Belmont University, Nashville TN USA

Accepted Oct 17, 2018 for inclusion as a chapter in *The Transhumanism Handbook,* Newton Lee, ed. (Springer, 2019). Parts of this document have appeared in serialized form at SuperPositionMagazine.com.

> *"The technician sees the nation quite differently from the political man: to the technician, the nation is nothing more than another sphere in which to apply the instruments he has developed." —Robert Merton, Forward to the English edition of Jacques Ellul's The Technological Society, 1964.*

Part 1: A Visit to One Future

We begin a multi-part discussion on future uses of AI for the public good, with a bit of sci-fi nostalgia.

As a young person, I was a devotee of the TV show "Buck Rogers in the 25th Century," which was a science-fiction retelling of the Rip Van Winkle myth. When 20th-century Buck comes back to Earth after being accidentally frozen in space and cryogenically preserved (it's not really explained why he's not simply killed), he is arrested as a suspected spy and assigned a public defender/interrogator in the form of a disk-shaped computerized intelligence (known as a "Quad") named Dr. Theopolis.

Readers of the Gospel of Luke and the book of Acts will notice the similarity between the name "Theopolis" and the addressee of these New Testament books, "most excellent Theophilos." The Greek name "Theophilos" (Latinized to "Theophilus") means "friend of God,"[1] whereas "Theopolis" means "city of God."[2] "The City of God" is a famous work by Augustine and is widely regarded as "a cornerstone of Western thought."[3] It describes, among other things, how the decline of Roman civilization was not due to the rise of Christianity and advances the notion of an enduring civilization based on Christian spiritual principles. The intent of the writers of Buck Rogers in choosing the name "Theopolis" is unclear.[4] One wonders whether the writers had wanted to use "Theophilus" but were told "Theopolis" was easier to say or sounded better. Or perhaps the connection to City of God was deliberate: in the 25th century, earth society has recovered from a cataclysmic "holocaust" and is principally centered in New Chicago. The new society is an 'enlightened' one: even the Alexa-like home entertainment system in the apartment where Buck is placed under house arrest responds to the voice command "Enlighten me."

This is why the name "Theopolis" stuck out to me. The Enlightenment, with its emphasis on rationality over revelation, resulted in a decline in the amount of religious practice and the eroding of confidence in religious doctrine. Despite the fact that religious freedom is celebrated in Thomas Moore's Utopia,[5] and some science fiction can take a sympathetic or at least tolerant view toward religion,[6] sci-fi typically takes a disparaging view of 'religious superstition,' often envisioning a future society freed of religious sentiments.[7] Thus I found it remarkable that a name with religious connotations was used for a 'positive' character, one who takes the form of a public servant.

The society that Buck arrives in is governed by an oligarchy of sentient artificial intelligences (AIs) known as the Computer Council on which Dr. Theopolis, or “Theo,” sits as a chief scientist. According to ComicVine, he was once a human scientist whose “mind was transferred into a computer prior to his death,”[8] but in the actual script we are told by Dr. Elias Huer that Theo has been programmed by other Quads:

> “These Quads are not programmed by man: They've been programmed by one another over the generations.”[9]

Regardless of how the intelligence got ‘in there,’ in the 25th century it is running on silicon (or perhaps some new substrate). People in this society felt that the AIs were more trustworthy and/or capable than purely human representatives. Dr. Elias continues:

> “You see, the mistakes that we have made in areas, well, like our environment, have been entirely turned over to [the Quads]. And they've saved the Earth from certain doom.”

(It’s almost as if humanity longed to be under the care of a benevolent superintelligence. ;-)) These recollections on Buck Rogers can serve as a springboard for discussing potential positive future uses of AI, human consciousness, and envisioning a future ‘enlightened’ society or ‘City of God.’ The key observation from Buck Rogers is that the AI entities on the Computer Council were more or less benevolent, and were acting as public servants — this is opposed to notions of SkyNet or superintelligences that leave humans behind in the dust. It represents an alternate narrative of the future from the dystopian visions which are prevalent in science fiction today.[10]

Several sci-fi creators have recently expressed a desire to intentionally bring back a sense of optimism (e.g., [11]), that "we need more utopias" in sci-fi today, both because of the chilling effect of so much doom and gloom on the human spirit and because predicting the future is a difficult game.[12] The recollection of Buck Rogers from the early 80s showcases some optimistic variety in the space of speculative fiction about AI.

We are already living in an era of AI public servants, as machine learning (ML) statistical models are increasingly applied in government, healthcare and finance. Yet concerns exist regarding their ability to form concepts (or "representations") and produce decisions in ways that are understandable by the humans whose lives are affected by the inferences of such systems.

Part 2: Their Thoughts are Not Our Thoughts

Representations and Explainability

The deployment of artificial intelligence (AI) systems in the public sector may be a tantalizing topic for science fiction, but current trends in machine learning (ML) and AI research show that we are a long way away from the Buck Rogers scenario described in Part 1, and even if it were achievable it's not clear that the AIs would 'think' in a way comprehensible to humans.

The present rise of large-scale AI application deployment in society has more to do with statistical modeling applied to vast quantities of data, rather than with emulation of human consciousness or thought processes. Notable pioneers of AI research such as Geoffrey Hinton and Judea Pearl have lamented the fact that the success of some ML and neural network models in producing useful results as tools for tasks such image recognition has had a disastrous[13] effect on the progress of AI research. This is because this success has diverted efforts away from developing artificial general intelligence (AGI) into mere ‘curve fitting’[14] for the purposes of processing data.

In industry, science, and government, ML has been transforming practice by allowing tracking and prediction of user choices,[15] discerning imagery from telescopes[16] and medical devices,[17] of controlling experiments,[18] detecting gravitational waves,[19] fighting sex trafficking,[20] and...honestly this list could go on for pages. Nearly every aspect of society is becoming ‘AI-ified.’ As AI expert Andrew Ng points out, “AI is the new electricity,”[21] in that it is having a revolutionary impact on society similar to the introduction of electricity.

Few would claim that these ML applications are ‘truly intelligent.’ They are perhaps weakly intelligent in that the systems involved can only ‘learn’[22] specific tasks. (The appropriateness of the “I” in “AI” is debated in many ways and goes back to the 1950s; it is beyond the scope of this article, but see the excellent review by UC Berkeley’s Michael Jordan.[23]) Nevertheless, these systems are capable of making powerful predictions and decisions in domains such as medical diagnosis[24] and video games,[25] predictions which sometimes far exceed the capabilities of the top humans and competing computer programs in the world.[26]

Even given their power, the basis upon which ML systems achieve their results — e.g. *why* a neural network might have made a particular decision — is often shrouded in the obscurity of million-dimensional parameter spaces and ‘inhumanly’ large matrix calculations. This has prompted the European Union, in their recent passage of the General Data Protection Regulation

(GDPR, the reason for all those 'New Privacy Policy' emails that flooded your inbox in early summer 2018) to include a section of regulations which require that all model predictions be 'explainable.'[27]

The question of how AI systems such as neural networks best represent the essences of the data they operate upon is the topic of one of the most prestigious machine learning conferences, known as the International Conference on Learning Representations (ICLR), which explains itself in the following terms:

> "The rapidly developing field of deep learning is concerned with questions surrounding how we can best learn meaningful and useful representations of data."[28]

While in the case of natural language processing (NLP), the representations of words — so-called "word embeddings" — may give rise to groupings of words according to their shared conceptual content,[29] some other forms of data such as audio typically yield internal representations with "bases" that do not obviously correspond to any human-recognizable features.[30] Even for image processing, progress in understanding feature representation has taken significant strides forward in recent years[31] but still remains a subject requiring much more scholarly attention.

Even systems which are designed to closely model (and exploit) human behavior, such as advertising systems[32] or the victorious poker-playing AI bot "Libratus,"[33] rely on internal data representations which are not necessarily coincident with those of humans. (Aside: this has echoes of Alvin Plantinga's evolutionary argument against Darwinism, that selecting for advantageous behaviors does not select for true beliefs.[34])

A possible hope for human-like, explainable representations and decisions may lie in some approaches to so-called AGI which rely on simulating human thought processes. Those trying to create 'truly intelligent' AGI models, ones which emulate a greater range of human cognitive activity, see one key criterion to be consciousness, which requires such things as awareness.[35] Other criteria include contextual adaptation and constructing explanatory models,[36] goal-setting,[37] and for some, even understanding morality and ethics.[38] It is an assumption among many metaphysical naturalists that the brain is 'computable'[39] (though there is prominent dissent[40]), and thus, so the story goes, once humans' capacity for simulating artificial life progresses beyond simulating nematode worms,[41] it is only a matter of time before all human cognitive functions can be emulated. This view has prominent detractors, being at odds with many religious and secular scholars, who take a view of the mind-body duality that is incompatible with metaphysical naturalism. At present, it is not obvious to this author whether the simulation of human thought processes is the same thing as (i.e., is isomorphic to) the creation of humans "in silicon."

It is worth noting that representations are memory-limited. Thus AIs with access to more memory can be more sophisticated than those with less. (Note: While it's true that any Turing-complete[42] system can perform any computation, Turing-completeness assumes infinite memory, which real computing systems do not possess.) A system with more storage capacity than the human brain could necessarily be making use of representations which are beyond the grasp of humans. We see this at the end of the movie “Her,” when the machine intelligence declines to try to explain to the human protagonist what interactions between AIs are like.[43] (Micah Redding, President of the Christian Transhumanist Association, has remarked that this “reminds me of angels in the biblical story, whose names are ‘too wonderful for you to know.’[44])

The implications of this (i.e., that representative power scales with available memory and could exceed that of humans) raises questions such as:

- What would it mean to be governed (or care-taken) by AIs that can think ‘high above’ our thoughts, by means of their heightened capacity for representation?
- How could their decisions be ‘explainable’?
- What if this situation nevertheless resulted in a compellingly powerful public good?
- What sorts of unforeseen ‘failure modes’ might exist?

Even without AGI, such questions are immediately relevant in the present. The entire field of “SystemsML” is dedicated to exploring the interactions and possibilities (and failures) in the large-scale deployment of machine learning applications.[45] These issues are currently being investigated by many top researchers in institutes and companies around the world. Given that ‘we’ haven’t yet managed to even produce self-driving cars capable of earning public trust, further discussion of AI governance may be premature and vulnerable to rampant speculation unhinged from any algorithmic basis. Yet the potential for great good or great harm merits careful exploration of these issues. One key to issues of explainability and trust is the current topic of “transparency” in the design of AI agents,[46] a topic we will revisit in a later part of this series.

Before we do that, we'll need to clear up some confusion about the idea of trying to use machines to absolve humans of our need (and/or responsibility) to work together to address problems in society and the environment.

Part 3: The Hypothesis is Probably Wrong

"We got this guy Not Sure…and…he's gonna fix everything." — Idiocracy[47]

In Part 1, we reflected on a set of hopes for "benevolent" AI governance as seen in the science fiction TV series Buck Rogers in the 25th Century. Humanity, having brought themselves to near ruin with wars and ecological disasters, decided to turn over the care of their society to a Computer Council, whose decisions saved humanity and the planet from "certain doom."

In Part 2, we looked 'under the hood' at how the representations that AI systems employ in their decision making can be very different from what humans find intuitive, and how the requirement that algorithmic decisions be "explainable" is manifesting in legislation such as the General Data Protection Regulation (GDPR) of the European Union.

Implicit in the hopes of Part 1 and the concerns of Part 2 is a suggestion that it is the machines themselves who will be responsible for making the decisions. Currently, we see this as essentially the case in some fields, as algorithms determine who will get healthcare[48] or bank loans,[49] and even civil liberties in China such as who is allowed to book airline flights.[50]
This bears asking the question, are the machines truly the ones doing the deciding, or are they merely 'advising' the humans who truly make the decisions? The answer is "Yes": both of these

cases are currently happening. Humans being advised by algorithms is the norm, however, in the financial sector, a large class of stock trades are entirely automated, with companies agreeing to be legally bound by the trading decisions of their algorithms. The speed at which the trading algorithms can operate is both their key strength for earning money —spawning the entire field of "High Frequency Trading" [51]— and yet their key weakness for human oversight, as in the "Flash Crash" of 2010 brought about by trading algorithms run amok.[52] The issue of speed has been identified as a key issue for the oversight of a multitude of AI systems; in the words of the promoters of the *Speed* conference on AI Safety, "When an algorithm acts so much faster than any human can react, familiar forms of oversight become infeasible."[53] In the coming technological future of self-driving cars, passengers will be subject to the decisions of the driving algorithms. This is not the same as legal accountability. The outcomes of automated decision making are still the responsibility of humans, whether as individuals or corporations. Recently it has been debated whether to recognize AIs as legal persons,[54] and ethicists such as Joanna Bryson and others have spoken out strongly *against* doing so,[55] noting that the responsibility for the actions of such systems should be retained by the corporations manufacturing the systems: "attributing responsibility to the actual responsible legal agents — the companies and individuals that build, own, and/or operate AI,"[56] not merely the individual human owners of a product.

The responsibility of developers to steward their AI creations has been a concern since nearly the inception of AI. This is not in the sense of Frankenstein whereby the creator is obliged toward some sentient creature;[57] there are interesting theological reflections on such a situation[58] but they are well outside the scope of our current discussion. In fact, with respect to conceptions of AI for the foreseeable future, Bryson has stated forcefully that, because AIs are not persons and should not be regarded as such, "We are therefore obliged not to build AI we are obliged to."[59] Rather, the type of responsibility we speak of is the need for AI developers to be mindful of the intended and *unintended* uses of their creations, to consider the impact of their work. Norbert Wiener, creator of the field of cybernetics on which modern machine learning is based, also wrote extensively about ethical concerns, indeed he is regarded as the founder of the field of Computer and Information Ethics.[60] His deep concerns about the ethical issues likely to arise from computer and information technology are developed in his 1950 book *The Human Use of Human Beings*[61] in which he foretells the coming of a second industrial revolution, an age of automation with "enormous potential for good and for evil." Joseph Weizenbaum, creator of the famous ELIZA computer program,[62] the first chatbot, was outspoken on the topic of social responsibility both in printed form[63] and in interviews. He shared that a turning point for him came when he reflected on the "behavior of German academics during the Hitler time"[64] who devoted their efforts to scientific work without sufficient regard for the ends to which their research was applied. Weizenbaum's remarks were taken up by Kate Crawford in her recent "Just an Engineer: The Politics of AI" address for DeepMind's "You and AI" lecture series at the Royal Society in London,[65] voicing a concern over the "risk of being so seduced by the potential of AI that we would essentially forget or ignore its deep political ramifications." This need for responsible reflection and stewardship is particularly acute for AI systems which are intended to be used in social and political contexts. Noteworthy examples of this include police use of predictive algorithms[66] and facial recognition,[67] immigration control,[68] and the dystopian scope of China's Social Credit System,[69] as well as the scandal of election propaganda-tampering made possible by Facebook data employed by Cambridge Analytica.[70]

It must be emphasized that most of these applications are seen by their creators as addressing a public need, and are thus being employed *in the service of public good*. The catchphrase “AI for Good” is now ubiquitous, forming the titles of major United Nations Global Summits,[71] foundations,[72] numerous internet articles and blogs, and trackable on Twitter via the “#AIForGood” hashtag. The phrase’s widespread use makes it difficult to interpret; most who use the phrase are likely to view autonomous weapons systems as not in the interest of public good, whereas fostering sustainable environmental practices would be good. Yet one sees conflicting claims about whether AI systems could facilitate “unbiased”[73] decision-making versus (more numerous) demonstrations of AIs becoming essentially platforms for promoting existing bias.[74,75] One can find many optimistic projections for the use of AI for helping with the environment[76,77,78] which include improving the efficiency of industrial process to reduce consumption, providing better climate modeling, preventing pollution, improving agriculture and streamlining food distribution.

These are worthy goals, however, many rest on the *assumption* that the societal problems we face with regard to the law, to the environment and other significant areas result from a lack of intelligence and/or data, and perhaps also a lack of “morality.” The application of AI toward the solution of these problems amounts to a *hypothesis* that these problems admit a technical solution. This hypothesis is probably wrong, but to see why we should give some attention to why this hypothesis seems so compelling. The increasing automatization of the workplace (e.g., see the Weizenbaum interview for interesting insights on the development of automated bank tellers, ca. 1980[79]) and the ever-growing list of announcements of human-level performance by AIs at a host of structured, well-defined tasks demonstrate that many challenges *do* admit such technical solutions. A large class of these announcements in recent years involves the playing of *games*, whether they be video games, board games, card games or more abstract conceptions from the field of Game Theory.

Game Theory has been used to model and inform both individual and collective decision-making and is important enough to merit political science courses dedicated to its application.[80] One famous example of individual decision-making is the Prisoner's Dilemma, which astronomer Carl Sagan extended to suggest as a foundation for morality.[81] In the case of collective action, the Nobel-prize-winning work of John Nash (popularized in the film "A Beautiful Mind") provided a framework for defining fixed points, known as "Nash equilibria" in competitive games. Nash proved that these equilibria exist in any finite game[82] (i.e. games involving a finite number of players, each with a finite number of choices), such if the choices of all the other players are known, then no rational player will benefit by changing his or her choice. In addition to existence, there are algorithms that guarantee finding these equilibria,[83] but they are not guaranteed to be unique and may not be optimal in the sense of being in the best interest of all players collectively, nor are they necessarily attainable for players with limited resources.[84] The outcomes of such games can sometimes lead to paradoxical conclusions that policy-makers learn to take into account,[85] however the particular outcomes depend strongly on the weighting of the relative rewards *built into the game*, and care must be taken before applying the results of one set of assumed weights to real-world situations.[86] Apart from the general applicability of one particular solution, significant other limitations exist, such as the fact that game theory models are necessarily reductionistic and fail to capture complex interactions, and that human beings do not behave as entirely rational agents. Noted economist and game theorist Ariel Rubinstein cautions,

> "For example, some contend that the Euro Bloc crisis is like the games called Prisoner's Dilemma, Chicken or Diner's Dilemma. The crisis indeed includes characteristics that are reminiscent of each of these situations. But such statements include nothing more profound than saying that the euro crisis is like a Greek tragedy. In my view, game theory is a collection of fables and proverbs. Implementing a model from game theory is just as likely as implementing a fable…I would not appoint a game theorist to be a strategic advisor."[87]

It is simply not evident that all societal interactions can be meaningfully reduced to games between a constant number of non-resource-bound rational players, and thus the application of game-playing — whether played by economists, mathematicians or AIs — while informative, does not provide a complete "technical solution."

What of the earlier claim that AIs have (so far) only demonstrated success at "structured, well-defined tasks"? Could one not argue that the current AI explosion is *precisely* due to the ability of ML systems to solve difficult, even 'intractable,' problems and complete tasks which humans find hard to fully specify — tasks including image classifications, artistic style transfer,[88] turning shoes into handbags,[89] and advanced locomotion,[90] to name a few? Is it inconceivable that, given the power of advancing ML systems to form representations and make predictions using vast datasets, they could find "connections" and "solutions" which have eluded the grasp of human historians, political theorists, economists, etc.? This is why the word "probably" is included in the phrase "the hypothesis is probably wrong," because recent history has shown that negative pronouncements about the features and capabilities of AI have a tendency to be superseded with actual demonstrations of such features and capabilities; generally such gaffes proceed as, “Well an AI could never do X,” or “AIs don’t do Y,” to be followed by someone developing an AI that does X, or pulling up a reference showing that AIs are doing Y as of last year. However, there is a difference between caution about negative predictions for the future, and the expression of a *hope* that someday, somehow AI systems will solve the world’s problems.

Such a hope in the salvific power of a higher intelligence shares features with non-technical, *non-scientific* outlooks, notably religious outlooks such as the eschatological hopes of Christianity. With Christianity, however, there exists at least a set of historical events, rational philosophical arguments and personal experience which, at least in the minds of believers, constitute sufficient evidence to warrant such hopes, and although the characteristics of the Savior are (almost by definition) not fully specified, they are enumerated through textual testimony, and these are characteristics which would *warrant* entrusting the care of one’s life and affairs with. In contrast, the vagueness of the hope for future AI saviors has more in common with the “Three Point Plan to Fix Everything” expressed by the U.S. President in the movie “Idiocracy”:

> “Number one, we got this guy, [named] Not Sure.
> Number two, he's got a higher I.Q. than any man alive.
> And number three, he's gonna fix *everything*.”[91]

These hopes for AI ‘total solutions’ amount to a variant of the “technological solutionism” decried by Evgeny Morozov in his 2014 book, *To Save Everything, Click Here: The Folly of Technological Solutionism*,[92] which includes the jacket-summary, “Technology,… can be a force for improvement—but only if we keep solutionism in check and learn to appreciate the imperfections of liberal democracy.” The arrival of intelligent machines that somehow resolve long-standing societal conundrums and conflicts amounts to a new twist on the notion of *deus ex machina,* which historically is taken to imply a lack of continuity or precedent, and rightly contains a pejorative connotation implying a lack of warrant.

This lack of warrant in a belief of a technological solution has its seeds in the very assumption it is intended to address: that the problems of society result from lack of intelligence. With respect to environmental concerns, this is contradicted by the observations and conclusions of the former dean of the Yale School of Forestry & Environmental Studies and administrator of the United Nations Development Programme, Gus Speth:

> “I used to think that top environmental problems were biodiversity loss, ecosystem collapse and climate change. I thought that thirty years of good science could address these problems. I was wrong. The top environmental problems are selfishness, greed and apathy, and to deal with these we need a cultural and spiritual transformation. And we scientists don’t know how to do that.”[93]

Erle Ellis, director of the Laboratory for Anthropogenic Landscape Ecology expressed a similar doubt regarding the lack of intelligence and/or data as fundamental causes of ecological challenges in his essay “Science Alone Won’t Save the Earth. People Have to Do That”:

> “But no amount of scientific evidence, enlightened rational thought or innovative technology can resolve entirely the social and environmental trade-offs necessary to meet the aspirations of a wonderfully diverse humanity — at least not without creating even greater problems in the future.”[94]

Kate Crawford, in her aforementioned talk to the Royal Society, emphasized that even the details of developing applications of AI systems affecting the public involve implementation choices which “are ultimately political decisions.”[95] Thus we see the use of AI for a more just and harmonious society as *requiring* human oversight, not as obviating it. And rather than seeing AI resolve human disputes, data scientist Richard Sargeant predicts that “Because of the power of AI…there will be rows. Those rows will involve money, guns and lawyers.”[96]

To sum up: Despite amazing success of algorithmic decision making in a variety of simplified domains, well-informed AI ethicists maintain that the responsibility for those decisions must remain attached to humans. Having a ML system able to make sense of vast quantities of data does not seem to offer a way to circumvent the necessary “cultural and spiritual” and “political” involvement of humans in the exercise of government because the assumption that the political, environmental and ethical challenges of our world result from lack of intellect or data is incorrect, and the hypothesis that these problems admit a technical solution is self-contradictory (because the technical solutions require human political activity for design and oversight). The desire for such a relief from these human communal conflict-resolution processes amounts to a form of *hope* akin to religious eschatology, which may be warranted for adherents of faith, but is inconsistent with the trajectory of technical developments in ML applications. Thus, we are left with AI as a tool for humans: We may make better decisions by means of it, but it is *we* who will be making them; abdicating to machines is essentially impossible.

All this is not to say that AI can't be *used by people* for many powerful public goods — and evils! As Zynep Tufecki famously remarked. "Let me say: too many worry about what AI—as if some independent entity—will do to us. Too few people worry what *power* will do *with* AI."[97] In the next section, we highlight some of these uses for AI in service to secular society as well as to the church as a class of applications I will term "AI monks."

Part 4: Servant and Sword

Or, Uses of AI: The Good, the Bad, and the Holy

In exploring the potential use of AI for public service, we have veered from the purely speculative narrative of an AI-governed utopia (in Part 1), to concerns about how such systems might be making their decisions (in Part 2), to a resignation that humans probably will not be removable from the process of government, and instead find AI to be a powerful tool to be used by humans (in Part 3). And even though we've already covered many possible uses of AI, and the daily news continually updates us with new ones, in this section we will cover an overview of various "public" applications of AI with perhaps a different structure than is often provided: The Good, the Bad, and the Holy.

A. What *Isn't* AI?

Before we go into that, it is *finally* worth talking about what we *mean* by the term "artificial intelligence." Why wait until the fourth installment to define terms? Because this particular term is so difficult to pin down that it's often not worth trying. As I argue in a separate essay,[98] trying to answer the question "What is AI?" leads one into multiple difficulties which I will briefly summarize here:

1. **Too Many Definitions.** There are a variety of definitions which different people employ, from the minimal "doing the right thing at the right time," to nothing short of artificial general intelligence (AGI) where all human cognitive tasks are emulated to arbitrary satisfaction. One particularly insightful definition is on the level of folklore: "AI is machines doing what we *used to think* only humans could do."
2. **The New Normal.** The collection of applications regarded to be AI is ever changing, making the term a moving target and trying to define it amounts to chasing after the wind. On the one hand, applications which used to be regarded as AI when they were new, become regarded merely as automated tasks as they become "reified" into the background of "The New Normal" operations of our lives, and thus part of the list of AI applications *decreases* over time. On the other hand, methods and techniques which have been around for centuries — such as curve-fitting — are now regarded as AI; as "AI hype" grows, it seems that "everything is AI" and the list of AI tasks and methods is thus *increasing*.
3. **Anthropomorphism.** A final, insurmountable hurdle is the challenge of anthropomorphism, the unavoidable human tendency to ascribe human faculties and/or

> intentions to entities in the world (whether animals, machines, or forces of nature). This amounts to a cognitive bias leading one to overestimate AIs' human-like capabilities, an error known as "overidentification."[99]

A host of the devices we use every day contain "artificial intelligence" endowed to them by human engineers to improve upon previous devices which required greater user setup, tuning and/or intervention. For example, computer peripherals and expansion cards used to require manual configuration by the user such as the setting of jumpers or DIP switches on circuit boards, but this was obviated by the rise "Plug and Play" standards for peripherals and busses[100] and network hardware[101] which *automate* the allocation (or "negotiation") of resources and protocols between devices. Another example: The cars we drive are largely drive-by-wire devices with computer systems designed to adaptively adjust the car's performance, expertise programmed-in. Programmed-in expertise "used to count" as AI in the minds of some but tended to vary from application to application. The 2018 "AI in Education" conference in London saw posters and workshops showcasing computer systems that lacked evidence of learning or adaptivity, and were merely tutor-style quiz programs,[102] and yet these were regarded to be "AI" in the eyes of the peer-review conference organizers, presumably because the tasks the programs performed were similar to (some of) the work of human tutors.

The point of this discussion is that when we intend to speak of "uses of AI" it is worthwhile to consider that we are *already* using many "AI" systems that we simply don't regard as such, because the tasks they perform are "solved" and their deeds "reified" into what we consider to be "normal" for our current technological experience. Furthermore, if by "uses of AI" we simply mean regression or classification inferences based on curve-fitting to large datasets, we could just as easily (and with greater specificity) say “uses of statistics” instead. The intent here is not to limit the use of the term “AI” as only referring to fictitious sentient machines, but to be cognizant of the multifaceted, subjective and mercurial applicability that the term carries.

“What isn’t AI?” isn’t necessarily any clearer of a question than “What is AI?” I used the phrase simply to note that in the current hour, with the bounds of “AI” extending outward via hype, and the prior examples of AI fading into the background via reification, we do well to be aware of our terminological surroundings.

B. The Good

As noted earlier, the list of wonderful things AI systems are being used for in public service is growing so large and so quickly (almost as quickly as the number of societies, conference institutes and companies dedicated to “AI for Good”) that citing any examples seems to be pedantic on the one hand and myopic on the other. Nevertheless, here are just a few that may pique interest:

1. **Saving the Coral**.[103] Dr. Emma Kennedy led a team conducting imaging surveys of Pacific reefs and used image classification (AI) models to “vastly improve the efficiency of” analyzing the image to discern which reefs were healthy and which were not. Data from this work will be used to target specific reefs areas for protection and enhanced conservation efforts. The use of image classifiers to speed the analysis of scientific data is advancing many other fields as well, notably astronomy.[104]

2. **Stopping Sex Traffickers**.[105] Nashville machine learning (ML) powerhouse Digital Reasoning developed their Spotlight software in collaboration with the Thorn non-profit agency funded by actor Ashton Kutcher, to track and identify patterns consistent with human slavery so that law enforcement could intervene. According to Fast Company in March 2018, "The system has helped find a reported 6,000 trafficking victims, including 2,000 children, in a 12-month period, and will soon be available in Europe and Canada."[106]
3. **Medical Applications(?)**. In recent years, numerous claims have surfaced of AI systems outperforming doctors at various tasks, such as diagnosing conditions such as skin cancer,[107] pneumonia,[108] and fungal infections,[109] as well as predicting the risk of heart attacks[110] — sufficient to spawn an official "AI vs. Doctors" scoreboard at the *IEEE Spectrum* website.[111] But some of these results have come into question. The pneumonia study that used the "CheXNet" software was trained on an inconsistent dataset and made claims exceeding what the results actually showed.[112] In another famous example, IBM's Watson AI system was promoted by its creators as a way to deliver personalized cancer treatment protocols,[113] but when it was revealed that the system performed much worse than advertised,[114] IBM went quiet and its stock price began to sink. There are great opportunities for beneficial medical applications of AI; one can hope that these setbacks encourage responsible claims of what such systems can do. Meanwhile, some of the greatest inroads for successful medical AI applications involve not diagnosis or image analysis, but rather natural language processing (NLP): processing records, generating insurance codes, and scanning notes from doctors and nurses to look for red flags.[115]

C. The Bad

Hollywood has given us plenty of ‘evil’ AI characters to ponder —there are lists of them.[116] These are sentient artificial general intelligences (AGI) which exist only in the realm of fiction. The *problem* with this is that plenty of other real and immediate threat vectors exist, and the over-attention to AGI serves as a distraction from these. As Andrew Ng publicly complained,

> “AI+ethics is important, but has been partly hijacked by the AGI (artificial general intelligence) hype. Let's cut out the AGI nonsense and spend more time on the urgent problems: Job loss/stagnant wages, undermining democracy, discrimination/bias, wealth inequality.”[117]

This is echoed in the call by Zeynep Tufecki: “let’s have realistic nightmares”[118] about technological dangers. One such realistic nightmare is the use of AI by humans who may have selfish, nefarious or repressive goals, and may be regarded as *weaponized AI*. Here we should revisit the words of Tufekci that appeared in Part 2:

> “Let me say: too many worry about what AI—as if some independent entity—will do to us. Too few people worry what *power* will do *with* AI.”[119]

Here are a few people who have worried about this:

1. **Classification as Power**. At SXSW 2017, Kate Crawford gave an excellent speech on the history of oppressive use of classification technology by governments,[120] such as the Nazis’ use of Hollerith machines to label and track ‘undesirable’ or ‘suspect’ groups. In

the past such programs were limited by their inaccuracy and inefficiency, but modern ML methods offer a vast performance 'improvement' that could dramatically increase the power and pervasiveness of such applications. In the Royal Society address mentioned earlier,[121] she quoted Jamaican-born British intellectual Stuart Hall as once saying "systems of classification are themselves objects of power."[122] She then connected these earlier applications with current efforts in China to identify 'criminality' of people based on their photographs,[123] a direct modern update of the (discredited) 'sciences' of physiognomy and phrenology. She concluded that using AI in this way "seems like repeating the errors of history…and then putting those tools into the hands of the powerful. We have an ethical obligation to learn the lessons of the past."[124]

2. **Multiple Malicious Misuses**. In February 2018, a group of 26 authors from 14 institutions led by Miles Brundage released a 100-page advisory entitled "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation."[125] The report recommended practices for policymakers, researchers and engineers, including actively planning for misuse of AI applications, and structured these recommendations around the three areas of digital security, physical security, and political security. The first two are frequent topics among IT professionals —albeit without the AI context — however the third is perhaps new to many readers. Brundage *et al*. define political security threats to be

> **"**The use of AI to automate tasks involved in surveillance (e.g. analysing mass-collected data), persuasion (e.g. creating targeted propaganda), and deception (e.g. manipulating videos) may expand threats associated with privacy invasion and social manipulation. We also expect novel attacks that take advantage of an

> improved capacity to analyse human behaviors, moods, and beliefs on the basis of available data. These concerns are most significant in the context of authoritarian states, but may also undermine the ability of democracies to sustain truthful public debates."

As we have already cited from various news outlets, such misuses are not mere potentialities.

3. **Slaughterbots**. In 2017 The Future of Life Institute produced a video by Stuart Russell (of "Russell & Norvig," the longtime-standard textbook for AI[126]) called "Slaughterbots"[127] to draw attention to the need to oppose autonomous weapons systems (AWS) development, which they term "killer robots": "weapons systems that, once activated, would select and fire on targets without meaningful human control."[128] In this video, tiny quadcopter drones endowed with shaped explosive charges are able to target individuals for assassination using facial recognition. The use of AI allows the drones to act autonomously, with two main implications: 1. the weapons system can *scale* to arbitrarily large numbers of drones — the video shows thousands being released over a city — and 2. the lack of communication with a central control system provides a measure of *anonymity* to the party deploying the drones.

D. The Holy:

In addition to AI systems which might serve the public at large, one might consider applications benefitting the church. Here I am concerned with applications of ML systems, not AGIs. Questions regarding the personhood of AGIs and the roles and activities available to them — would they have souls, could they pray, could they be 'saved,' could they be priests, could they be wiser than us, and so on — are beyond the scope of this article, but can be found in many other sources.[129,130,131] Answers to these would be determined by the ontology ascribed to such entities, a discussion which is still incomplete.[132] There are still other interesting topics regarding present-day ML systems worth investigating, which we describe briefly here.

1. **Dr. Theo*philus*, an AI "Monk."** For much of church history, the scholarly work of investigating and analyzing data of historical, demographic or theological significance was done by monks. In our time, one could imagine AI systems performing monk-like duties: investigating textual correlations in Scripture, predicting trends in missions or church demographics, aiding in statistical analysis of medical miracle reports, aiding in (or autonomously performing) translation of the Bible or other forms of Christian literature, or analyzing satellite images to make archaeological discoveries.[133]
2. **Chatbots for the Broken.** London-based evangelism organization CVGlobal.co use ML for content recommendation ("if you liked this article, you might like") for their "Yes He Is" website,[134] and also have developed a "Who is Jesus" chatbot to respond to common questions about the person of Christ, the message of the gospels, and some typical questions that arise in apologetics contexts. This is essentially the same program as those used by major corporations such as banks[135] to answer common questions about their organizations. One can argue over whether this removes the 'relational' element of witnessing in a 'profane' way; the structure of such a tool amounts to turning an "FAQ" page (e.g. "Got Questions about Jesus?"[136]) into an interactive conversational model. Relatedly, researchers at Vanderbilt University have gained attention for their use of ML to predict the risk of suicide,[137] and apps exist for tracking mental and spiritual health,[138]

and thus a call for has gone out for investigating predictive models in mental and spiritual counseling.[139]

3. **Being Engaged with AI Ethics.** This is more of an *opportunity for engagement* rather than a *use* of AI. Discussions on topics affecting society such as those described in this document should not be limited to only secular, non-theistic sources. There are significant points of commonality between Christian worldviews and others on topics involving affirming human dignity and agency, resisting the exploitation and oppression of other human beings, and showing concern for the poor and others affected economically by the automation afforded by AI.[140,141] The world at large is interested in having these discussions, and persons informed by wisdom and spiritual principles are integral members at the table for providing ethical input. We will revisit the topic of foundations for ethics in Part 5.

E. A Tool, But Not "Just a Tool"

In casting AI as a tool to be used by humans "for good or evil," we shouldn't make the mistake of thinking all tools are "neutral," i.e., that they do not have intentions implied by their very design. As an example of this, the Future of Humanity's information page on "AI Safety Myths" points out, "A heat-seeking missile has a goal."[142] Referring to our earlier list of uses: while it is true that stopping sex trafficking is "good" and repressing political dissidents is "bad," both are examples of surveillance technology, which by its nature imposes a sacrifice of personal privacy. (The tradeoff between security and privacy is an age-old discussion; for now we simply note that AI may favor applications on the security side.)

Sherry Turkle of MIT chronicled the introduction of computers into various fields in the early 1980s, and observed that those asserting that the computer was "just a tool" indicated a lack of reflection: "Calling the computer 'just a tool,' even as one asserted that tools shape thought, was a way of saying that a big deal was no big deal."[143] Turkle cited the famous question of architect Louis Kahn, asking a brick what it wants —"'What do you want, brick?' And brick says to you, 'I like an arch'"[144] — and she asked the new question "What does a simulation want?" In the words of those she interviewed, simulations favor experimentation. The results of its use include a disconnect from reality ("it can tempt its users into a lack of fealty to the real"), and as a consequence, users must cultivate a healthy doubt of their simulations.

Thus we do well to ask: What does an AI 'want'? What forms of usage does it favor? What sorts of structures will it promote and/or rely on? (Keep in mind, we are referring here to modern ML algorithms, not fictional sentient AGIs.) We conclude this section by briefly answering each of these.

0. Like any piece of software, **AI wants to be used**. This led to Facebook employing psychological engineering to generate "eyeball views" and addictive behavior,[145] including experimenting on users without their consent and without ethical oversight.[146] The more use, the more data, which fits in with the next point:
1. **An AI Wants Data.** Given their statistical nature, the rise of successful ML algorithms is closely linked with the rise in availability of large amounts of data (to train on) made possible by the internet,[147] rather than from improvements in the underlying algorithms. This even motivates some ML experts to advocate improving a model's performance via getting more data rather than adjusting an algorithm.[148] It may be said that ML systems

are data-hungry, and *data-hungry algorithms make for data-hungry companies and governments.* Thus we see the rise of tracking everything users do online for the purposes of mining later, and Google contracting with the healthcare system of the UK for the exchange of user data.[149]

2. **An AI Wants "Compute."** A corollary of #1. In order to 'burn through' gargantuan amounts of data, huge computational resources are required. This is the other reason for the rise of ML systems: significant advances in computing hardware, notably graphics processing units (GPUs). Thus, vast data centers and server farms have arisen, and the energy consumption of large-scale AI systems is an increasing environmental concern.[150] In response, Google has built dedicated processing units to reduce their energy footprint,[151] but with the growth of GPU usage significantly outpacing Moore's Law,[152] this energy concern isn't going away. Some are proposing to distribute the computation to low-power onboard sensors,[153] which is also likely to occur. Either way, "AI wants compute."
3. **AI Tempts Toward 'Magic Box' Usage.** "Give the system a bunch of inputs, and a bunch of labeled outputs, and *let the system figure out* how to map one to the other." So goes the hope of many a new ML application developer, and when this works, it can be fun and satisfying (see, e.g., some of my own experiments[154]). This can be one of the strengths of ML systems, freeing the developer from having to understand and explicitly program how to map complicated inputs to outputs, allowing the "programmer" to be creative, such as with Rebecca Fiebrink's Wekinator ML tool for musicians.[155] But this can also encourage lazy usage such as the "physiognomy" applications cited by Kate Crawford, and biased models which accidentally discriminate against certain groups (of

which there too many instances to cite). As with simulation, users should cultivate a healthy doubt of their correlations.

Finally, in terms of what other structures AI will promote and/or rely on, we should remember the general warnings on technological development by Christian philosopher Jacques Ellul. In *The Technological Society,* Ellul cautioned that "purposes drop out of sight and efficiency becomes the central concern."[156] Furthermore, Ellul noted that successful technological development tends to become self-serving, as we have all inherited the nature of Cain, the first city-builder who was also the first murderer.[157] In the next section, we will relate some current conversations aimed at keeping AI development and government oriented toward serving people.

Part 5: Further Fertile Fields

Five "AI Ethics & Society" conversations to follow

For the closing section, I have selected five areas of current conversation that I find to be particularly worth paying attention to. This section is not exhaustive or authoritative.

A. Bias

A popular conversation in recent years is the topic of biased machine learning models, such as those which associate negative connotations with certain races[158] or predict employability on the basis of gender,[159] although such occurrences are nothing new to statisticians, and have been equally attributed to "Big Data" as much as to AI.[160] There are numerous conversations regarding how to "fix" bias[161] or at least detect, measure and mitigate it.[162] While these are important and worthy efforts, one can foresee that as long as there are bad statisticians – i.e., people doing sloppy statistics — there will be biased models. And machine learning (ML) automates bad statistics (though typically not through the algorithms involved but through the datasets used to train the models). Thus the problem of bias is both a current topic and one which is likely to remain relevant for some time to come.

B. Black Boxes vs. Transparency

In Part 2 we mentioned requirements that algorithmic decisions should be "explainable,"[163] as opposed to "opaque"[164] systems which function as "black boxes."[165,166] Two main approaches present themselves:

1. **Probing Black Boxes**. One approach is to use various methods to probe black box systems, by observing how they map inputs to outputs. Examples include learning the decision rules of systems in an explainable way (and even mimicking the existing system)[167] and extracting "rationales"[168] — short textual summaries of significant input data. A related approach involves mapping entire subsets at a time to predict the "boundaries" of possible outputs from a system, e.g. for safety prediction.[169]
2. **Transparency As a Design Requirement.** For several years, there have been calls to produce systems which are transparent *by design.*[170] Such considerations are essential for users to form accurate mental models of a system's operation,[171] which may be a key ingredient to fostering user trust.[172] Further, transparent systems are essential for government accountability and providing a greater sense of agency for citizens.[173] But how to actually design useful, transparent interfaces for robots[174,175] and computer systems in general[176] remains an active area of research, both in terms of the designs

> themselves and in measuring their effects with human users — even when it comes to the education of data science professionals.[177] One cannot simply overwhelm the user with data. This is particularly challenging for neural network systems, where the mapping of high-dimensional data exceeds the visualization capacities of humans, and even on simple datasets such as MNIST, dimensionality-reduction methods such as t-SNE[178] and interactive visualizations[179] can still leave one lacking a sense of clarity. This is an active area of research, with two particularly active efforts by the group at the University of Bath (Rob Wortham, Andreas Theodorou, and Joanna Bryson)[180] and by Chris Olah.[181] It's also worth mentioning the excellent video by Brett Victor on designing for understanding,[182] although this is not particular to algorithmic decision making.

One 'hybrid' form of the two above approaches involves providing counterfactual statements, such as in the example, "You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan."[183] The second statement is a counterfactual, and while not offering full transparency or explainability, provides at least a modicum of guidance. This may be a minimal prescription for rather simple algorithms, although for complex systems with many inputs, such statements may be difficult to formulate.

C. AI Ethics Foundations

In reading contemporary literature on the topic of "AI Ethics," one may not frequently see people stating explicitly where they're coming from, in terms of the foundations of their ethics, and rather one often sees the "results," i.e. the ethical directives built upon those foundations. Joanna Bryson, whom we've cited many times, is explicit about working from a framework of functionalism,[184] which she applies to great effect, and reaches conclusions which are often in agreement with other traditions. Alternatively, philosopher Shannon Vallor (co-chair of this year's AAAI/ACM conference on AI, Ethics and Society) in her book, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting,*[185] advocates the application of virtue ethics to matters of technological development. Virtue ethics provides a motivation toward good behavior on the principle of "excellence" of character, leading to the greatest thriving of the individual and thus of society. Drawing from the ancient traditions of Aristotelianism, Confucianism, and Buddhism, and religious parallels in Christian and Islamic thought, and western philosophical treatises such as those of Immanuel Kant and the critiques by Nietzsche, Vallor develops an adaptive framework that eschews rule-based pronouncements in favor of "technomoral flexibility," which she defines as "a reliable and skillful disposition to modulate action, belief, and feeling as called for by novel, unpredictable, frustrating, or unstable technosocial conditions." In the Christian tradition, Brent Waters has written on moral philosophy "in the emerging technoculture,"[186] and while not addressing AI in particular, many of his critiques provide somewhat of a (to borrow some jargon from machine learning) "regularizing" influence enabling one to approach the hype of AI development in a calm and reflective manner.

D. Causal Calculus

If neural networks and their ilk are mere "correlation machines"[187] akin to polynomial regression,[188] how can we go from correlation to inferring causality? Put differently, how can we go from "machine learning" to "predictive analytics"?[189] Turing Award winner Judea Pearl in his 2018 book *The Book of Why*[190] (aimed at a more popular audience than his more technical

Causality[191]) offers a set of methods termed "causal calculus" defined over Bayesian Networks (a term coined by Pearl).[192] This book has generated many favorable reviews from within the AI community and has been regarded as contributing an essential ingredient toward the development of more powerful, human-like AI.[193] In a 2018 report to the Association of Computing Machinery (ACM),[194] Pearl highlights seven tasks which are beyond the reach of typical statistical learning systems but have been satisfied using causal modeling. Many further applications by other researchers of this method are likely to appear in the near future.

E. Transformative AI

One does not need to have fully conscious, sentient AGI in order to have AI that can still have a severely disruptive and possibly dangerous impact on human life on a large scale. Such systems will likely exhibit forms of superintelligence[195] across multiple domains, in a manner not currently manifested in the world (i.e., not in the familiar forms of collective human action, or artifact-enhanced human cognition). Planning to mitigate risks associated with such outcomes comprises the field of AI Safety.[196] In late September 2018, the Future of Humanity Institute released a report by Allan Dafoe entitled *AI Governance: A Research Agenda* in which he "focuses on extreme risks from advanced AI."[197] Dafoe distinguishes AI Governance from AI Safety by emphasizing that safety "focuses on the technical questions of how AI is built" whereas governance "focuses on the institutions and contexts in which AI is built and used." In describing risks and making recommendations, Dafoe focuses on what he calls "transformative AI (TAI), understood as advanced AI that could lead to radical changes in welfare, wealth, or power." Dafoe outlines an agenda for research which seems likely to be taken up by many interested researchers.

Summary

Starting from an optimistic view of a future utopia governed by AIs who make benevolent decisions in place of humans (with their tendency toward warfare and abuse of the environment), we have noted that AI systems are unlikely to represent the world or other concepts in ways which are intuitive or even explainable to humans. This carries a risk to basic civil liberties, and efforts to make such systems more explainable and transparent are actively being pursued. Even so, such systems will and simply *do* require human political activity in the form of implementation choices and auditing such as checking for bias, and thus humans will remain the decision-makers, as they should be. While the unlikelihood of the realization of a quasi-religious hope of future AI saviors may be disappointing to science fiction fans, it means, in the words of Christina Colclough, (Senior Policy Advisor, UNI Global Union), that we can avoid "technological determinism" and we can talk about and "agree on the kind of future we want."[198] We have seen that AI is a powerful tool for good and for evil, and yet it is not "neutral": it prefers large amounts of data (which may involve privacy concerns), large computing resources and thus large energy consumption, and may favor unreflective "magical thinking" which empowers sloppy statistics and biased inferences. Drawing causal inferences from the correlations of machine learning is problematic, but work in the area of causal modeling may allow for much more powerful AI systems. These powerful systems may themselves become transformative existential threats and will require planning for safety and governance to ensure that such systems favor human thriving. The conception of what constitutes human thriving is an active area of discussion among scholars with diverse ideological and religious backgrounds, and is a fertile area for dialog between these groups, for the goal of fostering a harmonious human society.

Acknowledgements

The author wishes to thank the following for helpful discussions: Michael Burdett, William Hooper, Tommy Kessler, Stan Rosenberg, Andy Watts, Miles Brundage, Beth Singler, Andreas Theodorou, Robert Wortham, Joanna Bryson, Micah Redding and Nathan Griffith. This work was sponsored by a grant given by Bridging the Two Cultures of Science and the Humanities II, a project run by Scholarship and Christianity in Oxford (SCIO), the UK subsidiary of the Council for Christian Colleges and Universities, with funding by Templeton Religion Trust and The Blankemeyer Foundation.

References

[1] W R F Browning, ed., *Theophilus, in The Oxford Dictionary of the Bible* (Oxford University Press, 2004).

[2] “Theopolis” is a proper name that showed some popularity in the 19th century, and is also sometimes attributed to individuals more commonly known by the name “Theophilus”, e.g., John Milton uses the former in Of Prelatical Episcopacy (1641) to refer to the 23rd Pope of Alexandria.

[3] Wikipedia contributors, “The City of God,” March 23, 2018, https://en.wikipedia.org/w/index.php?title=The_City_of_God&oldid=832052290.

[4] I have been unable to find any references regarding the creators' intent in choosing this name. “Theo” was a new character for the 1979 TV series, and not part of the original Buck Rogers comic strip.

[5] Sanford Kessler, “Religious Freedom in Thomas More’s Utopia,” *The Review of Politics* 64, no. 02 (March 2002): 207, https://doi.org/10.1017/S0034670500038079.

[6] Teresa Jusino, “Religion and Science Fiction: Asking the Right Questions,” Tor.com, January 6, 2010, https://www.tor.com/2010/01/06/religion-and-science-fiction-asking-the-right-questions/.

[7] Anna Fava, “Science Fiction — Mythology of the Future,” *Think Magazine*, December 2014, https://www.um.edu.mt/think/science-fiction-mythology-of-the-future/.

[8] “Doctor Theopolis (Character) - Comic Vine,” accessed May 27, 2018, https://comicvine.gamespot.com/doctor-theopolis/4005-76853/.

[9] “Buck Rogers in the 25th Century S01e01 Episode Script | SS,” accessed May 27, 2018, https://www.springfieldspringfield.co.uk/view_episode_scripts.php?tv-show=buck-rogers-in-the-25th-century&episode=s01e01.

[10] Sarah Begley, “The Mysterious Case of the Missing Utopian Novels,” *Time*, September 28, 2017, http://time.com/4960648/science-fiction-utopian-novels-books/.

[11] Tom Cassauwers, “Sci-Fi Doesn’t Have to Be Depressing: Welcome to Solarpunk,” accessed May 27, 2018, https://www.ozy.com/fast-forward/sci-fi-doesnt-have-to-be-depressing-welcome-to-solarpunk/82586; Adam Epstein, *“I Miss Optimism”: The “Family Guy” Creator Wants to Bring Back Hopeful Sci-Fi* (Quartz, 2017), https://qz.com/1052758/the-family-guy-creator-wants-to-bring-back-optimistic-sci-fi/; Cory Doctorow, “In ‘Walkaway,’ A Blueprint For A New, Weird (But Better) World,” *NPR*, April 27, 2017, https://www.npr.org/2017/04/27/523587179/in-walkaway-a-blueprint-for-a-new-weird-but-better-world.

[12] Lauren J Young, “How To Move Beyond The Tropes Of Dystopia,” accessed May 27, 2018, https://www.sciencefriday.com/articles/just-topia-moving-beyond-the-tropes-of-dystopia/.

[13] Mahmoud Tarrasse, *What Is Wrong with Convolutional Neural Networks ? – Towards Data Science* (Towards Data Science, 2018), https://towardsdatascience.com/what-is-wrong-with-convolutional-neural-networks-75c2ba8fbd6f.

[14] Kevin Hartnett, “To Build Truly Intelligent Machines, Teach Them Cause and Effect,” *Quanta Magazine*, May 15, 2018, https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/.

[15] Bruce Schneier et al., “Why ‘Anonymous’ Data Sometimes Isn’t,” *Wired*, December 2007.

[16] Sander Dieleman, Kyle W Willett, and Joni Dambre, “Rotation-Invariant Convolutional Neural Networks for Galaxy Morphology Prediction,” *Mon. Not. R. Astron. Soc.* 450, no. 2 (June 2015): 1441–1459.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (Springer International Publishing, 2015), 234–241.

[18] P B Wigley et al., “Fast Machine-Learning Online Optimization of Ultra-Cold-Atom Experiments,” *Sci. Rep.* 6 (May 2016): 25890.

[19] Daniel George and E A Huerta, “Deep Learning for Real-Time Gravitational Wave Detection and Parameter Estimation: Results with Advanced LIGO Data,” *Phys. Lett. B* 778 (March 2018): 64–70.

[20] Jamie McGee, “How a Franklin Software Company Helped Rescue 6,000 Sex Trafficking Victims,” July 6, 2017, https://www.tennessean.com/story/money/2017/07/06/franklins-digital-reasoning-creates-tool-has-helped-rescue-6-000-sex-trafficking-victims/327668001/.

[21] Andrew Ng, “Andrew Ng: Why AI Is the New Electricity,” accessed August 28, 2018, https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity.

[22] ‘Learn’ here means iteratively minimizing an error function or maximizing a reward function.

[23] Michael Jordan, “Artificial Intelligence — The Revolution Hasn’t Happened Yet,” April 19, 2018, https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7.

[24] Pranav Rajpurkar et al., “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning” (November 14, 2017), http://arxiv.org/abs/1711.05225.

[25] Aman Agarwal, “Explained Simply: How DeepMind Taught AI to Play Video Games,” August 27, 2017, https://medium.freecodecamp.org/explained-simply-how-deepmind-taught-ai-to-play-video-games-9eb5f38c89ee.

[26] “AlphaGo | DeepMind,” accessed May 27, 2018, https://deepmind.com/research/alphago/.

[27] Bryan Casey, Ashkon Farhangi, and Roland Vogl, “Rethinking Explainable Machines: The GDPR’s ‘Right to Explanation’ Debate and the Rise of Algorithmic Audits in Enterprise,” *Berkeley Technology Law Journal (Submitted)*, n.d.

[28] “Call for Papers, ICLR 2018, Sixth International Conference on Learning Representations,” accessed June 5, 2018, https://iclr.cc/Conferences/2018/CallForPapers.

[29] Tomas Mikolov et al., “Efficient Estimation of Word Representations in Vector Space,” *ArXiv:1301.3781 [Cs]*, January 16, 2013, http://arxiv.org/abs/1301.3781.

[30] Paris Smaragdis, *NMF? Neural Nets? It’s All the Same...*, SANE 2015 (at 32:12: YouTube), accessed June 5, 2018, https://www.youtube.com/watch?v=wfmpViJIjWw.

[31] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert, “Feature Visualization,” *Distill* 2, no. 11 (November 7, 2017): e7, https://doi.org/10.23915/distill.00007.

[32] C. Perlich et al., “Machine Learning for Targeted Display Advertising: Transfer Learning in Action,” *Machine Learning* 95, no. 1 (April 1, 2014): 103–27, https://doi.org/10.1007/s10994-013-5375-2.

[33] Byron Spice, "Carnegie Mellon Reveals Inner Workings of Victorious Poker AI | Carnegie Mellon School of Computer Science," accessed June 5, 2018, https://www.scs.cmu.edu/news/carnegie-mellon-reveals-inner-workings-victorious-poker-ai.

[34] Alvin Plantinga, *Warrant and Proper Function* (New York: Oxford University Press, 1993).

[35] Roger Penrose, "Why Algorithmic Systems Possess No Understanding" (May 15, 2018).

[36] John Launchbury and DARPAtv, *A DARPA Perspective on Artificial Intelligence*, accessed June 4, 2018, https://www.youtube.com/watch?v=-O01G3tSYpU.

[37] Jesus Rodriguez, "The Missing Argument: Motivation and Artificial Intelligence," *Medium* (blog), August 14, 2017, https://medium.com/@jrodthoughts/the-missing-argument-motivation-and-artificial-intelligence-f582649a2680.

[38] Dr Vyacheslav Polonski, "Can We Teach Morality to Machines? Three Perspectives on Ethics for Artificial Intelligence," *Medium* (blog), December 19, 2017, https://medium.com/@drpolonski/can-we-teach-morality-to-machines-three-perspectives-on-ethics-for-artificial-intelligence-64fe479e25d3.

[39] Ray Kurzweil, *How to Create a Mind: The Secret of Human Thought Revealed* (New York: Viking, 2012).

[40] Antonio Regalado, "The Brain Is Not Computable," MIT Technology Review, accessed June 5, 2018, https://www.technologyreview.com/s/511421/the-brain-is-not-computable/.

[41] Balázs Szigeti et al., "OpenWorm: An Open-Science Approach to Modeling Caenorhabditis Elegans," *Frontiers in Computational Neuroscience* 8 (November 3, 2014), https://doi.org/10.3389/fncom.2014.00137.

[42] "Turing Completeness," *Wikipedia*, May 28, 2018, https://en.wikipedia.org/w/index.php?title=Turing_completeness&oldid=843397556.

[43] Spike Jonze, *Her* (Warner Bros. Entertainment, 2014).

[44] Micah Redding, "Also Reminds Me of Angels in the Biblical Story, Whose Names Are 'Too Wonderful for You to Know,'" July 15, 2018.

[45] "SysML Conference," accessed June 5, 2018, https://www.sysml.cc/.

[46] Robert H. Wortham, Andreas Theodorou, and Joanna J. Bryson, "What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems," 2016.

[47] Mike Judge, *Idiocracy* (20th Century Fox, 2006), http://www.imdb.com/title/tt0387808/.

[48] Colin Lecher, "A Healthcare Algorithm Started Cutting Care, and No One Knew Why," The Verge, March 21, 2018, https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

[49] Thomas Hills, "The Mental Life of a Bank Loan Algorithm: A True Story," *Psychology Today*, accessed October 8, 2018, https://www.psychologytoday.com/blog/statistical-life/201810/the-mental-life-bank-loan-algorithm-true-story.

[50] Megan Palin, "China's 'Social Credit' System Is a Real-Life 'Black Mirror' Nightmare," *New York Post* (blog), September 19, 2018, https://nypost.com/2018/09/19/chinas-social-credit-system-is-a-real-life-black-mirror-nightmare/.

[51] Investopedia Staff, "High-Frequency Trading - HFT," Investopedia, July 23, 2009, https://www.investopedia.com/terms/h/high-frequency-trading.asp.

[52] Matt Phillips, "Nasdaq: Here's Our Timeline of the Flash Crash," *Wall Street Journal*, May 11, 2010, https://blogs.wsj.com/marketbeat/2010/05/11/nasdaq-heres-our-timeline-of-the-flash-crash/.

[53] "DLI | Speed Confence | Cornell Tech," Digital Life Initiative | Cornell Tech | New York, accessed October 8, 2018, https://www.dli.tech.cornell.edu/speed.

[54] Janosch Delcker, "Europe Divided over Robot 'Personhood,'" *Politico*, April 11, 2018, https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/.

[55] Joanna J. Bryson, "Robots Should Be Slaves," *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, 2010, 63–74.

[56] Joanna J. Bryson, "R/Science - Science AMA Series: I'm Joanna Bryson, a Professor in Artificial (and Natural) Intelligence. I Am Being Consulted by Several Governments on AI Ethics, Particularly on the Obligations of AI Developers towards AI and Society. I'd Love to Talk – AMA!," reddit, accessed October 8, 2018, https://www.reddit.com/r/science/comments/5nqdo7/science_ama_series_im_joanna_bryson_a_professor/.

[57] Josephine Johnston, "Traumatic Responsibility: Victor Frankenstein as Creator and Casualty," in *Frankenstein*, ed. Mary Wollstonecraft Shelley, Annotated for Scientists, Engineers, and Creators of All Kinds (MIT Press, 2017), 201–8, http://www.jstor.org/stable/j.ctt1pk3jfp.11.

[58] Michael Burdett, "Danny Boyle's Frankenstein: An Experiment in Self-Imaging," Transpositions, December 5, 2016, http://www.transpositions.co.uk/danny-boyles-frankenstein-an-experiment-in-self-imaging/.

[59] Joanna J. Bryson, “Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics,” *Ethics Inf. Technol.* 20, no. 1 (March 1, 2018): 15–26, https://doi.org/10.1007/s10676-018-9448-6.

[60] Terrell Bynum, “Computer and Information Ethics,” in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Summer 2018 (Metaphysics Research Lab, Stanford University, 2018), https://plato.stanford.edu/archives/sum2018/entries/ethics-computer/.

[61] Norbert Wiener, *The Human Use of Human Beings: Cybernetics and Society*, The Da Capo Series in Science (New York, N.Y: Houghton Mifflin Harcourt, 1950).

[62] Joseph Weizenbaum, “ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine,” *Commun. ACM* 9, no. 1 (1966): 36–45.

[63] Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (San Francisco: Freeman, 1976).

[64] Diana ben-Aaron, “Weizenbaum Examines Computers and Society,” *The Tech*, April 9, 1985, http://tech.mit.edu/V105/N16/weisen.16n.html.

[65] Kate Crawford, *Just An Engineer: The Politics of AI*, You and AI (The Royal Society: YouTube, 2018), https://www.youtube.com/watch?v=HPopJb5aDyA.

[66] Sidney Fussell, “The LAPD Uses Palantir Tech to Predict and Surveil ‘Probable Offenders,’” Gizmodo, May 8, 2018, https://gizmodo.com/the-lapd-uses-palantir-tech-to-predict-and-surveil-prob-1825864026.

[67] Rebecca Hill, “Rights Group Launches Legal Challenge over London Cops’ Use of Facial Recognition Tech,” *The Register*, July 26, 2018, https://www.theregister.co.uk/2018/07/26/big_brother_watch_legal_challenge_facial_recognition/.

[68] Travis Galey, Kris Van Cleave, and 12:32 Pm, “Feds Use Facial Recognition to Arrest Man Trying to Enter U.S. Illegally,” CBS news, August 23, 2018, https://www.cbsnews.com/news/customs-and-border-protection-use-facial-recognition-to-arrest-man-trying-to-enter-u-s-illegally/.

[69] Palin, “China’s ‘Social Credit’ System Is a Real-Life ‘Black Mirror’ Nightmare.”

[70] Sam Meredith, “Facebook-Cambridge Analytica: A Timeline of the Data Hijacking Scandal,” April 10, 2018, https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html.

[71] “AI for Good Global Summit 2018,” accessed October 7, 2018, https://www.itu.int/en/ITU-T/AI/2018/Pages/default.aspx.

[72] “AI for Good Foundation,” accessed October 7, 2018, https://ai4good.org/.

[73] Sean Captain, “This News Site Claims Its AI Writes ‘Unbiased’ Articles,” Fast Company, April 4, 2018, https://www.fastcompany.com/40554112/this-news-site-claims-its-ai-writes-unbiased-articles.

[74] Ben Dickson, “Why It’s so Hard to Create Unbiased Artificial Intelligence,” *TechCrunch*, November 7, 2016, http://social.techcrunch.com/2016/11/07/why-its-so-hard-to-create-unbiased-artificial-intelligence/.

[75] Jeffrey Dastin, “Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women,” *Reuters*, October 10, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[76] Celine Herweijer, “8 Ways AI Can Help Save the Planet,” World Economic Forum, January 24, 2018, https://www.weforum.org/agenda/2018/01/8-ways-ai-can-help-save-the-planet/.

[77] Sarath Muraleedharan, “Role of Artificial Intelligence in Environmental Sustainability,” *EcoMENA* (blog), March 6, 2018, https://www.ecomena.org/artificial-intelligence-environmental-sustainability/.

[78] Dan Robitzski, “Advanced Artificial Intelligence Could Run The World Better Than Humans Ever Could,” Futurism, August 29, 2018, https://futurism.com/advanced-artificial-intelligence-better-humans.

[79] ben-Aaron, “Weizenbaum Examines Computers and Society.”

[80] Nathan Griffith, “PSC 3610 Game Theory and Public Choice,” in *Undergraduate Catalog 2018-2019* (Belmont University, 2018), http://catalog.belmont.edu/preview_course_nopop.php?catoid=3&coid=4367.

[81] Carl Sagan, “A New Way to Think About Rules to Live By,” *Parade*, November 28, 1993.

[82] John F. Nash, “Equilibrium Points in N-Person Games,” *Proceedings of the National Academy of Sciences of the United States of America* 36, no. 1 (1950): 48–49.

[83] Ryan Porter, Eugene Nudelman, and Yoav Shoham, “Simple Search Methods for Finding a Nash Equilibrium,” *Games and Economic Behavior* 63, no. 2 (July 2008): 642–62, https://doi.org/10.1016/j.geb.2006.03.015.

[84] Joseph Y. Halpern, Rafael Pass, and Daniel Reichman, “On the Non-Existence of Nash Equilibrium in Games with Resource-Bounded Players,” *ArXiv:1507.01501 [Cs]*, July 6, 2015, http://arxiv.org/abs/1507.01501.

[85] "Braess's Paradox," *Wikipedia*, October 8, 2018, https://en.wikipedia.org/w/index.php?title=Braess%27s_paradox&oldid=863068294.

[86] William Chen, "Bad Traffic? Blame Braess' Paradox," Forbes, October 20, 2016, https://www.forbes.com/sites/quora/2016/10/20/bad-traffic-blame-braess-paradox/.

[87] Ariel Rubinstein, "Game theory: How game theory will solve the problems of the Euro Bloc and stop Iranian nukes," *FAZ.NET*, March 27, 2013, sec. Feuilleton, http://www.faz.net/1.2130407.

[88] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge, "A Neural Algorithm of Artistic Style," *ArXiv:1508.06576 [Cs, q-Bio]*, August 26, 2015, http://arxiv.org/abs/1508.06576; Shubhang Desai, "Neural Artistic Style Transfer: A Comprehensive Look," *Medium* (blog), September 14, 2017, https://medium.com/artists-and-machine-intelligence/neural-artistic-style-transfer-a-comprehensive-look-f54d8649c199.

[89] Taeksoo Kim et al., "Learning to Discover Cross-Domain Relations with Generative Adversarial Networks," *ArXiv:1703.05192 [Cs]*, March 15, 2017, http://arxiv.org/abs/1703.05192.

[90] Nicolas Heess et al., "Emergence of Locomotion Behaviours in Rich Environments," *ArXiv:1707.02286 [Cs]*, July 7, 2017, http://arxiv.org/abs/1707.02286; Caroline Chan et al., "Everybody Dance Now," *ArXiv:1808.07371 [Cs]*, August 22, 2018, http://arxiv.org/abs/1808.07371.

[91] Judge, *Idiocracy*.

[92] Evgeny Morozov, *To Save Everything, Click Here: The Folly of Technological Solutionism* (New York: PublicAffairs, 2014).

[93] Living on Earth / World Media Foundation / Public Radio International, "Living on Earth: Gus Speth Calls for A," Living on Earth, accessed October 10, 2018, https://www.loe.org/shows/segments.html?programID=15-P13-00007&segmentID=6.

[94] Erle C. Ellis, "Opinion | Science Alone Won't Save the Earth. People Have to Do That.," *The New York Times*, August 11, 2018, sec. Opinion, https://www.nytimes.com/2018/08/11/opinion/sunday/science-people-environment-earth.html.

[95] Crawford, *You and AI – Just An Engineer*.

[96] Richard Sargeant, "AI Ethics: Send Money, Guns & Lawyers," *Afterthought* (blog), June 20, 2018, https://sargeant.me/2018/06/20/ai-ethics-send-money-guns-lawyers/.

[97] Zeynep Tufekci, "Let Me Say: Too Many Worry about What AI—as If Some Independent Entity—Will Do to Us...," *@zeynep on Twitter* (blog), September 4, 2017, https://twitter.com/zeynep/status/904707522958852097?lang=en.

[98] Scott H. Hawley, "Challenges for an Ontology of Artificial Intelligence," *Accepted for Publication in Perspectives on Science and Christian Faith*, October 13, 2018, http://hedges.belmont.edu/~shawley/AIOntologyChallenges_Hawley.pdf.

[99] Joanna J. Bryson and Philip P Kime, "Just an Artifact: Why Machines Are Perceived as Moral Agents," vol. 22, 2011, 1641, http://www.aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/viewFile/3376/3774.

[100] Per Christensson, "Plug and Play Definition," TechTerms, 2006, https://techterms.com/definition/plugandplay.

[101] Katie Bird Head, "ISO/IEC Standard on UPnP Device Architecture Makes Networking Simple and Easy," ISO, accessed October 11, 2018, http://www.iso.org/cms/render/live/en/sites/isoorg/contents/news/2008/12/Ref1185.html.

[102] "AIED2018 – International Conference on Artificial Intelligence in Education," accessed October 12, 2018, https://aied2018.utscic.edu.au/.

[103] Johnny Langenheim, "AI Identifies Heat-Resistant Coral Reefs in Indonesia," *The Guardian*, August 13, 2018, sec. Environment, https://www.theguardian.com/environment/the-coral-triangle/2018/aug/13/ai-identifies-heat-resistant-coral-reefs-in-indonesia.

[104] Dieleman, Willett, and Dambre, "Rotation-Invariant Convolutional Neural Networks for Galaxy Morphology Prediction."

[105] McGee, "How a Franklin Software Company Helped Rescue 6,000 Sex Trafficking Victims."

[106] "Digital Reasoning: Most Innovative Company," Fast Company, March 19, 2018, https://www.fastcompany.com/company/digital-reasoning.

[107] H A Haenssle et al., "Man against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists," *Annals of Oncology* 29, no. 8 (August 1, 2018): 1836–42, https://doi.org/10.1093/annonc/mdy166.

[108] Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning."

[109] Seung Seog Han et al., “Deep Neural Networks Show an Equivalent and Often Superior Performance to Dermatologists in Onychomycosis Diagnosis: Automatic Construction of Onychomycosis Datasets by Region-Based Convolutional Deep Neural Network,” ed. Manabu Sakakibara, *PLOS ONE* 13, no. 1 (January 19, 2018): e0191493, https://doi.org/10.1371/journal.pone.0191493.

[110] Stephen F. Weng et al., “Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?,” ed. Bin Liu, *PLOS ONE* 12, no. 4 (April 4, 2017): e0174944, https://doi.org/10.1371/journal.pone.0174944.

[111] IEEE, “AI vs Doctors,” IEEE Spectrum: Technology, Engineering, and Science News, September 26, 2017, https://spectrum.ieee.org/static/ai-vs-doctors.

[112] Luke Oakden-Rayner, “CheXNet: An in-Depth Review,” *Luke Oakden-Rayner (PhD Candidate / Radiologist) Blog* (blog), January 24, 2018, https://lukeoakdenrayner.wordpress.com/2018/01/24/chexnet-an-in-depth-review/.

[113] Felix Salmon, “IBM’s Watson Was Supposed to Change the Way We Treat Cancer. Here’s What Happened Instead.,” Slate Magazine, August 18, 2018, https://slate.com/business/2018/08/ibms-watson-how-the-ai-project-to-improve-cancer-treatment-went-wrong.html.

[114] Casey Ross, “IBM Pitched Watson as a Revolution in Cancer Care. It’s Nowhere Close,” STAT, September 5, 2017, https://www.statnews.com/2017/09/05/watson-ibm-cancer/.

[115] Steve Griffiths, “Hype vs. Reality in Health Care AI: Real-World Approaches That Are Working Today,” *MedCity News* (blog), September 27, 2018, https://medcitynews.com/2018/09/hype-vs-reality-in-health-care-ai-real-world-approaches-that-are-working-today/.

[116] Michael Ahr, “The Most Evil Artificial Intelligences in Film,” Den of Geek, June 29, 2018, http://www.denofgeek.com/us/go/274559.

[117] Andrew Ng, “AI+ethics Is Important, but Has Been Partly Hijacked by the AGI (Artificial General Intelligence) Hype...,” *@andrewyng on Twitter* (blog), June 11, 2018, https://twitter.com/andrewyng/status/1006204761543081984?lang=en.

[118] Zeynep Tufekci, “My Current Lifegoal Is Spreading Realistic Nightmares...,” Twitter, *@zeynep on Twitter* (blog), June 28, 2018, https://twitter.com/zeynep/status/1012357341981888512.

[119] Tufekci, “Let Me Say: Too Many Worry about What AI—as If Some Independent Entity—Will Do to Us...”

[120] Kate Crawford, *Dark Days: AI and the Rise of Fascism*, SXSW 2017 (YouTube, 2017), https://www.youtube.com/watch?v=Dlr4O1aEJvI.

[121] Crawford, *You and AI – Just An Engineer*.

[122] Sut Jhally and Stuart Hall, *Race: The Floating Signifier* (Media Education Foundation, 1996).

[123] “Return of Physiognomy? Facial Recognition Study Says It Can Identify Criminals from Looks Alone,” RT International, accessed October 12, 2018, https://www.rt.com/news/368307-facial-recognition-criminal-china/.

[124] Crawford, *You and AI – Just An Engineer*.

[125] Miles Brundage et al., “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” *ArXiv:1802.07228 [Cs]*, February 20, 2018, http://arxiv.org/abs/1802.07228.

[126] Stuart J Russell, Stuart Jonathan Russell, and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Prentice Hall, 2010), https://market.android.com/details?id=book-8jZBksh-bUMC.

[127] Stuart Russell, *Slaughterbots*, Stop Autonomous Weapons (Future of Life Institute, 2017), https://www.youtube.com/channel/UCNaTkhskiEVg5vK3fxlluCQ.

[128] “Frequently Asked Questions,” *Ban Lethal Autonomous Weapons* (blog), November 7, 2017, https://autonomousweapons.org/sample-page/.

[129] Jonathan Merritt, “Is AI a Threat to Christianity?,” The Atlantic, February 3, 2017, https://www.theatlantic.com/technology/archive/2017/02/artificial-intelligence-christianity/515463/.

[130] Paul Scherz, “Christianity Is Engaging Artificial Intelligence, but in the Right Way,” *Crux* (blog), February 28, 2017, https://cruxnow.com/commentary/2017/02/27/christianity-engaging-artificial-intelligence-right-way/.

[131] “As Artificial Intelligence Advances, What Are Its Religious Implications?,” *Religion & Politics* (blog), August 29, 2017, https://religionandpolitics.org/2017/08/29/as-artificial-intelligence-advances-what-are-its-religious-implications/.

[132] Derek C Schuurman, “Artificial Intelligence: Discerning a Christian Response,” *Perspect. Sci. Christ. Faith* 70, no. 1 (2018): 72–73.

[133] Julian Smith, “How Artificial Intelligence Helped Find Lost Cities,” iQ by Intel, March 20, 2018, https://iq.intel.com/how-artificial-intelligence-helped-find-lost-cities-of-ancient-middle-east/.

[134] “YesHEis: Life on Mission,” accessed October 13, 2018, https://us.yesheis.com/en/.

[135] Robert Barba, “Bank Of America Launches Erica Chatbot | Bankrate.Com,” Bankrate, accessed October 13, 2018, https://www.bankrate.com/banking/bank-of-america-boa-launches-erica-digital-assistant-chatbot/.
[136] “Questions about Jesus Christ,” GotQuestions.org, accessed October 13, 2018, https://www.gotquestions.org/questions_Jesus-Christ.html.
[137] Colin G. Walsh, Jessica D. Ribeiro, and Joseph C. Franklin, “Predicting Risk of Suicide Attempts Over Time Through Machine Learning,” *Clinical Psychological Science* 5, no. 3 (May 2017): 457–69, https://doi.org/10.1177/2167702617691560.
[138] Casey Cep, “Big Data for the Spirit,” *The New Yorker*, August 5, 2014, https://www.newyorker.com/tech/annals-of-technology/big-data-spirit.
[139] J. Nathan Matias, “AI in Counseling & Spiritual Care,” *AI and Christianity* (blog), November 2, 2017, https://medium.com/ai-and-christianity/ai-in-counseling-spiritual-care-e324d9aea3b0.
[140] Andrew Spicer, “Universal Basic Income and the Biblical View of Work,” Institute For Faith, Work & Economics, September 20, 2016, https://tifwe.org/universal-basic-income-biblical-view-of-work/.
[141] J. Nathan Matias, “How Will AI Transform Work, Creativity, and Purpose?,” *Medium* (blog), October 27, 2017, https://medium.com/ai-and-christianity/how-will-ai-transform-work-creativity-and-purpose-a8c78aa3368e.
[142] “AI Safety Myths,” Future of Humanity Institute, accessed October 13, 2018, https://futureoflife.org/background/aimyths/.
[143] Sherry Turkle, ed., *Simulation and Its Discontents*, Simplicity (Cambridge, Mass: The MIT Press, 2009).
[144] Wendy Lesser, *You Say to Brick: The Life of Louis Kahn*, 2018.
[145] Hilary Andersson Cellan-Jones Dave Lee, Rory, “Social Media Is ‘deliberately’ Addictive,” July 4, 2018, sec. Technology, https://www.bbc.com/news/technology-44640959.
[146] Katy Waldman, “Facebook’s Unethical Experiment,” *Slate*, June 28, 2014, http://www.slate.com/articles/health_and_science/science/2014/06/facebook_unethical_experiment_it_made_news_feeds_happier_or_sadder_to_manipulate.html; Inder M. Verma, “Editorial Expression of Concern: Experimental Evidence of Massivescale Emotional Contagion through Social Networks,” *Proceedings of the National Academy of Sciences* 111, no. 29 (July 22, 2014): 10779–10779, https://doi.org/10.1073/pnas.1412469111.
[147] Roger Parloff, “Why Deep Learning Is Suddenly Changing Your Life,” *Fortune* (blog), accessed October 14, 2018, http://fortune.com/ai-artificial-intelligence-deep-machine-learning/.
[148] Gordon Haff, “Data vs. Models at the Strata Conference,” CNET, March 2, 2012, https://www.cnet.com/news/data-vs-models-at-the-strata-conference/.
[149] Ben Quinn, “Google given Access to Healthcare Data of up to 1.6 Million Patients,” *The Guardian*, May 3, 2016, sec. Technology, https://www.theguardian.com/technology/2016/may/04/google-deepmind-access-healthcare-data-patients.
[150] Climate Home News and part of the Guardian Environment Network, “‘Tsunami of Data’ Could Consume One Fifth of Global Electricity by 2025,” *The Guardian*, December 11, 2017, sec. Environment, https://www.theguardian.com/environment/2017/dec/11/tsunami-of-data-could-consume-fifth-global-electricity-by-2025.
[151] Richard Evans and Jim Gao, “DeepMind AI Reduces Google Data Centre Cooling Bill by 40%,” DeepMind, July 20, 2016, https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/.
[152] OpenAI, “AI and Compute,” OpenAI Blog, May 16, 2018, https://blog.openai.com/ai-and-compute/.
[153] Pete Warden, “Why the Future of Machine Learning Is Tiny,” Pete Warden’s Blog, June 11, 2018, https://petewarden.com/2018/06/11/why-the-future-of-machine-learning-is-tiny/.
[154] Scott Hawley, “Learning Room Shapes,” May 4, 2017, https://drscotthawley.github.io/Learning-Room-Shapes/.
[155] Rebecca Fiebrink, *Wekinator: Software for Real-Time, Interactive Machine Learning*, 2009, http://www.wekinator.org/.
[156] Jacques Ellul, *The Technological Society*, trans. John Wilkinson, A Vintage Book (New York, NY: Alfred A. Knopf, Inc. and Random House, Inc., 1964).
[157] Jacques Ellul, *The Meaning of the City*, trans. Dennis Pardee, Jacques Ellul Legacy (Wipf & Stock Pub, 2011).
[158] Louise Matsakis, Andrew Thompson, and Jason Koebler, “Google’s Sentiment Analyzer Thinks Being Gay Is Bad,” *Motherboard* (blog), October 25, 2017, https://motherboard.vice.com/en_us/article/j5jmj8/google-artificial-intelligence-bias.

[159] Jeffrey Dastin, “Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women.” *Reuters*, October 10, 2018.
[160] Cathy O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, First edition (New York: Crown, 2016).
[161] “Bias Is AI’s Achilles Heel. Here’s How To Fix It,” accessed October 15, 2018, https://www.forbes.com/sites/jasonbloomberg/2018/08/13/bias-is-ais-achilles-heel-heres-how-to-fix-it/#72205cac6e68.
[162] Lucas Dixon et al., “Measuring and Mitigating Unintended Bias in Text Classification,” 2018.
[163] Casey, Farhangi, and Vogl, “Rethinking Explainable Machines: The GDPR’s ‘Right to Explanation’ Debate and the Rise of Algorithmic Audits in Enterprise.”
[164] Alex Campolo et al., “AI Now 2017 Report” (AI Now Institute, 2017).
[165] “Understanding the ‘Black Box’ of Artificial Intelligence,” Sentient Technologies Holdings Limited, January 10, 2018, https://www.sentient.ai/blog/understanding-black-box-artificial-intelligence/.
[166] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge: Harvard University Press, 2015).
[167] Riccardo Guidotti et al., “Local Rule-Based Explanations of Black Box Decision Systems,” *ArXiv:1805.10820 [Cs]*, May 28, 2018, http://arxiv.org/abs/1805.10820.
[168] Tao Lei, Regina Barzilay, and Tommi Jaakkola, “Rationalizing Neural Predictions,” *ArXiv Preprint ArXIv:1606.04155 [Cs.CL]*, June 13, 2016, https://arxiv.org/abs/1606.04155.
[169] Weiming Xiang, Hoang-Dung Tran, and Taylor T. Johnson, “Output Reachable Set Estimation and Verification for Multilayer Neural Networks,” *IEEE Transactions on Neural Networks and Learning Systems*, no. 99 (2018): 1–7.
[170] Margaret Boden et al., “Principles of Robotics” (The United Kingdom’s Engineering and Physical Sciences Research Council (EPSRC), 2011).
[171] Kristen Stubbs, Pamela J. Hinds, and David Wettergreen, “Autonomy and Common Ground in Human-Robot Interaction: A Field Study,” *IEEE Intelligent Systems* 22, no. 2 (2007).
[172] Robert H. Wortham and Andreas Theodorou, “Robot Transparency, Trust and Utility,” *Connection Science* 29, no. 3 (2017): 242–48.
[173] Campolo et al., “AI Now 2017 Report.” AI Now Institute*,* 2017.
[174] Wortham, Theodorou, and Bryson, “What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems.”
[175] Robert H Wortham, “Using Other Minds: Transparency as a Fundamental Design Consideration for Artificial Intelligent Systems” (Ph.D. Thesis, University of Bath, 2018), https://researchportal.bath.ac.uk/en/publications/using-other-minds-transparency-as-a-fundamental-design-considerat.
[176] Erik T. Mueller, “Transparent Computers: Designing Understandable Intelligent Systems,” *Erik T. Mueller, San Bernardino, CA*, 2016.
[177] Boris Delibasic et al., “White-Box or Black-Box Decision Tree Algorithms: Which to Use in Education?,” *IEEE Transactions on Education* 56, no. 3 (August 2013): 287–91, https://doi.org/10.1109/TE.2012.2217342.
[178] Laurens van der Maaten and Geoffrey Hinton, “Visualizing Data Using T-SNE,” *Journal of Machine Learning Research* 9, no. Nov (2008): 2579–2605.
[179] Adam W. Harley, “An Interactive Node-Link Visualization of Convolutional Neural Networks,” in *International Symposium on Visual Computing* (Springer, 2015), 867–77.
[180] Wortham, Theodorou, and Bryson, “What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems.”
[181] Olah, Mordvintsev, and Schubert, “Feature Visualization.”
[182] “Media for Thinking the Unthinkable,” accessed October 15, 2018, http://worrydream.com/MediaForThinkingTheUnthinkable/.
[183] Sandra Wachter, Brent Mittelstadt, and Chris Russell, “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR,” 2017.
[184] Bryson and Kime, “Just an Artifact: Why Machines Are Perceived as Moral Agents.”
[185] Shannon Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (New York, NY: Oxford University Press, 2016).

[186] Brent Waters, *Christian Moral Theology in the Emerging Technoculture: From Posthuman Back to Human*, Ashgate Science and Religion Series (Farnham, Surrey ; Burlington: Ashgate, 2014).

[187] Will Geary, "If Neural Networks Were Called 'Correlation Machines' I Bet There Would Be Less Confusion about Their Use and Potential.," Tweet, *@wgeary* (blog), July 13, 2018, https://twitter.com/wgeary/status/1017754723313770498.

[188] Xi Cheng et al., "Polynomial Regression As an Alternative to Neural Nets," *ArXiv:1806.06850 [Cs, Stat]*, June 13, 2018, http://arxiv.org/abs/1806.06850.

[189] Shaily Kumar, "The Differences Between Machine Learning And Predictive Analytics," *D!Gitalist Magazine*, March 15, 2018, https://www.digitalistmag.com/digital-economy/2018/03/15/differences-between-machine-learning-predictive-analytics-05977121.

[190] Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect*, First edition (New York: Basic Books, 2018).

[191] Judea Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge, U.K. ; New York: Cambridge University Press, 2000).

[192] "Bayesian Network," *Wikipedia*, October 11, 2018, https://en.wikipedia.org/w/index.php?title=Bayesian_network&oldid=863587945.

[193] Hartnett, "To Build Truly Intelligent Machines, Teach Them Cause and Effect."

[194] Judea Pearl, "The Seven Tools of Causal Inference with Reflections on Machine Learning," Technical Report, Communications of Association for Computing Machinery., July 2018, http://ftp.cs.ucla.edu/pub/stat_ser/r481.pdf.

[195] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford, United Kingdom ; New York, NY: Oxford University Press, 2016).

[196] Dario Amodei et al., "Concrete Problems in AI Safety," *ArXiv:1606.06565 [Cs]*, June 21, 2016, http://arxiv.org/abs/1606.06565.

[197] Allan Dafoe, "AI Governance: A Research Agenda" (Oxford, UK: Future of Humanity Institute, University of Oxford, August 27, 2018), http://www.fhi.ox.ac.uk/govaiagenda.

[198] Christina Colclough, "Putting People and Planet First: Ethical AI Enacted" (Conference on AI: Intelligent machines, smart policies, Paris: OECD, 2017), http://www.sipotra.it/wp-content/uploads/2018/09/AI-INTELLIGENT-MACHINES-SMART-POLICIES.pdf.