Chapter 7

AI and Its Moral Concerns

Bohyun Kim
University of Rhode Island

Automating Decisions and Actions

The goal of artificial intelligence (AI) as a discipline is to create an artificial system—whether it be a piece of software or a machine with a physical body—that is as intelligent as a human in its performance, either broadly in all areas of human activities or narrowly in a specific activity, such as playing chess or driving.[1] The actual capability of most AI systems remained far below this ambitious goal for a long time. But with recent successes with machine learning and deep learning, the performance of some AI programs has started surpassing that of humans. In 2016, an AI program developed with the deep learning method, AlphaGo, astonished even its creators by winning four out of five Go matches with the eighteen-time world champion, Sedol Lee.[2] In 2020, Google's DeepMind unveiled Atari57, a deep reinforcement learning algorithm that reached superhuman levels of play in 57 classic Atari games.[3]

Early symbolic AI systems determined their outputs based upon given rules and logical inference. AI algorithms in these rule-based systems, also known as good old-fashioned AI (GOFAI), are pre-determined, predictable, and transparent. On the other hand, machine learning,

[1] Note that by 'as intelligent as a human,' I only mean AI at human-level performance in achieving a particular goal not general(/strong) AI. General AI—also known as 'artificial general intelligence (AGI)' and 'strong AI'—refers to AI with the ability to adapt to achieve any goals. By contrast, an AI system developed to perform only one or some activities in a specific domain is called a 'narrow (/weak) AI' system.

[2] AlphaGo can be said to be "as intelligent as humans," but only in playing Go, where it exceeds human capability. So, it does not qualify as general/strong AI in spite of its human-level intelligence in Go-playing. It is to be noted that general(/strong) AI and narrow(/weak) AI signify the difference in the scope of AI capability. General(/strong) AI is also a broader concept than human-like intelligence, either with its carbon-based substrate or with human-like understanding that relies on what we regard as uniquely human cognitive states such as consciousness, qualia, emotions, and so on. For more helpful descriptions of common terms in AI, see (Tegmark 2017, 39). For more on the match between AlphaGo and Sedol Lee, see (Koch 2016).

[3] Deep reinforcement learning is a type of deep learning that is goal-oriented and reward-based. See (Heaven 2020).

another approach in AI, enables an AI algorithm to evolve to identify a pattern through the so-called 'training' process, which relies on a large amount of data and statistics. Deep learning, one of the widely-used techniques in machine learning, further refines this training process using a 'neural network.'[4] Machine learning and deep learning have brought significant improvements to the performance of AI systems in areas such as translation, speech recognition, and detecting objects and predicting their movements. Some people assume that machine learning completely replaced GOFAI, but this is a misunderstanding. Symbolic reasoning and machine learning are two distinct but not mutually exclusive approaches in AI, and they can be used together (Knight 2019a).

With their limited intelligence and fully deterministic nature, early rule-based symbolic AI systems raised few ethical concerns.[5] AI systems that near or surpass human capability, on the other hand, are likely to be given the autonomy to make their own decisions without humans, even when their workings are not entirely transparent, and some of those decisions are distinctively moral in character. As humans, we are trained to recognize situations that demand moral decision-making. But how would an AI system be able to do so? Or, should they be? With self-driving cars and autonomous weapons systems under active development and testing, these are no longer idle questions.

The Trolley Problem

Recent advances of AI, such as autonomous cars, have brought new interest to *the trolley problem*, a thought experiment introduced by the British philosopher Philippa Foot in 1967. In the standard version of this problem, a runaway trolley barrels down a track where five unsuspecting people are standing. You happen to be standing next to a lever that switches the trolley onto a different track, where there is only one person. Those who are on either track will be killed if the trolley heads their way. Should you pull the lever, so that the runaway trolley would kill one person instead of five? Unlike a person, a machine does not panic or freeze and simply follows and executes the given instruction. This means that an AI-powered trolley may act morally as long as it is programmed properly.[6] The question itself remains, however. Should the AI-powered trolley be programmed to swerve or stay on course?

Different moral theories, such as virtue ethics, contractarianism, and moral relativism, take different positions. Here, I will consider utilitarianism and deontology. Since their tenets are relatively straightforward, most AI developers are likely to look towards those two moral theories for guidance and insight. Utilitarianism argues that the utility of an action is what makes an action moral. In this view, what generates the greatest amount of good is the most moral thing to do. If one regards five human lives as a greater good than one, then one acts morally by pulling the lever and diverting the trolley to the other track. By contrast, deontology claims that what determines whether an action is morally right or wrong is not its utility but moral rules. If an action is in accordance with those rules, then the action is morally right. Otherwise, it is morally

[4] Machine learning and deep learning have gained momentum because the cost of high-performance computing has significantly decreased and large data sets have become more widely available. For example, the data in the ImageNet contains more than 14 million hand-annotated images. The ImageNet data have been used for the well-known annual AI competition for object detection and image classification at large scale from 2010 to 2017. See `http://www.image-net.org/challenges/LSVRC/`.

[5] For an excellent history of AI research, see chapter 1, "What is Artificial Intelligence," of Boden 2016, 1-20.

[6] Programming here does not exclusively refer to a deep learning or machine learning approach.

wrong. If not to kill another human being is one of those moral rules, then killing someone is morally wrong even if it is to save more lives.

Note that these are highly simplified accounts of utilitarianism and deontology. The good in utilitarianism can be interpreted in many different ways, and the issue of conflicting moral rules is a perennial problem that deontological ethics grapples with.[7] For our purpose, however, these simplified accounts are sufficient to highlight the aspects in which the utilitarian and the deontological position appeal to and go against our moral intuition at the same time.

If a trolley cannot be stopped, saving five lives over one seems to be a right thing to do. Utilitarianism appears to get things right in this respect. However, it is hard to dispute that killing people is wrong. If killing is morally wrong no matter what, deontology seems to make more sense. With moral theories, things seem to get more confusing. Furthermore, consider the case in which one freezes and fails to pull the lever. According to utilitarianism, this would be morally wrong because it fails to maximize the greatest good, i.e. human lives. But how far should one go to maximize the good? Suppose there is a very large person on a footbridge over the trolley track, and one pushes that person off the footbridge onto the track, thus stopping the trolley and saving the five people. Would this count as a right thing to do? Utilitarianism may argue that. But in real life, many would consider throwing a person morally wrong but pulling the lever morally permissible.[8]

The problem with utilitarianism is that it treats the good as something inherently quantifiable, comparable, calculable, and additive. But not all considerations that we have to factor into moral decision-making are measurable in numbers. What if the five people on the track are helpless babies or murderers who just escaped from the prison? Would or should that affect our decision? Some of us would surely hesitate to save the lives of five murderers by sacrificing one innocent baby. But what if things were different and we were comparing five school children versus one baby or five babies versus one school child? No one can say for sure what is the morally right action in those cases.[9]

While the utilitarian position appears less persuasive in light of these considerations, deontology doesn't fare too well, either. Deontology emphasizes one's duty to observe moral rules. But what if those moral rules conflict with one another? Between the two moral rules, "do not kill a person" and "save lives," which one should trump the other? The conflict among values is common in life, and deontology faces difficulty in guiding how an intelligent agent is to act in a tricky situation such as *the trolley problem*.[10]

Understanding What Ethics Has to Offer

Now, let us consider AI-powered military robots and autonomous weapons systems since they present the moral dilemma in the trolley problem more convincingly due to the high stakes involved. Suppose that some engineers, following utilitarianism and interpreting victory as the ultimate good/utility, wish to program an unmanned aerial vehicle (UAV) to autonomously drop

[7] For an overview, see (Sinnott-Armstrong, 2019) and (Alexander and Moore, 2016).

[8] For an empirical study on this, see (Cushman, Young, and Hauser 2006). For the results of a similar survey that involves an autonomous car instead of a trolley, see (Bonnefon, Shariff, and Rahwan 2016).

[9] For an attempt to identify moral principles behind our moral intuition in different versions of the trolley problem and other similar cases, see (Thomson 1976).

[10] Some moral philosophers doubt the value of our moral intuition in constructing a moral theory. See (Singer 2005), for example. But a moral theory that clashes with common moral intuition is unlikely to be sought out as a guide to making an ethical decision.

bombs in order to maximize the chances of victory. That may result in sacrificing a greater number of civilians than necessary, and many will consider this to be morally wrong. Now imagine different engineers who, adopting deontology and following the moral principle of not killing people, program a UAV to autonomously act in a manner that minimizes casualties. This may lead to defeat on the battlefield, because minimizing casualties may not be always advantageous to winning a war. From these examples, we can see that philosophical insights from utilitarianism and deontology may provide little practical guidance on how to program autonomous AI systems to act morally.

Ethicists seek abstract principles that can be generalized. For this reason, they are interested in borderline cases that reveal subtle differences in our moral intuition and varying moral theories. Their goal is to define what is moral and investigate how moral reasoning works or should work. By contrast, engineers and programmers pursue practical solutions to real-life problems and look for guidelines that will help with implementing those solutions. Their focus is on creating a set of constraints and *if-then* statements, which will allow a machine to identify and process morally relevant considerations, so that it can determine and execute an action that is not only rational but also ethical in the given situation.[11]

On the other hand, the goal of military commanders and soldiers is to end a conflict, bring peace, and facilitate restoring and establishing universally recognized human values such as freedom, equality, justice, and self-determination. In order to achieve this goal, they must make the best strategic decisions and take the most appropriate actions. In deciding on those actions, they are also responsible for abiding by the principles of *jus in bello* and for not abdicating their moral responsibility, protecting civilians and minimizing harm, violence, and destruction as much as possible.[12] The goal of military commanders and soldiers, therefore, differs from those of moral philosophers or of the engineers who build autonomous weapons. They are obligated to make quick decisions in a life-or-death situation while working with AI-powered military systems.

These different goals and interests explain why moral philosophers' discussion on *the trolley problem* may be disappointing to AI programmers or military commanders and soldiers. Ethics does not provide an easy answer to the question of how one should program moral decision-making into intelligent machines. Nor does it prescribe the right moral decision in a battlefield. But taking this as a shortcoming of ethics is missing the point. The role of moral philosophy is not to make decision-making easier but to highlight and articulate the difficulty and complexity involved in it.

Ethical Challenges from Autonomous AI Systems

The complexity of ethical questions means that dealing with the morality of an action by an autonomous AI system will require more than a clever engineering or programming solution. The fact that ethics does not eliminate the inherent ambiguity in many moral decisions should not lead to the dismissal of ethical challenges from autonomous AI systems. By injecting the capacity for autonomous decision-making into machines, AI can fundamentally transform any given field. For example, AI-powered military robots are not just another kind of weapon. When widely deployed, they can change the nature of war itself. Described below are some of the significant ethical challenges that autonomous AI systems such as military robots present. Note that

[11] Note that this moral decision-making process can be modeled with a rule-based symbolic AI approach, a machine learning approach, or a combination of both. See Vincent Conitzer et al. 2017.

[12] For the principles of *jus in bello*, see International Committee of the Red Cross 2015.

in spite of these ethical concerns, autonomous AI systems are likely to continue to be developed and adopted in many areas as a way to increase efficiency and lower cost.

(a) Moral desensitization

AI-powered military robots are more capable than merely remotely-operated weapons. They can identify a target and initiate an attack on their own. Due to their autonomy, military robots can significantly increase the distance between the party that kills and the party that gets killed (Sharkey 2012). This increase, however, may lead people to surrender their own moral responsibility to a machine, thereby resulting in the loss of humanity, which is a serious moral risk (Davis 2007). The more autonomous military robots become, the less responsibility humans will feel regarding their life-or-death decisions.

(b) Unintended outcome

The side that deploys AI-powered military robots is likely to suffer fewer casualties itself while inflicting more casualties on the enemy side. This may make the military more inclined to start a war. Ironically, when everyone thinks and acts this way, the number of wars and the overall amount of violence and destruction in the world will only increase.[13]

(c) Surrender of moral agency

AI-powered military robots may fail to distinguish innocents from combatants and kill the former. In such a case, can we be justified in letting robots take the lives of other human beings? Some may argue that only humans should decide to kill other humans, not machines (Davis 2007). Is it permissible for people to delegate such a decision to AI?

(d) Opacity in decision-making

Machine learning is used to build many AI systems today. Instead of prescribing a pre-determined algorithm, a machine learning system goes through a so-called 'training' process to produce the final algorithm from a large amount of data. For example, a machine learning system may generate an algorithm that successfully recognizes cats in a photo after going through millions of photos that show cats in many different postures from various angles.[14] But the resulting algorithm is a complex mathematical formula and not something that humans can easily decipher. This means that the inner workings of a machine learning AI system and its decision-making process is opaque to human understanding, even to those who built the system itself (Knight 2017). In cases where the actions of an AI system can have grave consequences such as a military robot, such opacity becomes a serious problem.[15]

[13] (Kahn 2012) also argues that the resulting increase in the number of wars by the use of military robots will be morally bad.

[14] Google's research team created an AI algorithm that learned how to recognize a cat in 2012. The neural network behind this algorithm had an array of 16,000 processors and more than one billion connections. Unlabeled random thumbnail images from 10 million YouTube videos allowed this algorithm to learn to identify cats by itself. See Markoff 2012 and Clark 2012.

[15] This black-box nature of AI systems powered by machine learning has raised great concern among many AI researchers in recent years. This is problematic in all areas where these AI systems are used for decision-making, not just in military operations. The gravity of decisions made in a military operation makes this problem even more troublesome. Fortunately, some AI researchers including those in the US Department of Defense are actively working to make AI systems explainable. But until such research bears fruit and AI systems become fully explainable, their military use means accepting many unknown variables and unforeseeable consequences. See Turek n.d.

AI Applications for Libraries

Do these ethical concerns outlined above apply to libraries? To answer that, let us first take a look at how AI, particularly machine learning, may apply to library services and operations. AI-powered digital assistants are likely to mediate a library user's information search, discovery, and retrieval activities in the near future.

In recent years, machine learning and deep learning have brought significant improvement to natural language processing (NLP), which deals with analyzing large amounts of natural language data to make the interaction between people and machines in natural languages possible. For instance, Google Assistant's new feature 'duplex' was shown to successfully make a phone reservation with restaurant staff in 2018 (Welch 2018). Google's real-time translation capability for 44 different languages was introduced to Google Assist-enabled Android and iOS phones in 2019 (Rincon 2019).

As digital assistants become capable of handling more sophisticated language tasks, their use as a flexible voice user interface will only increase. Such digital assistants will be able to directly interact with library systems and applications, automatically interpret a query, and return results that they deem to be most relevant. Those digital assistants can also be equipped to handle the library's traditional reference or readers' advisory service. Integrated into a humanoid robot body, they may even greet library patrons at the entrance and answer directional questions about the library building.

Cataloging, abstracting, and indexing are other areas where AI will be actively utilized. Currently, those tasks are performed by skilled professionals. But as AI applications become more sophisticated, we may see many of those tasks partially or fully automated and handed over to AI systems. Machine learning and deep learning can be used to extract key information from a large number of documents or from information-rich visual materials, such as maps and video recordings, and generate metadata or a summary.

Since machine learning is new to libraries, there are a relatively small number of machine learning applications developed for libraries' use. They are likely to grow in number. Yewno, Quartolio, and Iris.ai are examples of the commercial products developed with machine learning and deep learning techniques.[16] Yewno Discover displays the connections between different concepts or works in library materials. Quartolio targets researchers looking to discover untapped research opportunities based upon a large amount of data that includes articles, clinical trials, patents, and notes. Similarly, Iris.ai helps researchers identify and review a large amount of research papers and patents and extracts key information from them. Kira identifies, extracts, and analyzes text in contracts and other legal documents.[17] None of these applications performs fully automated decision-making nor incorporates the digital assistant feature. But this is an area on which information systems vendors are increasingly focusing their efforts.

Libraries themselves are also experimenting with AI to test its potential for library services and operations. Some are focusing on using AI, particularly the voice user interface aspect of the digital assistant, in order to improve existing services. The University of Oklahoma Libraries have been building an Alexa application to provide basic reference service to their students.[18]

[16] See `https://www.yewno.com/education`, `https://quartolio.com/`, and `https://iris.ai/`.

[17] See `https://kirasystems.com/`. Law firms are adopting similar products to automate and expedite their legal work, and law librarians are discussing how the use of AI may change their work. See Marr 2018 and Talley 2016.

[18] University of Oklahoma Libraries are building an Alexa application that will provide some basic reference service to their students. Also, their PAIR registry attempts to compile all AI-related projects at libraries. See `https://pair.libraries.ou.edu`.

At the University of Pretoria Library in South Africa, a robot named 'Libby' already interacts with patrons by providing guidance, answering questions, conducting surveys, and displaying marketing videos (Mahlangu 2019).

Other libraries are applying AI to extract information from digital materials and automate metadata generation to enhance their discovery and use. The Library of Congress has worked on detecting features, such as railroads in maps, using the convolutional neural network model, and issued a solicitation for a machine learning and deep learning pilot program that will maximize the use of its digital collections in 2019.[19] Indiana University Libraries, AVP, University of Texas Austin School of Information, and the New York Public Library are jointly developing the Audiovisual Metadata Platform (AMP), using many AI tools in order to automatically generate metadata for audiovisual materials, which collection managers can use to supplement their archival description and processing workflows.[20]

Some libraries are also testing out AI as a tool for evaluating services and operations. The University of Rochester Libraries applied deep learning to the library's space assessment to determine the optimal staffing level and building hours. The University of Illinois Urbana-Champaign Libraries used machine learning to conduct sentiment analysis on their reference chat log (Blewer, Kim, and Phetteplace 2018).

Ethical Challenges from the Personalized and Automated Information Environment

Do these current and future AI applications for libraries pose ethical challenges similar to those that we discussed earlier? Since information query, discovery, and retrieval rarely involve life-or-death situations, stakes seem to be certainly lower. But an AI-driven automated information environment does raise its own distinct ethical challenges.

(i) Intellectual isolation and bigotry hampering civic discourse

Many AI applications that assist with information seeking activities promise a higher level of personalization. But a highly personalized information environment often traps people in their own so-called 'filter bubble,' as we have been increasingly seeing in today's social media channels, news websites, and commercial search engines, where such personalization is provided by machine learning and deep learning.[21] Sophisticated AI algorithms are already curating and pushing information feeds based upon the person's past search and click behavior. The result is that information seekers are provided with information that conforms and reinforces their existing beliefs and interests. Views that are novel or contrast with their existing beliefs are suppressed and become invisible without them even realizing.

Such lack of exposure to opposing views leads information users to intellectual isolation and even bigotry. Highly personalized information environments powered by AI can actively restrict ways in which people develop balanced and informed opinions, thereby intensifying and perpetuating social discord and disrupting civic discourse. Under such conditions, prejudices, discrim-

[19] See Blewer, Kim, and Phetteplace 2018 and Price 2019.

[20] The AMP wiki is `https://wiki.dlib.indiana.edu/pages/viewpage.action?pageId=531699941`. The Audiovisual Metadata Platform Pilot Development (AMPPD) project was presented at Code4Lib 2020 (Averkamp and Hardesty 2020).

[21] See Pariser 2012.

ination, and other unjust social practices are likely to increase, and this in turn will have more negative impact on those with fewer privileges. Intellectual isolation and bigotry has a distinctly moral impact on society.

(ii) Weakening of cognitive agency and autonomy

We have seen earlier that AI-powered digital assistants are likely to mediate people's information search, discovery, and retrieval activities in the near future. As those digital assistants become more capable, they will go beyond listing available information. They will further choose what they deem to be most relevant to users and proceed to recommend or autonomously execute the best course of action.[22] Other AI-driven features, such as extracting key information or generating a summary of a large amount of information, are also likely to be included in future information systems, and they may deliver key information or summaries even before the request is made based upon constant monitoring of the user's activities.

In such a scenario, an information seeker's cognitive agency is likely be undermined. Crucial to cognitive agency is the mental capacity to critically review a variety of information, judge what is and is not relevant, and interpret how they relate to other existing beliefs and opinions. If AI assumes those tasks, the opportunities for information seekers to exercise their own cognitive agency will surely decrease. Cognitive deskilling and the subsequent weakening of people's agency in the AI -powered automated information environment presents an ethical challenge because such agency is necessary for a person to be a fully functioning moral agent in society.[23]

(iii) Social impact of scholarship and research from flawed AI algorithms

Previously, we have seen that deep learning applications are opaque to human understanding. This lack of transparency and explainability raises a question of whether it is moral to rely on AI-powered military robots for life-or-death decisions. Does the AI-powered information environment have a similar problem?

Machine learning applications base their recommendations and predictions upon the patterns in past data. Their predictions and recommendations are in this sense inherently conservative. They also become outdated when they fail to reflect new social views and material conditions that no longer fit the past patterns. Furthermore, each data set is a social construct that reflects particular values and choices such as who decided to collect the data and for what purpose; who labeled data; what criteria or beliefs guided such labeling; what taxonomies were used and why (Davis 2020). No data set can capture all variables and elements of the phenomenon that it describes. Furthermore, data sets used for training machine learning and deep learning algorithms may not be representational samples for all relevant subgroups. In such a case, an algorithm trained by such a data set will produce skewed results. Creating a large data set is also costly. Consequently, developers often simply take the data sets available to them. Those data sets are likely to come with inherent limitations such as omissions, inaccuracies, errors, and hidden biases.

[22] Needless to say, this is a highly simplified scenario. Those features can also be built in the information system itself rather than being delivered by a digital assistant.

[23] Outside of the automated information environment, AI has a strong potential to engender moral deskilling. Vallor (2015) points out that automated weapons will lead to soldiers' moral deskilling in the use of military force; new media practices of multitasking may result in deskilling in moral attention; and social robots can cause moral deskilling in practices of human caregiving.

AI algorithms trained with these flawed data sets can fail unexpectedly, revealing those limitations. For example, it has been reported that the success rate of a facial recognition algorithm plunges from 99% to 35% when the group of subjects changes from white men to dark-skinned women because it was trained mostly with the photographs of white men (Lohr 2018). Adopting such a faulty algorithm for any real-life use at a large scale would be entirely unethical. For the context of libraries, imagine using such a face-recognition algorithm to generate metadata for digitized historical photographs or a similarly flawed audio transcription algorithm to transcribe archival audio recordings.

Just like those faulty algorithms, an AI-powered automated information environment can produce information, recommendations, and predictions affected by similar limitations existing in many data sets. The more seamless such an information environment is, the more invisible those limitations become. Automated information systems from libraries may not be involved in decisions that have a direct and immediate impact on people's lives, such as setting a bail amount or determining the Medicaid payment to be paid.[24] But automated information systems that are widely adopted and used for research and scholarship will impact real-life policies and regulations in areas such as healthcare and the economy. Undiscovered flaws will undermine the validity of the scholarly output that utilized those automated information systems and can further inflict serious harm on certain groups of people through those policies and regulations.

Moral Intelligence and Rethinking the Role of AI

In this chapter, I discussed four significant ethical challenges that automating decisions and actions with AI presents: (a) moral desensitization; (b) unintended outcomes; (c) surrender of moral agency; (d) opacity in decision-making.[25] I also examined somewhat different but equally significant ethical challenges in relation to the AI-powered automated information environment, which is likely to surround us in the future: (i) intellectual isolation and bigotry hampering civic discourse; (ii) weakening of cognitive agency and autonomy; (iii) social impact of scholarship and research based upon flawed AI algorithms.

In the near future, libraries will be acquiring, building, customizing, and implementing many personalized and automated information systems. Given this, the challenges related to the AI-powered automated information environment are highly relevant to them. At present, libraries are at an early stage in developing AI applications and applying machine learning and deep learning techniques to improve library services, systems, and operations. But the general issues of hidden biases and the lack of explainability in machine learning and deep learning are already gaining awareness in the library community.

As we have seen in *the trolley problem*, whether a certain action is moral is not a line that can be drawn with absolute clarity. It is entirely possible for fully-functioning moral agents to make different judgements. In addition, there is the matter of morality that our tools and systems display. This is called "machine morality" in relation to AI systems.

Wallach and Allen (2009) argue that there are three distinct levels of machine morality: operational morality, functional morality, and full moral agency (26). Operational morality is found in systems that are low in both autonomy and ethical sensitivity. At this level of machine morality, a machine or a tool is given a mechanism that prevents its immoral use, but the mechanism

[24] See Tashea 2017 and Stanley 2017.

[25] This is by no means an exhaustive list. User privacy and potential surveillance are examples of other important ethical challenges, which I do not discuss here.

is within the full control of the user. Such operational morality exists in a gun with a childproof safety mechanism, for example. A gun with a safety mechanism is neither autonomous nor sensitive to ethical concerns related to its use. By contrast, machines with functional morality do possess a certain level of autonomy and ethical sensitivity. This category includes AI systems with significant autonomy and little ethical sensitivity or those with little autonomy and high ethical sensitivity. An autonomous drone would fall under the former type, while *MedEthEx*, an ethical decision-support AI recommendation system for clinicians, would be of the latter. Lastly, Wallach and Allen regard systems with high autonomy and high ethical sensitivity as having full moral agency, as much as humans do. This means that those systems would have a mental representation of values and the capacity for moral reasoning. Such machines can be held morally responsible for their actions.

We do not know whether AI will be able to produce such a machine with full moral agency. If the current direction to automate more and more human tasks for cost savings and efficiency at scale continues, however, most of the more sophisticated AI applications to come will be of the kind with functional morality, particularly the kind that combines a relatively high level of autonomy and a lower level of ethical sensitivity.

In the beginning of this chapter, I mentioned that the goal of AI is to create an artificial system—whether it be a piece of software or a machine with a physical body—that is as intelligent as a human in its performance, either broadly in all areas of human activities or narrowly in a specific activity. But what does "as intelligent as a human" exactly mean? If morality is an integral component of human-level intelligence, AI research needs to pay more attention to intelligence not only in accomplishing a goal but also in doing so ethically.[26] In that light, it is meaningful to ask what level of autonomy and ethical sensitivity a given AI system is equipped with, and what level of machine morality is appropriate for its purpose.

In designing an AI system, it would be helpful to consider what level of autonomy and ethical sensitivity would be best suited for its purpose and whether it is feasible to provide that level of machine morality for the system in question. In general, the narrower the function or the domain of an AI system is, the easier it will be to equip it with an appropriate level of autonomy and ethical sensitivity. In evaluating and designing an AI system, it will be important to test the actual outcome against the anticipated outcome in different types of cases in order to identify potential problems. System-wide audits to detect well-known biases, such as gender discrimination or racism, can serve as an effective strategy.[27] Other undetected problems may surface only after the AI system is deployed. Having a mechanism to continually test an AI algorithm to identify those unnoticed problems and feeding the test result back into the algorithm for retraining will be another way to deal with algorithmic biases. Those who build AI systems will also benefit from consulting existing principles and guidelines such as FAT/ML's "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms."[28]

We may also want to rethink how and where we apply AI. We and our society do not have

[26] Here, I regard intelligence as the ability to accomplish complex goals following Tegmark 2017. For more discussion on intelligence and goals, see Chapter 2 and Chapter 7.

[27] These audits are far from foolproof, but the detection of hidden biases will be crucial in making AI algorithms more accountable and their decisions more ethical. A debiasing algorithm can also be used during the training stage of an AI algorithm to reduce hidden biases in training data. See Amini et al. 2019, Knight 2019b, and Courtland 2018.

[28] See `https://www.fatml.org/resources/principles-for-accountable-algorithms`. Other principles and guidelines include "Ethics Guidelines for Trustworthy AI" (`https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai`) and "Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability" (`https://ainowinstitute.org/aiareport2018.pdf`).

to use AI to equip all our systems and machines with human- or superhuman-level performance. This is particularly so if the pursuit of such human- or superhuman-level performance is likely to increase unethical decisions that negatively impact a significant number of people. We do not have to task AI with always automating away human work and decisions as much as possible. What if we reframe AI's role as helping people become more intelligent and more capable where they struggle or experience disadvantages, such as critical thinking, civic participation, healthy living, financial literacy, dyslexia, or hearing loss? What kind of AI-driven information systems and environments would be created if libraries approach AI with such intention from the beginning?

References

Alexander, Larry, and Michael Moore. 2016. "Deontological Ethics." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/.

Amini, Alexander, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. 2019. "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure." In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 289–295. AIES '19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3306618.3314243.

Averkamp, Shawn, and Julie Hardesty. 2020. "AI Is Such a Tool: Keeping Your Machine Learning Outputs in Check." Presented at the Code4lib Conference, Pittsburgh, PA, March 11. https://2020.code4lib.org/talks/AI-is-such-a-tool-Keeping-your-machine-learning-outputs-in-check.

Blewer, Ashley, Bohyun Kim, and Eric Phetteplace. 2018. "Reflections on Code4Lib 2018." *ACRL TechConnect* (blog). March 12, 2018. https://acrl.ala.org/techconnect/post/reflections-on-code4lib-2018/.

Boden, Margaret A. 2016. *AI: Its Nature and Future*. Oxford: Oxford University Press.

Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. 2016. "The Social Dilemma of Autonomous Vehicles." *Science* 352 (6293): 1573–76. https://doi.org/10.1126/science.aaf2654.

Clark, Liat. 2012. "Google's Artificial Brain Learns to Find Cat Videos." *Wired*, June 26, 2012. https://www.wired.com/2012/06/google-x-neural-network/.

Conitzer, Vincent, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. 2017. "Moral Decision Making Frameworks for Artificial Intelligence." In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4831–4835. AAAI'17. San Francisco, California, USA: AAAI Press.

Courtland, Rachel. 2018. "Bias Detectives: The Researchers Striving to Make Algorithms Fair." *Nature* 558 (7710): 357–60. https://doi.org/10.1038/d41586-018-05469-3.

Cushman, Fiery, Liane Young, and Marc Hauser. 2006. "The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm." *Psychological Science* 17 (12): 1082–89.

Davis, Daniel L. 2007. "Who Decides: Man or Machine?" *Armed Forces Journal*, November. http://armedforcesjournal.com/who-decides-man-or-machine/.

Davis, Hannah. 2020. "A Dataset Is a Worldview." Towards Data Science. March 5, 2020. https://towardsdatascience.com/a-dataset-is-a-worldview-5328216dd44d.

Foot, Philippa. 1967. "The Problem of Abortion and the Doctrine of Double Effect." *Oxford Review* 5: 5–15.

Heaven, Will Douglas. 2020. "DeepMind's AI Can Now Play All 57 Atari Games—but It's Still Not Versatile Enough." *MIT Technology Review*, April 1, 2020. https://www.technologyreview.com/2020/04/01/974997.

International Committee of the Red Cross. 2015. "What Are Jus Ad Bellum and Jus in Bello?" January 22, 2015. https://www.icrc.org/en/document/what-are-jus-ad-bellum-and-jus-bello-0.

Kahn, Leonard. 2012. "Military Robots and The Likelihood of Armed Combat." In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and George A. Bekey, 274–92. Intelligent Robotics and Autonomous Agents. Cambridge, Mass.: MIT Press.

Knight, Will. 2017. "The Dark Secret at the Heart of AI." *MIT Technology Review*, April 11, 2017. https://www.technologyreview.com/2017/04/11/5113.

———. 2019a. "Two Rival AI Approaches Combine to Let Machines Learn about the World like a Child." *MIT Technology Review*, April 8, 2019. https://www.technologyreview.com/2019/04/08/103223.

———. 2019b. "AI Is Biased. Here's How Scientists Are Trying to Fix It." *Wired*, December 19, 2019. https://www.wired.com/story/ai-biased-how-scientists-trying-fix/.

Koch, Christof. 2016. "How the Computer Beat the Go Master." *Scientific American*. March 19, 2016. https://www.scientificamerican.com/article/how-the-computer-beat-the-go-master/.

Lohr, Steve. 2018. "Facial Recognition Is Accurate, If You're a White Guy." *New York Times*, February 9, 2018. https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html.

Mahlangu, Isaac. 2019. "Meet Libby - the New Robot Library Assistant at the University of Pretoria's Hatfield Campus." *SowetanLIVE*. June 4, 2019. https://www.sowetanlive.co.za/news/south-africa/2019-06-04-meet-libby-the-new-robot-library-assistant-at-the-university-of-pretorias-hatfield-campus/.

Markoff, John. 2012. "How Many Computers to Identify a Cat? 16,000." *New York Times*, June 25, 2012.

Marr, Bernard. 2018. "How AI And Machine Learning Are Transforming Law Firms And The Legal Sector." *Forbes*, May 23, 2018. https://www.forbes.com/sites/bernardmarr/2018/05/23/how-ai-and-machine-learning-are-transforming-law-firms-and-the-legal-sector/.

Pariser, Eli. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think.* New York: Penguin Press.

Price, Gary. 2019. "The Library of Congress Posts Solicitation For a Machine Learning/Deep Learning Pilot Program to 'Maximize the Use of Its Digital Collection.' " LJ InfoDOCKET. June 13, 2019. https://www.infodocket.com/2019/06/13/library-of-congress-posts-solicitation-for-a-machine-learning-deep-learning-pilot-program-to-maximize-the-use-of-its-digital-collection-library-is-looking-for-r/.

Rincon, Lilian. 2019. "Interpreter Mode Brings Real-Time Translation to Your Phone." *Google Blog* (blog). December 12, 2019. https://www.blog.google/products/assista

nt/interpreter-mode-brings-real-time-translation-your-phone/.

Sharkey, Noel. 2012. "Killing Made Easy: From Joysticks to Politics." In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and George A. Bekey, 111–28. Intelligent Robotics and Autonomous Agents. Cambridge, Mass.: MIT Press.

Singer, Peter. 2005. "Ethics and Intuitions." *The Journal of Ethics* 9 (3/4): 331–52.

Sinnott-Armstrong, Walter. 2019. "Consequentialism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/consequentialism/.

Stanley, Jay. 2017. "Pitfalls of Artificial Intelligence Decisionmaking Highlighted In Idaho ACLU Case." *American Civil Liberties Union* (blog). June 2, 2017. https://www.aclu.org/blog/privacy-technology/pitfalls-artificial-intelligence-decisionmaking-highlighted-idaho-aclu-case.

Talley, Nancy B. 2016. "Imagining the Use of Intelligent Agents and Artificial Intelligence in Academic Law Libraries." *Law Library Journal* 108 (3): 383–402.

Tashea, Jason. 2017. "Courts Are Using AI to Sentence Criminals. That Must Stop Now." *Wired*, April 17, 2017. https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/.

Tegmark, Max. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Alfred Knopf.

Thomson, Judith Jarvis. 1976. "Killing, Letting Die, and the Trolley Problem." *The Monist* 59 (2): 204–17.

Turek, Matt. n.d. "Explainable Artificial Intelligence." Defense Advanced Research Projects Agency. https://www.darpa.mil/program/explainable-artificial-intelligence.

Vallor, Shannon. 2015. "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character." *Philosophy & Technology* 28 (1): 107–24. https://doi.org/10.1007/s13347-014-0156-9.

Wallach, Wendell. 2009. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Welch, Chris. 2018. "Google Just Gave a Stunning Demo of Assistant Making an Actual Phone Call." The Verge. May 8, 2018. https://www.theverge.com/2018/5/8/17332070/google-assistant-makes-phone-call-demo-duplex-io-2018.