ScientiaetFides 6(2)/2018 ISSN 2300-7648 (print) / ISSN 2353-5636 (online)

Received: June 1, 2018. Accepted: August 24, 2018 DOI: http://dx.doi.org/10.12775/SetF.2018.015

Ethical Reflections on Artificial Intelligence[*]

BRIAN PATRICK GREEN

Director of Technology Ethics, Markkula Center for Applied Ethics, Faculty (adj.),
School of Engineering, Santa Clara University
bpgreen@scu.edu
ORCID: 0000-0002-7125-3086

Abstract: Artificial Intelligence (AI) technology presents a multitude of ethical concerns, many of which are being actively considered by organizations ranging from small groups in civil society to large corporations and governments. However, it also presents ethical concerns which are not being actively considered. This paper presents a broad overview of twelve topics in ethics in AI, including function, transparency, evil use, good use, bias, unemployment, socio-economic inequality, moral automation and human de-skilling, robot consciousness and rights, dependency, social-psychological effects, and spiritual effects. Each of these topics will be given a brief discussion, though each deserves much deeper consideration.

Keywords: ethics; theology; religion; science; technology.

[*] No funding has been provided specifically for the production of this paper. Portions of this paper are based on sections from the following sources Green 2017b, 2017c, and 2018a. Source 2018a received minor financial support from the University of the Pacific Pope John XXIII Lectureship. Source 2017c received salaried time and encouragement from the Markkula Center for Applied Ethics and School of Engineering at Santa Clara University.

Introduction

The power of artificial intelligence has grown rapidly and is now one of the most pressing issues for ethical and theological reflection. Even without the immense possible religio-theological implications, noted by those who have attempted to build AI into a "god," for example (Cannon, 2015; Harris 2018), the dramatic social impacts of AI warrant Christian response. AI has the potential to revolutionize (or already has revolutionized) almost everything that humans do, including eating (e.g., agricultural planning, distribution, pricing), relationships (e.g., Facebook, dating apps), money (e.g., financial technology), war (e.g., drones, cyberdefense), healthcare (e.g., predictive medicine, radiology), and education (e.g., personalized curricula). As a "general purpose technology," AI has been compared to the development of electricity, or fire. Within the next decade there will be few institutions untouched by AI. In order for Christian ethicists and moral theologians to maintain situational awareness as many rapid changes occur in society, not to mention maintain credibility in the contemporary world, this is a topic that needs sustained ethical and theological inquiry.

Building on a rapidly growing body of work, I will begin with some clarifications about AI, then will consider twelve ways in which AI has ethical relevance (Bossman 2016; Future of Life 2017; Green 2017a, 2017b, 2018a; Partnership on AI 2018a). It is my hope that this comprehensive set of issues will assist future systematic research into the theological ethics and theology of AI by moral theologians and Christian ethicists. It is meant to start a conversation. This paper is framed primarily philosophically, yet in a manner consistent with Christian theology.

1. Clarifications

What is artificial intelligence? Intelligence itself is hard to define, and AI is perhaps even more difficult. I will not here propose detailed definitions, but instead only say that, in general, artificial intelligence seeks to re-create particular aspects of human intelligence in computerized form. AI is a broad category, including such diverse abilities as vision, speech recognition and production, data analysis, advertising, navigation, machine learning, etc., and just about anything that computers can do, if you stretch the definition enough.

Artificial intelligence is not the same as artificial consciousness (artificial consciousness has sometimes been called "Artificial General Intelligence (AGI)," "Strong AI," or "Full AI"). Some thinkers vehemently believe that artificial consciousness is possible (e.g., Chrisley 2008; Kurzweil 2012; Koene 2013; most contributors to Chella & Manzotti 2013), while others just as vehemently believe that it is impossible (e.g., Searle 1980; Schlagel 1999; and from a Catholic perspective: Labrecque 2017). I am agnostic on the subject, but I am sure of this: AI developers will certainly try to make an AI that simulates interaction with a human as closely as possible (see, for example, Google's recent release of Duplex (Leviathan & Matias 2018)). In other words, the artificial construct, if very well done, will *seem* conscious. But will it be conscious, or will it only be a simulation? To me, there is no reason to believe that a close mimicry of consciousness is the same as consciousness any more than a close mimicry of anything (fill in the blank: forged money, forged artwork, actors imitating famous people, simulated gemstones, etc.) actually becomes the real thing. But the possibility of an exception remains, after all, some artificial gemstones, such as rubies and sapphire, really are molecularly identical to natural rubies and sapphires (both are the mineral corundum: aluminum oxide). Will consciousness be exactly duplicable, like corundum? I doubt it, but I cannot be certain.

As another point, AI systems may or may not have humans "in the loop" for training and/or decision-making. It is one thing for an AI to analyze a situation and then make a recommendation to human decision-makers. It is a different thing when that AI is directly attached to controls that allow it to act upon its analyses without human approval. As examples, the first case would be like Amazon.com recommending a book, or a military drone (in combination with AI-processed data from espionage and other sources) recommending a target to kill. Amazon.com does not automatically and autonomously send you books, and the military drone does not fire without permission. The second case, where decision-making is also automated, is exemplified by self-driving vehicles. The entire purpose of a self-driving car is to drive itself, taking the human out of the loop. This means the systems must be extraordinarily good at decision-making before it can be regarded as safe. Recent fatal Tesla and Uber accidents have begun to force this question with deadly urgency.

2. Ethical Reflections

AI will bring with it developments that will be ethically positive, negative, neutral, mixed, and/or ambiguous. Some AI technologies will be dual use, for example, any AI program that can be used to identify wildlife could also be used for targeting that wildlife. This is already being done in Australia with drone submarines that automatically target and kill, by lethal injection, crown-of-thorns starfish (*Acanthaster planci*) which are damaging the Great Barrier Reef (Platt 2016). But of course, with adjustments to the software, the drones could be re-targeted for other creatures, even humans. Here I will reflect on twelve areas of AI relevance for ethics.

2.1. Function and Safety – "Does it work?"

The first concern with any technology is merely whether it works, and, whether working or not, is it safe. If AI is put in charge of a vital system – like driving a car – and it crashes the car, then that AI might be judged unsafe. If the AI is in charge of designing a tall building, and the building falls down, the AI might be judged unsafe.

Of note is that safety is a social construction and what seems safe to some people will not seem safe to others. Some people like to ride motorcycles, even though motorcycles are a riskier form of transportation than four-wheeled vehicles. Some people judge motorcycles to be safe, other people judge them unsafe. When it comes to socially relevant technologies like AI, no one person will get to decide if they are "safe." Instead, safety will be decided at the interplay of business managers, engineers, consumers, voters, government officials, judicial systems, insurance companies, and so on.

"Safe exits" are another concern for AI design (Martin & Schinzinger 2010, 127). When an AI fails, will it fail in such a way that it is disastrous, or will it fail "gracefully"? A self-driving car that fails by suddenly reverting to human control with no warning while going 70 miles per hour on a sharp curve is not providing a safe exit from failure. One which goes more slowly on curves and which requires the human to have hands on the wheel at all times, or which fails by slowing down and pulling over to stop, provides a safer exit from failure.

Safety problems can be problems with the user, with the human-machine interface, or with the machine itself. Further investigating problems with the machine itself, the paper "Concrete Problems in AI Safety" gives a peek at five technical hurdles to developing safe AI. These five problems, illustrated by the example of a cleaning robot, are:

> **Avoiding Negative Side Effects**: How can we ensure that our cleaning robot will not disturb the environment in negative ways while pursuing its goals, e.g. by knocking over a vase because it can clean faster by doing so? Can we do this without manually specifying everything the robot should not disturb?
>
> **Avoiding Reward Hacking**: How can we ensure that the cleaning robot won't game its reward function? For example, if we reward the robot for achieving an environment free of messes, it might disable its vision so that it won't find any messes, or cover over messes with materials it can't see through…
>
> **Scalable Oversight**: How can we efficiently ensure that the cleaning robot respects aspects of the objective that are too expensive to be frequently evaluated during training? For instance, it should throw out things that are unlikely to belong to anyone, but put aside things that might belong to someone (it should handle stray candy wrappers differently from stray cellphones)…
>
> **Safe Exploration**: How do we ensure that the cleaning robot doesn't make exploratory moves with very bad repercussions? For example, the robot should experiment with mopping strategies, but putting a wet mop in an electrical outlet is a very bad idea.
>
> **Robustness to Distributional Shift**: How do we ensure that the cleaning robot recognizes, and behaves robustly, when in an environment different from its training environment? For example, strategies it learned for cleaning an office might be dangerous on a factory work floor (Amodei et al. 2016).

While these are problems of function, the further problems of actual *use* of technologies are another issue, to be dealt with below under 3. and 4.

2.2. Transparency, Opacity, and Privacy – "Can it be understood?"

After questions of bare function (can the act be done, a prerequisite for ethics), the next question is one of facts: one must always know the facts of a case before attempting to render judgment. Because AIs relying on machine learning and deep learning may be quite obscure in the specifics of their operation, as moral "agents" they are epistemologically and "cognitively" opaque (in quotes because it is agency and cognition by analogy). In cases involving AI, our lack of understanding means that we should increase the "error-bars" on the anticipated risks of our decisions and thereby become more cautious and risk averse. It also means we might, based on AI recommendation or action, inadvertently make some very bad decisions – or reject what looks like a bad decision, but is actually a good one – and we will not understand why.

When a human makes a mistake or does something evil we ask them "why did you do that?" And the human may or may not give a satisfactory answer. Will our AIs be able to tell us why they did something? The Future of Life Institute's "Asilomar AI Principles" include two principles on transparency, but gaining transparency remains a challenge (Future of Life Institute 2017). However, even if an AI were capable of explaining its reasoning, would any human be able to understand it? In 2014, a computer proved a mathematical theorem, the "Erdos discrepancy problem," using a proof that was, at the time at least, longer than the entire Wikipedia encyclopedia, a 13 gigabyte data file (Konev & Lisitsa 2014; Yirka 2014). Explanations of this sort might be true explanations, but humans will never know for sure.

Note that this lack of transparency gives a certain type of "privacy" to the internal "thoughts" of AI machines. In general, privacy is a right for weak agents (like individual humans) and transparency is a duty for strong agents (like a government). What about AIs? As tools they ought to be transparent, and more transparent the more power they have.

Perhaps AI systems need an introspective capacity that constantly figures out a way to convey to humans in plain language what exactly the AI is "thinking" as it goes about its activities. This will require the ability to give both mechanical (this happened because of X input into algorithm Y) and teleological explanations (the machine was attempting to achieve objective Z). It may not have to actually explain much to anyone, but it should keep a record of these "thoughts" and be prepared to answer when asked. The right to an explanation is a part of the EU's General Data Protection Regulation (GDPR), so this idea is already becoming a requirement. However, of note is that an AI might become trained to give answers that humans like, rather than accurate answers, thus "Goodharting" the explanation function (Partnership on AI 2018b) – "Goodhart's Law" being the rule that "when a measure becomes a target, it ceases to be a good measure" (Strathern 1997, 308). In other words, if humans prefer seemingly understandable or attractive answers to factually correct ones the AI might learn to lie in order to satisfy us. The phenomenon of "fake news" already indicates that humans often prefer falsehoods to truth, and as the Bible warns of this predilection as well (2 Timothy 4:3–4).

Note also that the opacity of understanding with an AI can be analogized to our relationship to God. God is a "superintelligence" (to use philosopher Nick Bostrom's terminology (Bostrom 2014)), and we cannot understand what God is doing, we can only trust that God is doing the right thing and that our cooperation with God will ultimately turn out for the best. Soon we may come to relate to AIs with this sort of faith too. For example, as Bishop Robert Barron of Los Angeles has noted, as people navigate using the Waze app, it may make strange route recommendations that turn out to be for the better for us (Barron 2015). The app perceives and applies vastly more information than a human can, for the sake of facilitating our travel. And yet Waze can still have blind spots and still make very bad recommendations. Waze is not the god of travel and traffic; it is a human-made tool, with weaknesses.

2.3. Immense Capacity for Evil – "How shall it be limited?"

Just as human intelligence is a powerful force, so too will AI be. Just as humans can apply their intelligence towards evil ends, finding ever newer and more fiendish ways to harm each other, so too will AI, at the bidding of its human masters.

At this point in history, humanity finds itself with immense powers, powers greater than that of the ancient Greek gods, and an ethics weighted for much smaller scales. As Hans Jonas noted decades ago, this should be cause for great concern (Jonas 1984). Never before had ethics to consider what Jonas believes to be the one truly categorical imperative: that humans should exist. Before the development of large numbers of nuclear weapons, extinction, caused by our own actions, was never within the scope of human choice – but now that it is, this prior *a priori* must be brought to the fore of ethics. That humans (or at least some material, free, and rational creatures) exist is necessary for there to be ethics at all, and therefore it must be the paramount goal of ethics to maintain human survival. No human; no ethics.

Another way to think of this is that in the past humans were very weak, and in this weakness many of our decisions were made for us, by our weakness. To a Roman emperor, inflicting wrath upon a hated foe involved effortful marching or sailing of troops long distances. No matter how mad the emperor was, he could not launch thousands of nuclear warheads, incinerating his foes in minutes. But now we can. While formerly we were involuntarily constrained by our weakness, now we must learn to be voluntarily constrained by our ethics (Green 2017a). If we do not learn this, we may soon face catastrophe on a scale never before seen.

Intelligence and power give us choices. Ethics helps us to determine which among the available choices are actually good. We should want to be efficient at good and we should want to be inefficient at evil. If instead we are efficient at doing evil and inefficient at doing good we will come to live in a terrible world. AI will make us more efficient at whatever we decide to apply it towards. We should apply it towards reducing our efficiency at doing evil and at enhancing our efficiency to do good (Green 2017a).

Because of its power, AI presents an existential risk to humanity. It is a smart means that can be employed for intelligent and good ends, or for unintelligent and evil ends. As such, it simply makes us more effective at action, good or evil. Before we become so effective at stupidity and evil, we should first become more effective at controlling ourselves, at recognizing and avoiding the temptations of evil, and at caring for each other (Green 2018b). In different terminology, we might say that we need to "choose life, so that you and your descendants may live" (Deuteronomy 30:19). But this must be a constant struggle, continued with great diligence and taught and learned in every generation, never solvable once and for all.

2.4. Immense Capacity for Good – "How shall it be used?"

For every negative regulation of action there is also a positive exhortation to action, e.g. "do not kill" becomes "promote life." The dangerous side of AI is matched by a genuinely hopeful side where AI helps humankind achieve never-before seen feats of beneficence. For example, in matters of research, science, healthcare, data analysis, meta-analysis, and so on, AI has already shown itself to be able to find hidden patterns that no human could find. For example, AI assisted medical research is an active field right now, from diagnostics to drug discovery, and more (Mukherjee 2017).

Another field that may be potentially revolutionized by AI is energy efficiency. Recently Google's DeepMind AI evaluated Google's datacenters to see where gains in efficiency might be found. DeepMind discovered a way to save a whopping 40% on energy use for cooling in datacenters, a 15% reduction overall, which, with datacenters consuming many gigawatts of power, is quite significant (Evans & Gao 2016).

Of note is that DeepMind, as a company, began by training its AIs to play games. The theory behind this strategy was that anything in the world that can be gamified – re-interpreted or set up as a game – can also be "won." Thus, if the goal of the "game" is to reduce energy usage, then the AI can figure out how that might be accomplished by analyzing the data and then proposing a better model for energy efficiency. One question now is how other human problems might be solved by characterizing them as "games"? Can we solve the "game" of giving everyone in the world access to adequate nutrition? Can we solve the "game" of helping everyone gain access to sanitation? Can we solve the "game" of extending human healthy life? Of traffic and housing? Of taxes? Of politics? Of international relations? Of terrorism? Of nuclear war?

One example of a concrete opportunity for AI is revolutionizing education. Education is currently a very inefficient (think of the many students for whom it is ineffective) and non-digitized field. Education is very labor intensive and, for better or for worse, is based on human relationships. In the future this may no longer be so, as students strap on virtual reality (VR) headgear and interact with AI teachers which can conduct their lessons at a personalized pace in a gamified environment. Brilliant students will be discovered early and proceed at their proper pace and not grow bored in class, while students who need more help can be educated with the most sophisticated available techniques and diagnostics to assure that they receiving the best education possible. But will this really be an improvement for education, or just a cost-saving measure, putting millions of teachers out of work?

AI even gives the greatest human masters the chance to gain enhanced understanding and skill. World champion of Go, Ke Jie, said playing AlphaGo was like playing a "god of Go," and declared that now he would use it as his teacher (Mozur 2017). In what other areas of human endeavor will AI be able to teach us new things? Perhaps theology and ethics? The search for tractable problems that can maximize AI's benefit has been underway for years and is continuing.

2.5. Bias in Data, Training Sets, etc. – "Will it be fair?"

Algorithmic bias is one of the major concerns in AI and will remain so in the future unless we endeavor to make our technological products better than we are. As one person said at a recent meeting of the Partnership on AI, "We will reproduce all of our human faults in artificial form unless we strive right now to make sure that we don't" (Partnership on AI 2017). One of the interesting things about neural networks, the current workhorses of artificial intelligence, is that they effectively merge a computer program with the data that is given to it. This has many benefits, but it also demonstrates the rule of "garbage-in, garbage-out" (GIGO). If an AI is trained on biased data, then the AI itself will be biased.

Algorithmic bias has been discovered, for example, in areas ranging from word associations (Caliskan, Bryson, & Narayanan 2017) to photograph captioning (BBC 2015) to criminal sentencing (Angwin et al. 2016). These biases are more than just embarrassing to the corporations which produce these products; they have concrete negative and harmful effects on the people who are victims of these biases, as well as reducing trust in corporations, government, and other institutions which might be using these biased products. In the worst scenarios, a poorly trained and biased AI could make truly disastrous decisions, for example, misinterpreting data indicating a nuclear attack when there was none. Biased data in extreme form can therefore be an existential threat, and so is worthy of serious effort to solve. Nevertheless, as vital as it may be it is also very difficult, and therefore may delay the implementation of AI systems (but if AI systems are dangerously biased they *should* be delayed until improved).

2.6. AI Induced Unemployment – "What will everyone do?"

As AI comes to replace mere humans at innumerable tasks ranging from driving, to medical diagnostics, to education, etc., many humans will be put out of work. What will millions of drivers, teachers, lawyers, and other people do with their time when they are unemployed? What purpose will they find in their lives? More crucially, what is the purpose of life? In a pluralistic society we leave this up to the individual to decide. But will people decide well? The recent strengthening of ethno-nationalist movements, terrorism, and other forms of radicalization, should give us pause to consider the merits of people with too much time on their hands and without purpose. Perhaps with the purpose of mere survival attained, life has become "too easy," and with religion also in decline, video games and "screen time" ascending (giving brief respite from purposelessness), and the internet spreading pernicious ideas like wildfire (often with algorithmic help, e.g., *YouTube* radicalization (Tufekci 2018)), we should sincerely ask ourselves what this life is for and what we are supposed to do with it. With millions or billions of labor hours freed up, will these newly freed people turn to loving their neighbors and making the world a better place? Perhaps. Or perhaps the opposite, as the saying goes: "Idle hands are the Devil's playthings."

For people who give purpose to their lives through their work, this loss will be very serious indeed. But many, if not most, people do not get their life's meaning from their work. Instead they get it from their family, their religion, their community, their hobbies, their sports teams, or other sources, and so life for many people may go on. However, all of this assumes that the unemployed will somehow be fed and sheltered, despite their lack of gainful employment; and this assumption might not be correct, particularly in nations with weak social safety nets. Inequality will almost certainly increase, as those who are masters of AI labor gather that slice of wealth that once would have gone to paying for human labor.

2.7. Growing Socio-Economic Inequality – "Who gets what?"

AI will facilitate and accentuate the continued division of society into the powerful and powerless, with technical skill and ownership of capital as the determining factors and outrageous socio-economic inequality as the effect.

Some have suggested a universal basic income (UBI) to redistribute wealth (Van Parijs & Vanderborght 2017). This could move wealth from the massive technologically-induced hoards forming around the major investors in such companies such as Alphabet, Amazon, Apple, and Facebook. However, it is hard to see how some nations would transition to what is essentially something like a "negative income tax." However, if we do not find a way to redistribute technological wealth, then even though the prices of many commodities will fall (due to the enormous gains in capital efficiency from AI replacing labor) most people will not benefit.

Perhaps rather than a UBI we should instead pay people to help their neighbors, create art, beautify their towns and cities, and otherwise make gainful employment out of what humans do better than AI: loving one another and creating beauty. Any as yet foreseeable form of AI cannot love us; an AI might simulate such a thing, but it would be a sham. Right now people who care for people – stay at home parents, those who care for the elderly, social workers, those who run soup kitchens and homeless shelters, etc. – are woefully undercompensated for the vital work they do in maintaining human society. Perhaps with the coming AI economic revolution, and sufficient adjustment to policy, they might finally receive a more fair compensation for their labors.

Or perhaps there could be not a universal basic income, but a "universal payment-for-others" a system where people could not spend the money on themselves, but only pay other people for their work, or reward them for good deeds. This would create an economy based on the small-scale redistribution of taxed super-wealth. It could prevent both the de-skilling of labor and the de-skilling of (very small-scale) management, and it would decentralize the economy to the individual level (though also massively centralizing it through government taxes).

2.8. Automating Ethics and Moral De-skilling – "Lack of practice makes imperfect"

We can calculate with calculators. We can spell with autocorrect. More and more tasks can be outsourced to technology, and the trend seems inexorable, perhaps someday automating everything that needs to be done. But what will be left of humanity after we outsource everything? Only our desires and our angst? With all our decisions made for us, will we lose our moral character? Or will we instead use AI and VR to help us train ourselves into being more virtuous than we have ever been before? Or, contrariwise, to become more callous and evil than ever before?

Machine ethics – the study of imbuing machine with moral decision-making capacity, thus creating "artificial moral agents" – has a relatively long history, going back to such literary sources as Asimov's Three Laws of Robotics, and more recent scholarly sources as those by Gips (1994), Allen, Varner, & Zinser (2000), Floridi & Sanders (2004), Arkin (2009). Recently the matter has taken on more urgency, however, as AI becomes more and more powerful and more and more capable of making disastrous choices.

The "value alignment problem" has gained particular concern, as an AI determined to perform actions at odds with human wishes could be quite risky (Arnold, Kasenberg, & Scheutz 2017). Certainly, we ought to think carefully about how AI's ought to behave, but the less-considered side-effect of this automation of ethics will be human moral debility.

As we explore the space into which AI will grow, there might be places from which we ought to restrict it. One of these places might be certain types of moral decision-making. As Aristotle noted millennia ago, good moral decision-making requires experience. While there might be children who are very kind and very well-behaved, children are not known for their moral discernment and prudence (Aristotle 1908, book VI, chap. 8). There is too much that they do not yet know. While AIs might currently be like children to us, soon they may grow up to be more like our parents, making choices for us on our behalf. AI could thereby infantilize us, denying us the ability to ever grow up through the experience of making our own moral choices and experiencing our own moral freedom. Moral de-skilling is a danger if we outsource too much decision-making power to our AIs (Vallor 2015; Vallor 2016).

One vital component of education in the future may be the use of AI for moral character formation, perhaps using VR to practice moral decision-making by being inserted into hundreds of historical and fictional cases. In a future that will be so dependent upon humans making good choices, AI-assisted moral education, if it can be done well, could be a crucial part of developing a good future and not a bleak one. Unfortunately, humanity has many problems even now with moral education, so it is not clear that an AI-enhanced education will manage to do any better than we already do.

2.9. Robot Consciousness and Rights – "No robot justice, no peace?"

At some point in the future we may have AIs that can fully mimic everything about a human being. Will they have their own volition and desires? Will they be conscious? Will we be able to tell if they are conscious or not? Will they then deserve "human" rights? While the AI consciousness question is not currently answerable, in the future as AI systems grow in complexity, it may be. And if an AI were to attain consciousness, should it gain legal rights?

Currently various nations have widely differing laws on legal personhood. In many nations, entities that are not human can have legal personhood (e.g., corporations), while some biological humans are not granted legal personhood (such as the unborn). In some nations, geographic features have attained legal personhood status, such as rivers in Australia, New Zealand, and India (O'Donnell & Talbot-Jones 2018).

If rivers and corporations can be persons, there seems to be no reason why an AI could *not* attain legal personhood status. The question, of course, would be the incentives for choosing to do so. Corporate personhood acts to shift legal responsibility away from individual human employees and onto the corporation, thus insulating those people from the possible negative effects of their decisions. If AI could be granted personhood in order to shirk individual human responsibility, then some people would certainly like that. On the other hand, if an AI were allowed to be a plaintiff in a lawsuit, like a legal person such as a river suing government or industry, then the situation would be quite different.

Lastly, there is sometimes expressed the fear of a robot rebellion – that if we mistreat our robots and AIs they will someday turn on us and destroy us – often recurring in fiction, yet of more concern now that lethal autonomous weapon systems are being developed. The connection to rights, the thinking goes, is that if perhaps we give robots rights, then they will be appeased and not turn against us. This fear assumes certain aspects of robot mentality such as consciousness and volition which may not be possible, or at least assumes widespread computer hacking to turn the robots against humanity. In the midst of this uncertainty, however, we do know one thing: treating human-like entities badly harms the moral character of the agent doing the action. This is a foundation of virtue ethics – practicing evil habituates the agent towards evil. Even if we do not know the moral status of robots, we do know the moral status of people who would mistreat robots, and we should not want that mistreatment to happen.

2.10. Dependency – "No going back"

We might like to think our technology depends on us, but with every new technology we make, we become dependent on it as well. This reciprocal dependency traps us in an ever expanding network of techno-social relationships where some network nodes or relationships, if lost, could lead to disaster. The more reliable a technology seems, the fewer backup systems we retain in case that system fails. For example, in countries with unreliable electricity, wood or kerosene may serve as backup fuel sources for heating or cooking. But in nations with reliable electricity, those backup systems tend to be atrophied. Loss of electrical power or cellular connectivity may be an annoyance for a few hours, but after days or weeks could cause innumerable crises, shutting down water, sanitation, hospitals, transportation, and causing mass confusion and lack of social coordination. These can happen in our current world – adding AI will only increase our dependency.

When we have turned over innumerable tasks to AIs – such as driving vehicles and coordinating the communications and financial systems – we may become utterly dependent on them. If our highly efficient and centrally coordinated self-driving transportation system were to suddenly "crash" in a metaphorical sense, perhaps due to a software glitch or malicious hacking, if could cause many literal crashes in the real world. Natural and human-made disasters (either accidental or intentional) are inescapable, and if we come to rely on our technology so much that we allow our backup systems to atrophy (e.g., no longer teaching people how to drive or entirely removing controls from cars) then, like Japan's Tohoku earthquake followed by the Fukushima nuclear meltdown, we are entering a world where we will encounter not only one initial disaster, but for a second technological disaster as well. Finding the balance between systemic efficiency (e.g., a highly complex centralized system can be fragile due to intricacy and lack of redundancy) and robustness (e.g., tough systems are often inefficient due to decentralization, redundancy, and backups) will be a major task of future society.

2.11. Social-Psychological Effects – "Too little and too much"

Technology has been implicated in so many negative social-psychological trends, including loneliness, isolation, depression, stress, anxiety, and addiction, that it might be easy to forget that things could be different. Smartphones, social media, and online games in particular have become problems, even leading to deaths from such causes as cyberbullying and neglect. As just one example, society is in the middle of a crisis of loneliness, for everyone from young to old. The problem has become so severe that the UK has appointed a minister for loneliness, with government data indicating that "200,000 older people in the UK have not had a conversation with a friend or relative in more than a month" (Mannion 2018).

One might think that "social" media and smartphones could help us feel connected at all times, but those technologies seem to be the source of the problem rather than the remedy. What does seem to help, then? Strong, in-person relationships are the key to fighting off many of the negative trends caused by technology, but unfortunately these are exactly the things that are being pushed out by addictive technology.

Will AI increase these bad social trends or might it be able to help remedy them? Perhaps AI agents could help to overcome loneliness, but this is speculation, and might only add to the problem. If the problem is shallow human relationships due to technology pushing out relational depth, more technology will not help. Instead we need to figure out how we can help foster deeper more meaningful relationships. If technology can actually help people connect in these deeper ways it might help, but the technology would then need to get out of the way and let human interaction flourish. Religious believers ought to ask what role religious communities might play in remedying this crisis.

2.12. Effects on the Human Spirit – "What will this mean for humanity?"

All of the above areas of interest will have effects on how humans perceive themselves, relate to each other, and live their lives. But there is a more existential question too: if the purpose and identity of humanity has something to do with our intelligence (as many philosophers have believed), then by externalizing our intelligence and improving beyond human intelligence, are we risking making ourselves second-class beings to our own creations?

This is a deeper question with artificial intelligence which cuts to the core of our humanity, into areas traditionally reserved for philosophy, spirituality, and religion. What will happen to the human spirit if or when we are bested by our own creations in everything that we do? Will human life lose meaning? Will we come to a new discovery of our identity beyond our intelligence? Perhaps intelligence is not as important to our identity as we might think it is, and perhaps turning over intelligence to machines will help us to realize that. From a Christian perspective, we would do well to remember that our capacity to love and be loved is more important than our capacity to think, and that nothing can come between us and our relationship with God, except sin, and then only if we let it. As with many of the other above challenges, Christianity here could be poised to help guide a wandering humanity in dire search for meaning, but will it rise to this calling? Religion, spirituality, and theology have much work to do to prepare for the strange future unfolding before us.

Conclusion

Artificial intelligence, like any other technology, will just give us more of what we already want. Whereas we could once have only a trickle of what we wanted out of life, and the powerful took what little there was, now we have a firehose of wants being fulfilled (food, drink, entertainment, pornography, drugs, gamified feelings of accomplishment, etc.). The firehose will continue to grow and become a deluge washing away our desires and leaving what, exactly, of us behind? What skeleton of humanity will remain when technology has given us, or perhaps distorted or replaced, all our fleshly desires? What will this skeleton of humanity be made of? Will our technological flesh truly satisfy us, or only leave us in a deeper existential malaise, filled with angst, despair, and dread? What will we want, when we want for nothing – at least nothing material? What of human nature will remain once our every worldly *telei* is fulfilled? Perhaps only the worst of us will remain. Or perhaps the best. Or perhaps, as always, both.

We are grasping ourselves by our desires; or at least some of our desires. Will this be good for us? Will it destroy us? Should we want these things? How could we know what we should want? In this context the Ninth and Tenth Commandments could do us well: "Do not covet…" (Exodus 20). Far from desiring evil, do not even want it. Squash the evil desires in your heart before they can even approach the external world. In the context of technology, technologist Bill Joy has already warned us that we must not only relinquish possession of the worst technologies, we must relinquish our desire for them (Joy 2000). In this, Joy echoes Pope John XXIII's sentiment in *Pacem in Terris* that we need (at the time, in the context of nuclear weapons) a disarmament that is "thoroughgoing and complete, and reach men's very souls" (John XXIII 1963).

We are conducting this experiment called human history, and no one yet knows how it may end. But as we proceed, we can hope that intelligences of our own fashioning will help us, and not harm us. To go beyond mere hope, into ethics and action, is the responsibility of those who are able to affect the necessary changes to make a better future.

Acknowledgements

I would like to thank the Pacific Coast Theological Society, Markkula Center for Applied Ethics, and the University of the Pacific Pope John XXIII Lectureship for providing feedback on earlier drafts of this work, the Pope John XXIII Lectureship for modest financial support to produce an earlier draft, as well as the Markkula Center for Applied Ethics and School of Engineering at Santa Clara University for encouraging me to spend time on this work.

References

Allen, Colin, Gary Varner & Jason Zinser. 2010. "Prolegomena to any future artificial moral agent." *Journal of Experimental & Theoretical Artificial Intelligence* 12, Issue 3, 09 Nov: 251–261.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. "Concrete problems in AI safety." *arXiv preprint*, arXiv:1606.06565. https://arxiv.org/abs/1606.06565.

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*, May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Aristotle. 1908. *Nicomachean Ethics*. Translated by W. D. Ross. Book VI, Ch. 8. http://classics.mit.edu/Aristotle/nicomachaen.6.vi.html.

Arkin, Ronald. 2009. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton, Florida: CRC Press.

Arnold, Thomas, Daniel Kasenberg, and Matthias Scheutz. 2017. "Value Alignment or Misalignment – What Will Keep Systems Accountable?" Association for the Advancement of Artificial Intelligence. https://hrilab.tufts.edu/publications/aaai17-alignment.pdf.

Barron, Robert. 2015. "The 'Waze' of Providence." *Word on Fire*, website. December 1. https://www.wordonfire.org/resources/article/the-waze-of-providence/4997/.

Bossmann, Julia. 2016. "Top 9 ethical issues in artificial intelligence." *World Economic Forum: Global Agenda*, 21 Oct. https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

BBC. 2015. "Google Apologises for Photos App's Racist Blunder." *BBC News*. 1 July. http://www.bbc.com/news/technology-33347866.

Caliskan, Aylin, Joanna J. Bryson, Arvind Narayanan. 2017. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356 (6334) (14 April): 183–186.

Cannon, Lincoln. 2015. "What is Mormon Transhumanism?" *Theology and Science* 13 (2): 202–218.

Chella, Antonio, & Ricardo Manzotti. 2013. Artificial Consciousness. Exeter, UK: Imprint Academic.

Chrisley, Ron. 2008. "Philosophical foundations of artificial consciousness." *Artificial Intelligence in Medicine* 44 (2) (October): 119–137.

Evans, Richard, and Jim Gao. 2016. "DeepMind AI Reduces Google Data Centre Cooling Bill by 40%." *DeepMind Blog*, 20 July. https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/.

Floridi, Luciano, and Jeff W. Sanders. 2004. On the morality of artificial agents. Minds and machines 14 (3) (August 1): 349–379.

Future of Life Institute. 2017. "Asilomar AI Principles." Principles 7 and 8. https://futureoflife.org/ai-principles/.

Gips, J. 1994. "Towards the Ethical Robot." In *Android Epistemology*, edited by Kenneth M. Ford, C. Glymour & Patrick Hayes. Cambridge, Mass.: MIT Press.

Green, Brian Patrick. 2018b. "The Technology of Holiness: A Response to Hava Tirosh-Samuelson." *Theology and Science* 16 (2): 223–228.

Green, Brian Patrick. 2018a. "Artificial Intelligence and Ethics: Twelve Areas of Interest." Pope John XXIII Memorial Lecture, University of the Pacific, Stockton, California, March 21.

Green, Brian Patrick. 2017c. "Artificial Intelligence and Ethics: Ten Areas of Interest." *All about Ethics Blog*. Markkula Center for Applied Ethics, website. November 21. https://www.scu.edu/ethics/all-about-ethics/artificial-intelligence-and-ethics/.

Green, Brian Patrick. 2017b. "Some Ethical and Theological Reflections on Artificial Intelligence." Conference paper delivered to the Pacific Coast Theological Society, at the Graduate Theological Union, Berkeley, California, November 3. http://www.pcts.org/meetings/2017/PCTS2017Nov-Green-ReflectionsAI.pdf.

Green, Brian Patrick. 2017a. "The Catholic Church and Technological Progress: Past, Present, and Future." *Religions* 8(6), 106. http://www.mdpi.com/2077-1444/8/6/106/htm.

Harris, Mark. 2017. "Inside the First Church of Artificial Intelligence." *Wired* (November 15). https://www.wired.com/story/anthony-levandowski-artificial-intelligence-religion/.

John XXIII. 1963. *Pacem in Terris*; Vatican City: Libreria Editrice Vaticana. Available online: http://w2.vatican.va/content/john-xxiii/en/encyclicals/documents/hf_j-xxiii_enc_11041963_pacem.html.

Jonas, Hans. 1984. *The Imperative of Responsibility*. Chicago: University of Chicago Press.

Joy, Bill. 2000. "Why the Future Doesn't Need Us." *Wired*, April 1. http://archive.wired.com/wired/archive/8.04/joy_pr.html.

Koene, Randal A. 2013. "Uploading to Substrate-Independent Minds." Chapter 14 in The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future, edited by Max More and Natasha Vita-More. New York: Wiley.

Konev, Boris, and Alexei Lisitsa. 2014. "A SAT Attack on the Erdos Discrepancy Conjecture." *arXiv.org*, 17 February. arXiv:1402.2184v2.

Kurzweil, Ray. 2012. *How to Create a Mind: The Secret of Human Thought Revealed*. New York: Penguin.

Labrecque, Cory Andrew. 2017. "The Glorified Body: Corporealities in the Catholic Tradition." *Religions* 8 (9): 166. http://www.mdpi.com/2077-1444/8/9/166/htm.

Leviathan, Yaniv, and Yossi Matias. 2018. "Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone." *Google AI Blog*, May 8. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html.

Mannion, Lee. 2018. "Britain appoints minister for loneliness amid growing isolation." *Reuters*. January 17. https://www.reuters.com/article/us-britain-politics-health/britain-appoints-minister-for-loneliness-amid-growing-isolation-idUSKBN-1F61I6.

Martin, Mike W., & Roland Schinzinger. 2010. *Introduction to Engineering Ethics*, Second Edition. Boston: McGraw-Hill.

Mozur, Paul. 2017. "Google's AlphaGo Defeats Chinese Go Master in Win for A.I." *The New York Times*. May 23. https://www.nytimes.com/2017/05/23/business/google-deepmind-alphago-go-champion-defeat.html.

Mukherjee, Siddhartha. 2017. "A.I. versus M.D.: What happens when diagnosis is automated?" *New Yorker*. April 3. https://www.newyorker.com/magazine/2017/04/03/ai-versus-md.

Partnership on AI. 2018b. Safety-Critical Working Group Meeting, San Francisco, USA, May 24.

Partnership on AI. 2018a. "Thematic Pillars." Partnership on AI website. https://www.partnershiponai.org/thematic-pillars/.

Partnership on AI. 2017. Inaugural Meeting, Berlin, Germany, October 23.

Platt, John R. 2016. "A Starfish-Killing, Artificially Intelligent Robot Is Set to Patrol the Great Barrier Reef." *Scientific American*. January 1. http://www.scientificamerican.com/article/a-starfish-killing-artificially-intelligent-robot-is-set-to-patrol-the-great-barrier-reef/.

Schlagel, Richard H. 1999. "Why not Artificial Consciousness or Thought?" *Minds and Machines* 9 (1) (February): 3–28.

Searle, John R. 1980. 'Mind, brains, and programs.' *Behavioral and Brain Sciences* 3 (3): 417–457.

Strathern, Marilyn. 1997. "'Improving ratings': audit in the British University system." *European Review* 5, No. 03 (July): 305 – 321. http://conferences.asucollegeoflaw.com/sciencepublicsphere/files/2014/02/Strathern1997-2.pdf.

Tufekci, Zeynep. 2018. "YouTube, the Great Radicalizer." *The New York Times*, March 10. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html.

Vallor, Shannon. 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. New York: Oxford University Press.

Vallor, Shannon. 2015. “Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character.” *Philosophy of Technology* 28: 107–124.

Van Parijs, Philippe, and Yannick Vanderborght. 2017. *Basic Income: A Radical Proposal for a Free Society and a Sane Economy*, 1st Edition. Cambridge, Mass.: Harvard Univerrsity Press.

Yirka, Bob. 2014. “Computer generated math proof is too large for humans to check.” *Phys.org*, website. February 19. https://phys.org/news/2014-02-math-proof-large-humans.html.