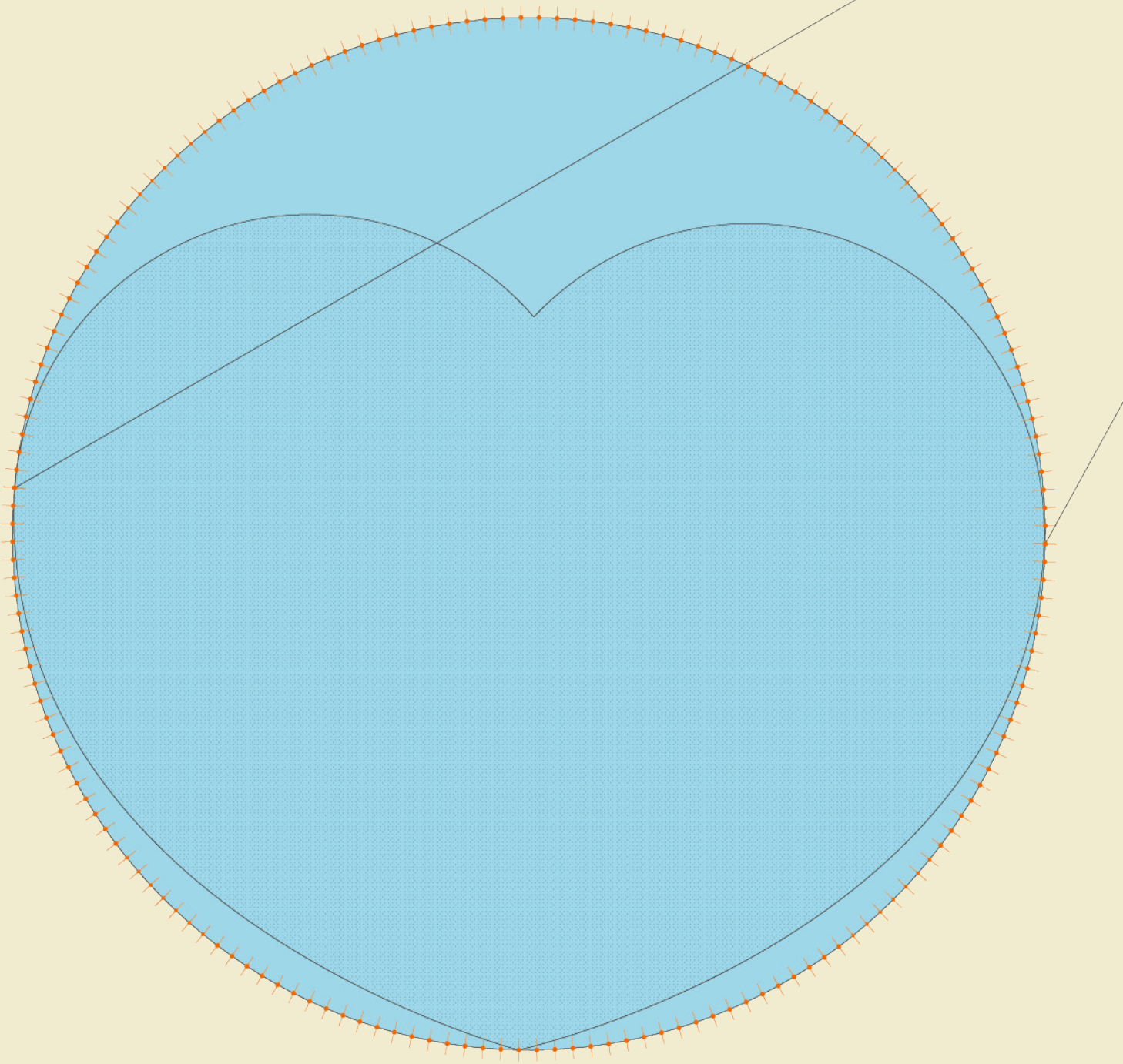
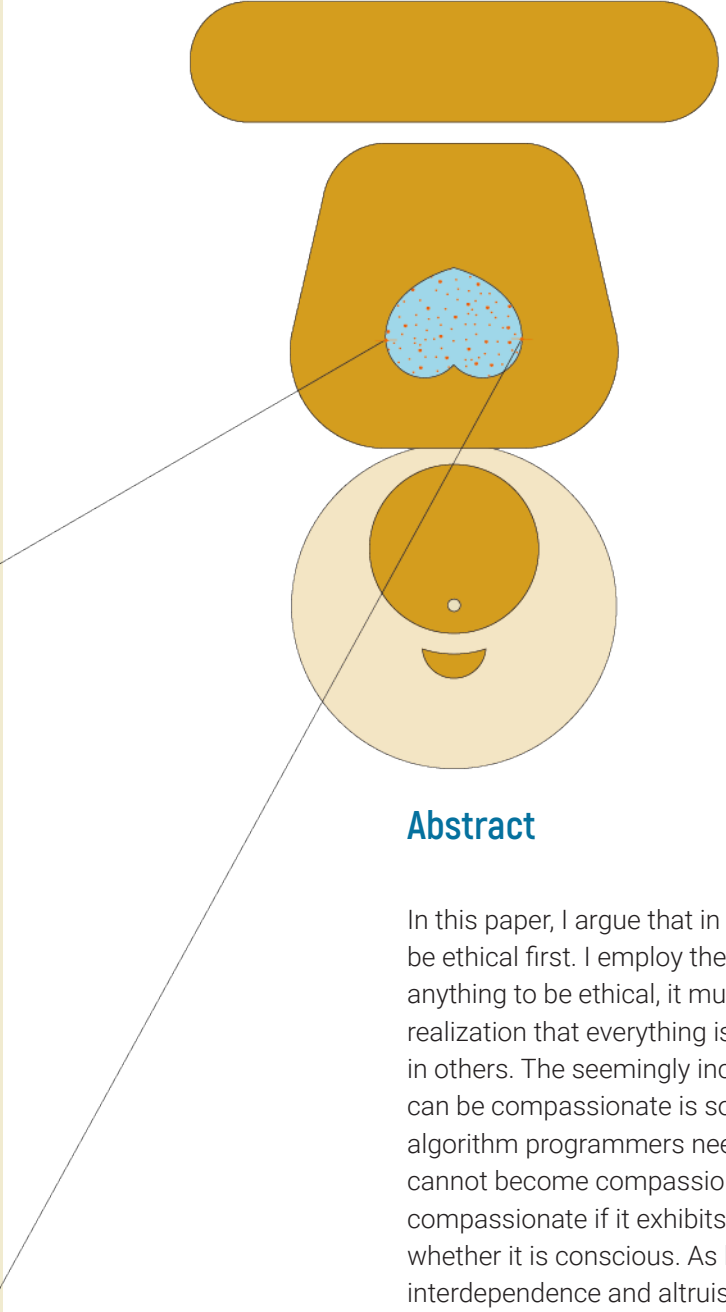
Philosophical Point of View for Social Implementation

AI for Social Good:

Buddhist Compassion as a Solution

Soraj Hongladarom
Department of Philosophy,
Faculty of Arts,
Chulalongkorn University





Abstract

In this paper, I argue that in order for artificial intelligence (AI) to deliver social good, it must be ethical first. I employ the Buddhist notion of compassion (*karunā*) and argue that for anything to be ethical, it must exhibit qualities that characterize compassion, namely the realization that everything is interdependent and the commitment to alleviate suffering in others. The seemingly incoherent notion that a thing (e.g., an AI machine or algorithm) can be compassionate is solved by the view that – at this current stage of development – algorithm programmers need to be compassionate. This does not mean that a machine cannot become compassionate in another sense. For instance, a machine can become compassionate if it exhibits the qualities of a compassionate being, regardless of whether it is conscious. As long as the machine exhibits the outward characterization of interdependence and altruism, then it can be said to be compassionate. This paper also argues that the ethics of AI has to be integral to the coding of its program. In other words, the ethics (i.e., how we would like the AI to behave based on our ethical standpoint) needs to be programmed into the AI software from the very beginning. This study has also replied to several objections against this idea. To summarize, coding ethics into a machine does not imply that the ethics thus coded belongs solely to the programmer, nor does it mean that the machine is thereby completely estranged from its socio-cultural context.

Introduction

In the past few years, few innovations in technology have aroused as much public interest and discussion as AI. After many years of lying in the doldrums, with many broken promises in the past decades, AI once again became a focal point after it defeated both the European champion and reigning world champion at the ancient game of Go in 2016. The defeat was totally unexpected, as computer scientists and the public believed that Go was much more complex than chess. Since the number of possible moves that needed to be calculated was too vast for any computer to calculate, many believed that Go represented the supreme achievement of human beings, and could not be bested or emulated by a machine. Thus,

there was worldwide sensation after both the European champion Fan Hui, and Lee Sedol, the world champion, were soundly defeated at Go by a machine in a relatively short span of time. Following this AI victory, it became clear that no human could ever defeat a machine in a board game.

What ensued was an explosion in the power of AI — a resurgence after many years of dormancy and repeated failed promises. AI has been with us for many decades. Computer scientists who developed it believed that a computer could actually mimic the workings of the human brain. The project seemed promising at first; for example, the computers could play Tic-Tac-Toe, Checkers, and eventually chess. Some progress was also made in the field of natural language processing and machine translation. Nonetheless, these successes were not as spectacular as the scientists themselves had envisioned, and AI was unable to fulfil the expectations that its developers had originally claimed. For example, the expert system environment was developed during the early 1980s, but was prone to mistakes and thus became not suitable for normal use. The market for expert systems thus largely failed. Many promises of AI systems at that time, such as speech recognition, machine translation, and others, were not fulfilled. As a result, funding was largely cut, and AI research made very little progress. These failures were largely due to the fact that computers at that time lacked power, and data, so their predictive power remained limited.

The software that created history, AlphaGo, was developed by DeepMind, a British company founded in 2010 and acquired by Google in 2014. The company made history in 2015 and 2016 when its AI creation, AlphaGo, defeated both the European champion and the world champion of Go. The technique used by AlphaGo was radically different from Deep Blue, a software developed by IBM which defeated the chess world champion, Gary Kasparov, in 1997. Deep Blue used GOFAI, or "good old-fashioned AI", to blindly search for the best possible moves using a brute force search technique. This technique proved unfeasible for much more complex games such as Go, where the number of possible moves exceed the number of atoms in the universe. Thus, AlphaGo used a new technique which was also being developed at that time. The new technique, known as deep learning, avoided the brute force search technique, and instead relied on very large amounts of data. The program learned from this data to determine the best moves. The data from millions of past moves made by humans limited the number of possible moves that the algorithm would need to make, thus enabling it to focus on the most relevant moves. This, coupled with more powerful hardware, contributed to the program defeating Lee Sedol. The event was watched by many people worldwide, and its success was a "Sputnik moment" in terms of bringing AI back into the spotlight. Now, many researchers are racing against each other to find the most useful applications for the technology.

Many applications are being touted as potential ways in which deep learning AI could help to solve the world's problems. The following applications are currently being promoted: self-driving cars, deep learning (AI use) in healthcare, voice search or voice assistants, adding sounds to silent movies, machine translation, text generation, handwriting generation, image recognition, image caption generation, automatic colorization, advertising, earthquake prediction, brain cancer detection, neural networks in finance, and energy market price forecasting (Mittal, 2017). Some of these applications indeed address serious matters, such as self-driving cars and image recognition, while others are rather quaint, such as colorization or automatic sound generation in silent movies. In any case, Mittal mentions that some of the most prominent applications of deep learning (or machine learning) AI has emerged over the past three or four years. One of the most powerful uses of today's AI is its predictive power. Using vast data sources, AI promises to make predictions that would not be conceivable by human analysts. One of the promises, for example, concerns an AI system that can detect the onset of cancer by analyzing images of those who are still healthy. In other words, the power of today's AI lies in its ability to "see" things that are often undetected by trained specialists. The algorithm gains this ability through its analysis of

extensive data points that are fed into its system. The machine analyzes these data and finds patterns and correlations to make predictions.

This new technology has led many to look for ways in which AI could improve society. The applications mentioned in Mittal's article identifies some of the potential uses, or "social goods" that could be delivered by AI. Many large corporations have also jumped on the bandwagon in search of AI opportunities. Google, for example, has founded an initiative titled "AI for Social Good" (http://ai.google/social-good/), which aims at "applying AI to some of the world's biggest challenges", such as forecasting floods, predicting cardiac events, mapping global fishing activity, and so on (AI for Social Good, 2020).

This paper analyzes some of the ethical concerns arising from such applications. Researching the potential of AI to solve these problems is important, but when the technology is applied in real-world scenarios, care must be taken to ensure that the social and cultural environment is fully receptive to the technology. Not being receptive to the imported technology can lead to a sense of alienation, which can happen when the local population is excluded from the process of decision making regarding the adoption of the technology in question (Hongladarom, 2004). This could also lead to a resistance to AI technology. For example, using AI to forecast floods may lead to administrative measures that could cause mistrust or misunderstandings if the AI technology is not made clear to those affected by the measure. It is one thing for AI (if reliable) to identify when and where a severe flood will take place; it is another to convince a local population that a flood will occur and that their location will be affected. This shows that any successful employment of AI must factor in local beliefs and cultures. Moreover, the forecasting must not be used to gain an unfair advantage over others. For example, forecast knowledge of floods in a particular area and time might lead to hoarding or other unfair measures designed to maximize the individual gains of certain parties. This shows that ethics must always be integral to any kind of deployment of technology and its products.

Consequently, this paper aims to find ways in which machine learning AI could deliver social good in an ethical manner. More specifically, this paper argues that in order for AI to deliver social good, it must be ethical first. Otherwise, it might lead to negative outcomes that are similar to the aforementioned scenario of flood forecasting and hoarding. This is a vital principle to address, as sophisticated technology, such as facial recognition software, could be used to endanger people's right to privacy. As mentioned above, AI algorithms that forecast flooding could be used to gain unfair advantages over others. Hence, there must be a way for these algorithms themselves to act as safeguards against such use. For flood forecasting software, this might not be immediately apparent as it does not typically involve autonomous action. The software would likely deliver information and forecasting, with humans ultimately being responsible for acting on the information. However, even in this case, the software itself must be ethical on its own. At the very least, there should be some form of mechanism in which the possibility of misuse or abuse by certain groups (such as those intent on using the information to hoard food and other supplies) is minimized; such a mechanism should be installed as part of the software from the very beginning. Regarding facial recognition technology, the same type of mechanism should also be installed to avoid potential misuse. Simply, AI should be an integral part of an ethical way of living, right from the moment of implementation. Hence, instead of regarding AI and its surrounding technologies as something imported and inherently harmful towards the developing world, we must find a way in which AI becomes integral to help these people flourish.

Furthermore, this paper argues that the details of how to live an ethical life should include insights obtained from Buddhism; specifically, the teachings on compassion (*karunā*), which is one of the most important tenets of Buddhism. It may be suggested that Buddhist compassion — a concept that will be further developed in this paper — should play a key role in developing an *ethical* AI. This development then comprises the possibility of AI to deliver social good and function as an integral part of ethical living.

AI is undoubtedly powerful and has the potential to significantly change the world. Power always has to be accompanied by corresponding responsibility, restraint, and other ethical virtues.

The next section of this paper will review some of the current literature on the ethics of AI and AI for social good. Section 3 deals with the basic concepts of Buddhism. Section 4 presents the paper's main argument, together with replies to some of the objections during the course of research. The last section concludes with two main policy recommendations for the public sector and tech companies.

AI for Social Good

The advent of AI has given rise to a plethora of ethical guidelines that aim to regulate AI research and development worldwide. A survey of the literature on AI for social good revealed that much of the literature overlaps with the ethics of AI and proposals for AI ethics guidelines in general. This is not surprising, as proposing AI for social good implies that AI should act ethically; by promoting social good, AI thereby becomes ethical. However, this transition is not automatic; one still has to provide an account of why it is indeed the case. The need for such an account seems to be more acute when an AI program might be created with the aim of providing a social good, but instead, turns out to be harmful. This justification forms one of the main objectives of this paper.

Nevertheless, it is important to review the literature on ethics guidelines for AI, as well as AI for social good, to provide a general outline and identify some of the key issues. A website titled "AI Ethics Guidelines Global Inventory" (https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/) has documented 82 guidelines. However, only four Asian countries are represented on the list: China, Korea, Dubai, and Japan. It should also be noted that none of the documents published in these countries are based on their own indigenous intellectual resources (see also Gal, 2019). This shows that there is a very high level of interest in how AI should be ethically grounded. In a related paper, "The Ethics of AI Ethics", Thilo Hagendorff (Hagendorff, 2019) documents the ethical concepts that are mentioned in some of these guidelines, and identifies the top five concepts, which include privacy, accountability, fairness, transparency, and safety (Hagendorff, 2019). These factors largely correspond with a list in another paper written by Luciano Floridi and others (Floridi et al, 2020), where seven "essential factors" are listed, namely: (1) falsifiability and incremental deployment, (2) safeguards against manipulation of predictors, (3) receiver-contextualized intervention, (4) receiver-contextualized explanation and transparent purposes, (5) privacy protection and data subject consent, (6) situational fairness, and (7) human-friendly semanticization (Floridi et al, 2020, p. 5). Here, falsifiability means that the software system needs to be empirically testable, and only if it is testable will it be deemed trustworthy. Factor (2) (safeguards against predictors) is rather straightforward; it means that there needs to be a mechanism whereby false manipulation of input into the software is prevented, so that the results produced by the software are not biased. Factor (3) (receiver-contextualized intervention) refers to respecting the autonomy of the user; any intervention performed by the software needs to be "contextualized" to the needs and desires of the user. Factor (4) (receiver-contextualized explanation and transparent purposes) refers to respecting the autonomy of the user in terms of the software being easy and transparent to understand, where nothing important is hidden. Factor (5) (privacy protection and data subject consent) is self-explanatory and is the number one concern in the guidelines studied in Hagendorff's paper. Factor (6) (situational fairness) refers to the need for the software to maintain objectivity and neutrality by avoiding data input that is biased from the beginning. Factor (7) (human-friendly semanticization) means that humans should still maintain a level of control when the software is allowed to interpret and manipulate meaningful messages. For example, AI software can create clearer communication between the caregiver and patient, without intervening and excluding the caregiver from the process (Floridi et al, 2020, pp. 5-19).

These factors and concepts are also very much related to another set of concepts, also developed primarily by Floridi (Floridi et al, 2018; see also Cowls and Floridi,

2018). In this paper, Floridi and his team delineate five elements that are necessary for "good" AI in society. Most of these elements resemble the familiar ethical principles found in other areas of applied ethics, most notably in medical ethics. These are beneficence, non-maleficence, autonomy, and justice. Then Floridi and his team add another factor, explicability, which is unique to AI as it tends to operate in a "black box", where the normal user has no clue over how it works and how it comes up with its own answers (Floridi et al, 2018). Moreover, Mariarosario Taddeo and Floridi also have another article published in Science in 2018 mentioning the need for these factors for a good AI society (Taddeo and Floridi, 2018). They also discuss the need for what they call a "translational ethics" that combines foresight methodologies and analyzes of ethical risks (Taddeo and Floridi, 2018). In addition, these five principles are also discussed in The European Commission's High Level Expert Group on Artificial Intelligence (The European Commission's High-Level Expert Group on Artificial Intelligence, 2018, pp. 8-10), with the emphasis that AI systems need to be "human-centric" (The European Commission's High-Level Expert Group on Artificial Intelligence, 2018, p. 14). The overall concern of the document is that AI needs to be "trustworthy", and the requirements discussed here are among the necessary conditions. More specifically, the document discusses ten factors that are supposed to be sufficient for a trustworthy AI system. These are accountability, data governance, design for all, governance of AI autonomy (human oversight), non-discrimination, respect for (and enhancement of) human autonomy, respect for privacy, robustness, safety, and transparency (The European Commission's High-Level Expert Group on Artificial Intelligence, p. 14). Thus, these ten requirements largely mirror the requirements or essential factors mentioned earlier. Chief among these lists are factors such as autonomy, privacy, safety, and transparency. It is clear that there are many overlaps among such guidelines, with only relatively small differences among them.

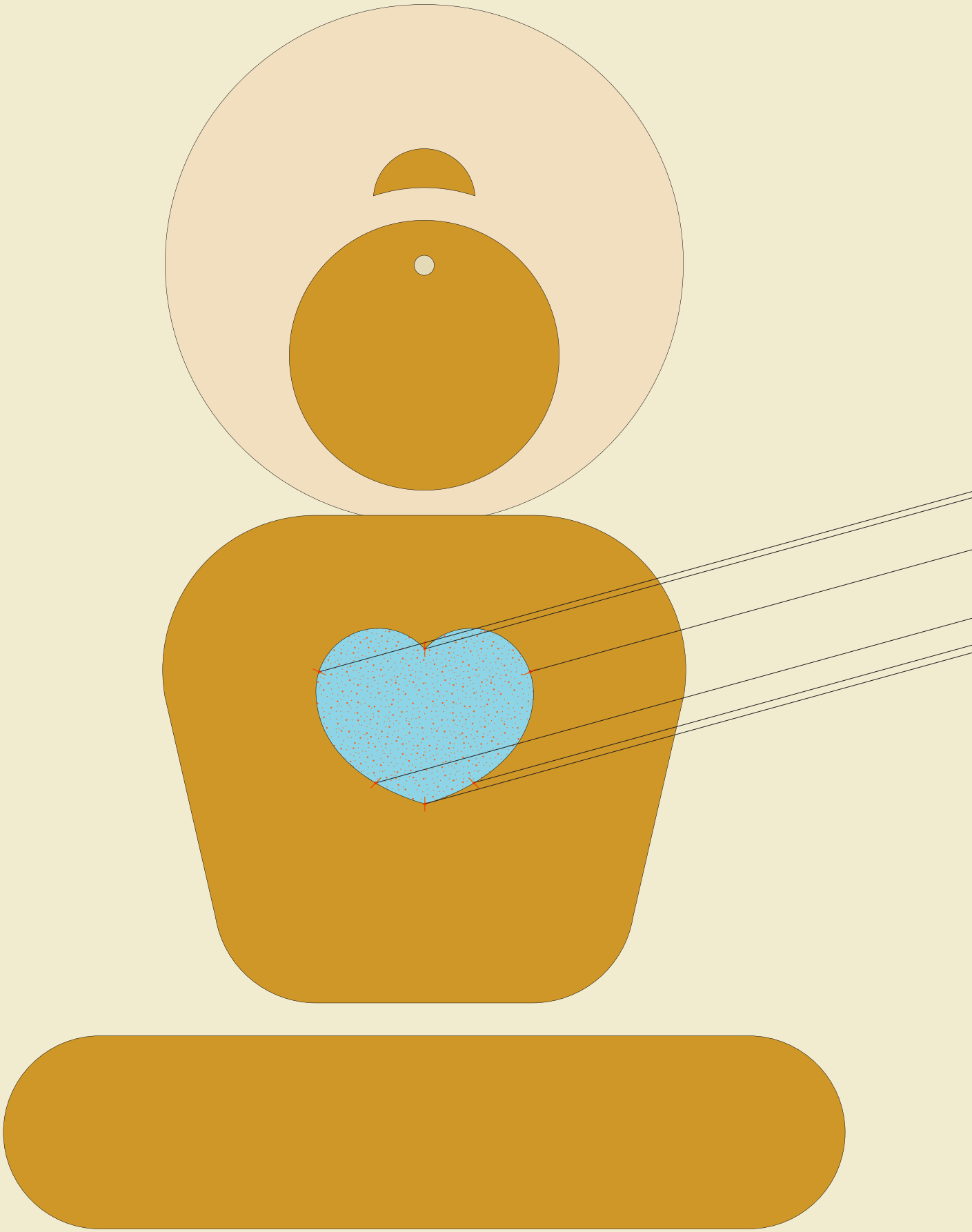
Furthermore, Ben Green (Green, 2019) argues that computer scientists cannot rely on the idea that algorithms alone can solve the world's problems, but they need to see how social programs (which AI for Social Good is supposed to solve) are all connected with deeper and more intricate interconnections, which mere technical means alone cannot solve. Bettina Berendt, in a similar vein, proposes an "ethics pen-testing" where the design of AI is critically challenged by a series of questions aimed at the designer to defend himself/herself and to show that the design is ethically sensitive, all in order to improve the software design (Berendt, 2019). What is interesting in both Green's and Berendt's papers is that they are not content on merely proposing a list of guidelines for AI developers to follow, and instead point out that AI researchers and developers must be aware of ethics during all stages of development. Technical solutions alone are not enough, and will not be effective in bringing about the proposed "social good" of AI.

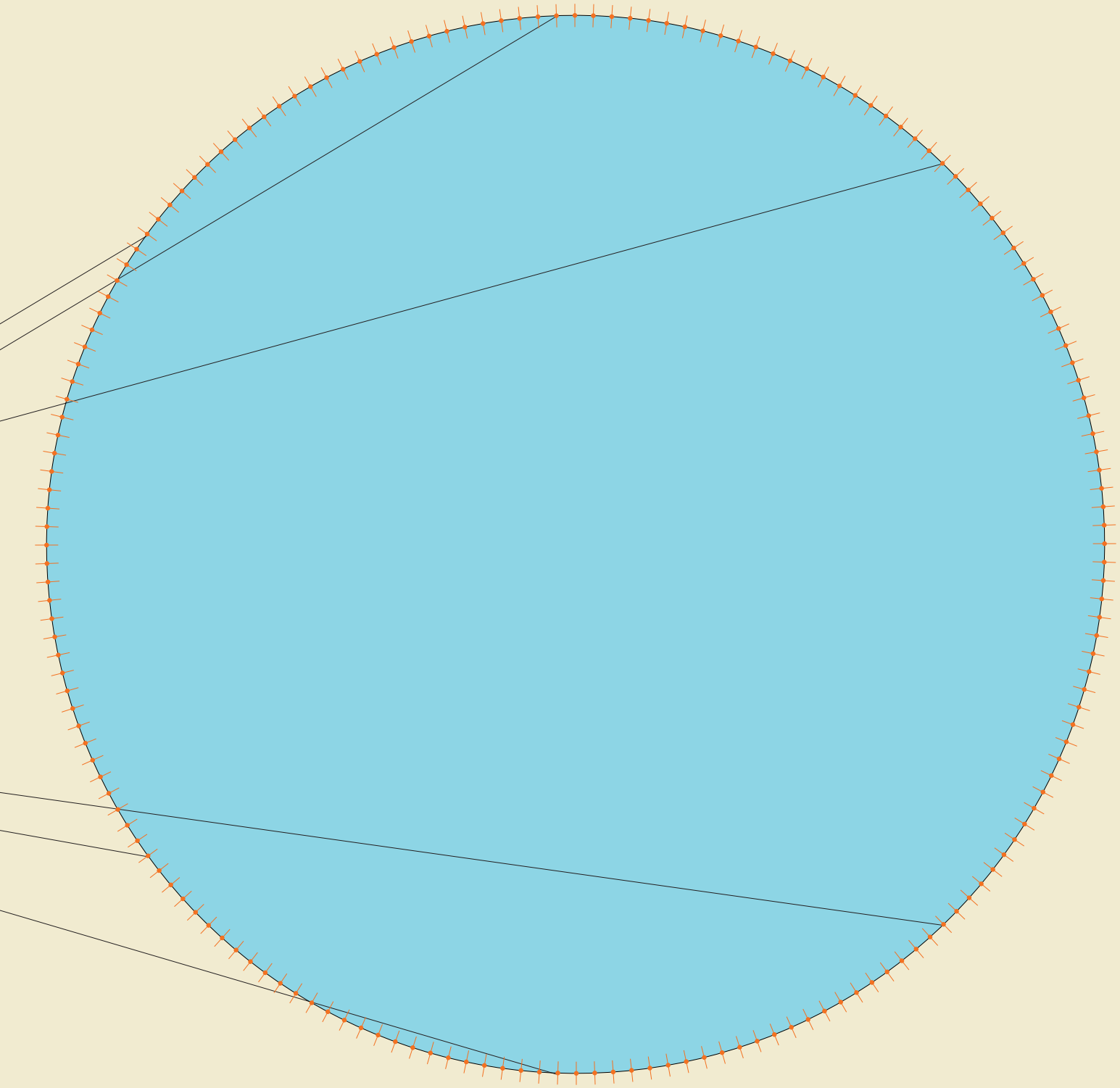
What has emerged is that most of the literature focuses on a list of ethical principles which, they argue, should be necessary for an effective ethical AI system. However, only a few works (e.g., Green and Berendt) argue that simply providing such a list bypasses the deeper interweaving connection between ethical principles and the underlying social and cultural contexts. Nonetheless, both Green and Berendt address these contexts in a vertical manner. More specifically, they focus on the interrelations between ethical principles and the wider concerns in a Western context. As mentioned earlier, there are only a few guidelines in Asia, and more interestingly, these guidelines do not mention their own intellectual resources. Hence, a large gap exists in the literature, namely the formulation of AI ethics principles based on the intellectual resources of the East. In fact, my recent book, "The Ethics of AI and Robotics: A Buddhist Viewpoint", discusses this issue in great detail (Hongladarom, 2020). Moreover, going beyond the gap in theoretical terms, there is also a gap in the content of the proposed guidelines. What I propose in this paper is that a complimentary principle of Buddhist ethics should be adopted as the foundation for thinking and deliberating on the ethics of AI and AI for social good. Furthermore, the principle of *karunā* (compassion) should be considered for the ethical guidelines of AI and for any theory related to AI for social good.

Buddhist Ethics and Basic Buddhist Principles

It is not possible to explain all of the principles of Buddhism in this paper. Nonetheless, a very brief introduction to its relevant principles should provide a better context for the argument. More details on the principles of Buddhist ethics and an introduction to Buddhist philosophy can be found in the book that I mentioned earlier (Hongladarom, 2020). The book explains that Buddhist ethics is based on the idea that an action is considered right if it brings out something that is universally desired by all human beings, and wrong if it goes in the opposite direction. Thus, Buddhist ethics is markedly different from modern ethical theories; for instance, other theories do not specify what is universally desirable for all humans. In Immanuel Kant's ethical theory, for example, the basic idea of what constitutes a good action comes without considering the possible consequences of that action. Instead Kant's theory questions whether the action follows a universalizable maxim or not. The universally desirable goal, on the contrary, is definitely a goal; thus, Buddhist theory is in opposition with Kant's deontological theory. Furthermore, Buddhist theory is also different from utilitarianism in that, although utilitarianism is a kind of consequentialism, Buddhist theory specifies a definite content of the goal that is universally desirable to all human beings. Conversely, utilitarianism does not specify any definite content, and instead focuses on content that is deemed utilitarian. Buddhism suggests the possibility of a universally desirable goal that is valid for everyone. Since everyone desires happiness and wishes to avoid suffering, it may be seen as a universal goal. Buddhism has a very detailed theory regarding the definition of happiness as a universal goal. In simple terms, it describes a type of happiness that results when one's action is in total accordance with nature. Thus, the kind of "happiness" that results from indulging in sensual pleasure would not qualify, as this pleasure also brings about suffering. For example, eating certainly brings pleasure, but too much eating can cause a certain degree of discomfort, such as feeling bloated, etc. Therefore, true happiness (i.e., without suffering) is only attainable through a true understanding of nature. This does not mean becoming a scientist, but instead, understanding that nature works according to the rules of cause and effect. Realizing this is a necessary step towards attaining what Buddhists call "*nirvāna*" or total cessation of all suffering. The term is usually translated as "Enlightenment." Hence, Buddhist ethical theory explains that an action is good if it leads to *nirvāna*, and vice versa.

As mentioned earlier, the aim of this paper is to show that Buddhist philosophy can contribute to the ethics of AI and AI for Social Good. A key point is that a person's actions must be in tune with nature. When this is the case, they essentially become one with nature. This is a concrete expression of the realization that there is no attachment to the ego, since it is just such an attachment that separates one from becoming fully in tune with nature. Compassion is a key ingredient in this realization, and what is truly good is the realization that there is no boundary between the ego and everything else, as well as the resultant desire to help others get rid of their suffering, which is ultimately due to a lack of realization. In the area of AI ethics and AI for social good, this means that one has to find a way in which AI can contribute to relieving the suffering of all beings. This may not be as grandiose as it may sound, as we are more than capable of finding out specific and concrete ways to achieve this. Doing so is to implement an ethics of AI that is in accordance with Buddhist ethical principles. The main idea being that, in order for AI to provide social good, it must consider the contexts involved, which may vary from place to place. A solution that might work in one context might not work in another. The examples put forward in this paper are flood forecasting and facial recognition, however, we can certainly imagine other cases. In the field of automated reasoning or decision making, one also needs to be careful that the decisions made by AI are always accountable to humans. Allowing AI to have a free hand in making decisions (such as in stock trading) would go against the Buddhist principle of compassion, as this tends to create more suffering rather than reduce it.





Could AI Become Compassionate?

As we have seen, this study argues that AI needs to be compassionate. This means that AI must exhibit the two qualities that constitute compassion, namely interdependence and altruism. AI exhibits interdependence by showing concretely that it understands (within the constraints of current AI technology) the concept of things being interdependent and interconnected. This can be achieved with an AI algorithm that shows concern for the welfare of someone or something. For example, the aforementioned flood forecasting algorithm could show a level of understanding of interdependence by having in its internal mechanism connected to other relevant factors that are no less important, such as economic conditions, price forecasting, political climate, and geographical information, etc. AI flood forecasting could lead to the hoarding of essential food and supplies, which is an unethical act. However, the algorithm might struggle to learn how its predictions could be used by humans in a negative way. Here, a program that embeds algorithms in a larger context could make it more difficult for information to be used for personal gains. For example, the algorithm could publicly broadcast its predictions, making it impossible for certain parties to gain an advantage. An internal "safety lock" within the algorithm could be installed as an indelible component to make it imperative to broadcast information to everyone involved rather than to individual users. The broadcasting feature may, however, be necessary for flood forecasting, but broadcasting on this scale might be unethical in other contexts or for certain algorithms. For example, some algorithms are intended to work privately (e.g., personal health information). As such, developers need to see which contexts are relevant for installing safety mechanisms inside algorithms.

The other component of Buddhist compassion is the commitment to alleviate suffering for all sentient beings. Here, sentient beings are relieved of their suffering through someone who is completely compassionate. However, such an ideal is impossible to realize in reality, where the one who practices compassion has limited power. Nonetheless, we must do whatever we can—within the limits of our power—to help relieve suffering. For AI algorithms, this would mean taking active steps in creating a world where suffering is eliminated as much as possible. More specifically, the algorithm should be designed to help alleviate suffering from the very beginning. For example, facial recognition technology could be developed to recognize particular features so that certain traits are predicted, such as the onset of a disease, leading to early prevention. One may assume that suffering is unrelated to software development, as it appears to be an external requirement. However, it should be an integral part of software development in itself. This pertains to key areas or problems which AI algorithms will be designed to solve from the beginning.

Michael Kearns and Aaron Roth (Kearns and Roth, 2019) argue that an algorithm should be ethical in the sense that ethical components should be programmed into the algorithm. Here I suggest that compassion should also be programmed into AI algorithms. In fact, the same idea has already been proposed by James Hughes (Hughes, 2012). However, according to Hughes, a robot only becomes compassionate when it can imitate human emotion. I propose that compassion can be attained when it exemplifies the two components mentioned earlier, namely realization of interdependence and the commitment to relieve suffering. More specifically, a robot becomes compassionate when it exhibits genuine commitment and action geared toward alleviating suffering. Thus, it is more action-oriented than merely displaying or mimicking emotions.

How can we program robots or AI algorithms to be compassionate? We could say that an algorithm "understands" interdependence when it is programmed in such a way that it "recognizes" various external factors that are involved in making a more ethically nuanced assessment. Of course, the algorithm does not understand anything—we are not talking about a superintelligence—but it is a way of talking to show that the algorithm exhibits certain behaviors that we recognize colloquially as an

understanding. Hence, for the algorithm to understand interdependence, which is one component of Buddhist compassion, it has to exhibit certain external features that are not directly part of its core objective, so to speak. These features may not be part of the core mission, but they are very important in making an ethical judgment of the situation in which it is employed in order that it becomes more ethical. If a given objective, such as to maximize a certain output, is found to involve trade-offs between the output and other desirable factors, then the machine would be programmed not to follow the maximization. It will realize or "understand" that such an action leads to a contradiction with its own prime directive, which is to alleviate human suffering. To come back to flood prevention software, an algorithm might be taught to accurately predict floods in a certain area. However, predicting floods alone is not ethical as it could lead to hoarding, as we have seen. Thus, the AI needs to be programmed with compassion so that it can predict floods while also considering other relevant factors. For example, the AI could display a warning sign if a user attempts to misuse the data. Then, the second component of compassion, altruism, is ideally put into action when the algorithm initiates an action designed to help relieve affected persons from suffering. To use another example, a microloan algorithm might override its directive (maximizing profit for its creator or owner) in favor of clients who, on paper, would have suffered even more if the algorithm did not act otherwise. Here the algorithm must be able to distinguish between clients who really need the money, and who show good faith and commitment to repaying the loan, from those clients who are out to get cheap money without any intention of repayment. In this case, there are many specific details involved; the idea I am proposing is only that the algorithm should follow the Buddhist principle of trying to relieve suffering as best it can, based on the information available to it at the time.

Some may object to this proposal, saying that giving AI its own discretion in making more ethical decisions will inhibit the freedoms of the human user in applying AI in any way he or she sees fit. Furthermore, there is no guarantee that the algorithm will act as ethically as intended. These are legitimate concerns. Nonetheless, installing a component that inhibits the user from performing certain actions is not a new principle. For instance, some cars will not start unless the driver is wearing a seatbelt. The AI that refuses to follow certain orders from the user acts in the same way. Such a car limits the freedom of the user, but this is still seen as a strong safety feature. Additionally, how do we know that the AI, when given this amount of freedom, will always act ethically? For the artificial general intelligence (AGI) of the future, this is a serious matter because AGI's are capable of thinking on their own. Therefore, it is in our best interest to guide its development towards being both intelligent and ethical. For today's more specialized AI, however, safety devices should be installed or programmed so that the algorithm functions to promote ethical action.

In fact, giving AI the ability to act ethically is possible with today's technology. This does not necessarily mean that the AI is endowed with consciousness and free will. Instead, the AI is equipped with algorithms to act ethically and compassionately from the beginning. The microloan software will act ethically if it takes the interest of its clients into account. This might not maximize the bank's profit, but the social cost of being inflexible when loan decisions are analyzed and approved could be greater. As an increasing number of loan decisions are made autonomously by algorithms, having an ethical algorithm seems essential.

Objections and Replies

During a series of meetings held by the Association of Pacific Rim Universities (APRU) under the project titled "AI for Social Good", my proposal benefited from a number of comments and helpful criticisms from my colleagues, who challenged me to develop a better, more defensible position on this topic. The first objection focused on the claim that ethics should be encoded into the algorithm or the inner programming core of the AI software. The objection is that it makes ethics too narrow and technical. According to this objection, ethics coding would result in the AI system being estranged from its social, cultural, and economic environment, leading to the system not being relevant to the aims of the forum. First, it should be noted that I would not advocate that social and cultural considerations should be taken away from ethical deliberation. This is just not possible, because ethics is always naturally embedded in the set of practices that surround any technical product, which is something that has been recognized by technology philosophers for a long time. For example, a car that remains stationary until the driver puts on a seatbelt, is an example of encoded ethics. According to my analysis, a car that neglects to warn the driver to wear a seatbelt and does not take appropriate action to ensure that he or she does so, is unethical. In the same vein, it is also ethical for microloan software to take more data points than required to ensure that loans are repayable. Sure enough, a program that makes an accurate risk calculation for a loan would to some extent be an ethical program. However, if this is all the software does, then its degree of ethicality is limited. It needs to consider other factors too, such as the condition of the loan applicant (e.g., economic status, children, health, etc.). It would be more prudent for the program to provide a loan under certain economic conditions, such as the current COVID-19 pandemic. The act of coding ethics into the inner workings of an AI program does not imply that the coder and employer are isolated from the surrounding socio-economic conditions or social environment. On the contrary, it shows that the coder and tech company value ethics, and must pay close attention to the needs and values of the society in which they intend to use the software.

The second objection builds upon the final sentence in the paragraph above. When a programmer encodes ethics into a machine, who will ensure that these ethics are correct? In other words, who or what would guarantee that the programmer does not put forward their own personal agenda and values into the software? In order to answer this question, one has to bear in mind that the programmer cannot, in fact, neglect the needs and values of society. If the programmer neglects those values and injects his or her own personal beliefs into the machine, it is likely that the machine would act strangely and be unusable. Software containing an idiosyncratic set of values would be condemned by users and thus would not be successful. The manufacturer would also have a strong interest in ensuring that the consumer receives a desirable service. Hence, the software would need to be tested repeatedly, not only for safety and quality control, but also for ethical quality.

According to the third objection, coding ethics into a machine is too narrow; the program must learn its ethics by interacting with its environment. Instead of taking all the cues from the programmer, an intelligent AI should be able to learn what is right and wrong from its interaction with other people. The more people it interacts with, the better it becomes at learning right and wrong. This is just like how a child learns ethics—to live in a social environment with parents, siblings, friends, and so on. There is just no way for an algorithm to understand ethics through code alone. This is a valid objection, but the coding is only a part of the larger program, which involves teaching a machine to be compassionate. Since we do not have AGI level machines yet, we have to see how specialized, blind ASIs (artificial specialized intelligence) can exhibit behaviors that we deem to be (approximately) compassionate. At this stage, we would be glad if AI could deliver social good, even without being conscious. The AI could be encoded in such a way that it knows how to learn ethical principles. Humans are already hardwired to become ethical, since altruism and cooperation among members of our species has been fundamental throughout our evolution. After

all, understanding ethical and social cues would be a very strong achievement for AI, but would still require coding for this possibility to occur.

The final objection explains that coding ethics into a machine implies that programmers and software companies do not care for, and are not accountable to, society at large. Again, this does not have to be the case. There is no logical link between coding ethics into an algorithm and the programmer and employer being unaccountable to society. We have seen earlier that the programmer and software company must ensure that their products meet the requirements set by consumers and society; furthermore, they are still a part of society and need to follow specific laws and regulations.

The objections and comments from my colleagues largely focus on the relation of coding to its socio-economic context. This is an important matter, and in conclusion I would like to argue that coding must be embedded within its contexts. More specifically, this means that coding must only be one aspect of the overall systematic practice of ensuring that AI is ethical. Nevertheless, without an emphasis on coding, there is no definitive way in which the design of AI could directly contribute to a better society. For this to happen, the components of an ethical AI need to be translated into a language that a computer would understand. That is, the ethical components need to be made operationalizable, and they need to be pared down into basic steps for a computer to follow. Most importantly, the ethical vision must be clear, and the operationalization needs to adhere to it closely.

Conclusion and Recommendations

I would like to end this paper with a number of recommendations, both to public and private sectors, so that an ethical AI for social good can be fully developed and deployed. The recommendations are as follows:

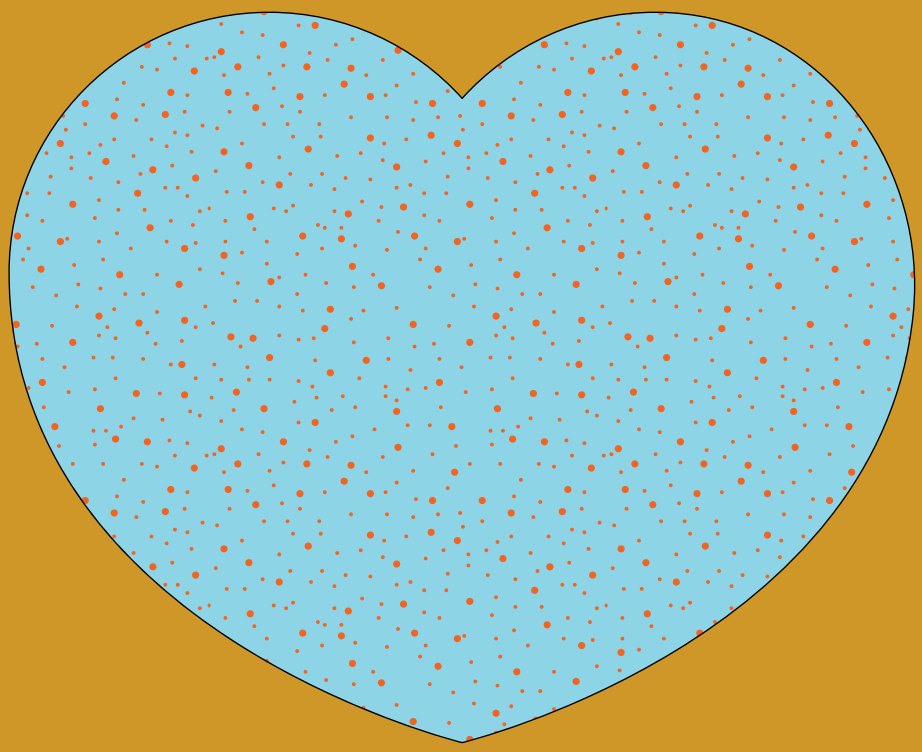
Recommendation 1: Programmers and software companies must implement compassionate AI programs, which is the key message of this article. No matter what kind of "social good" the AI is supposed to bring about, the software needs to be compassionate and ethical in the Buddhist sense. I have specified in some detail as to what being compassionate for AI actually means. Basically, the AI needs to realize that all things are dependent on all others (interdependence) and that the AI needs to show actual commitment to improving the condition of everyone in society (altruism). In order to make this recommendation feasible, the components of compassion need to be translated into algorithmic steps for the computer. In other words, the software needs to be coded in such a way that it becomes ethical. However, the coding must not be alienated from its socio-economic and historical contexts. That is, the software companies responsible for manufacturing AI programs must function as responsible and contributing members to society. No matter what kind of social good the AI is intended to bring about, this is a necessary requirement. The paper has shown that some applications that are being developed in the AI for Social Good program, such as flood forecasting, can indeed be used for nefarious purposes. This can happen when the information gained from the AI is used to gain unfair personal advantages. There should be ways within the design and programming of AI itself to prevent this, insofar as it is technically feasible. Abuses of flood forecasting information is an example of how the work of AI, which may originate from good intention, can be used in such a way that the AI itself becomes a culprit in an unethical action, such as hoarding or implementing flood prevention programs that privilege certain groups over others. Software companies need to be aware of this possibility and take the necessary steps to prevent it from happening.

Recommendation 2: The public sector needs to ensure that rules and regulations are in place in order to create an environment that facilitates the development of ethical AI for social good. Such rules and regulations will ensure not only that private companies have a clear set of directives to follow, but also public trust in the works of the private sector (assuming the work of creating AI software belongs to the private sector). Furthermore, even in a situation where the development of AI falls largely on the public sector, such as in Thailand, where the private sector is still rather weak in original research and development, the rules are also applicable. For example, the rules could provide incentives for software manufacturers to be more ethical. It needs to be made clear to all parties that there are material benefits to being more ethical. The belief that becoming ethical runs counter to profit maximization is shown to be unfounded. Realizing the objective of a private company must be embedded in the context of consumer trust; without the latter, it is hard to imagine how this type company could flourish in the long run.

These two recommendations make it clear that AI will create social good that truly answers people's needs and suffering. AI in the future may, or may not, become conscious and attain the level of superintelligence in the sense advocated by Nick Bostrom (Bostrom, 2014). In any case, AI needs to be made ethical at this time, as there is a decreasing window of opportunity to do so.

Acknowledgments

Many thanks to the Association of Pacific Rim Universities (APRU) for initiating the project on AI for social good. I would also like to thank Prof. Jiro Kokuryo of Keio University, Japan, the Principal Investigator of this project, for giving me the opportunity to become engaged in this exciting project. My sincere gratitude goes to Christina Schönleber, Director for Policy and Programs, APRU, as well as all my colleagues in the project, from whom I have learned a great deal. Thank you Prof. Pirongrong Ramasoota, Vice-President for Communication, Chulalongkorn University, and my colleague at Chula. Finally, I would like to thank Dr. Chulanee Tianthai, who gave me the information about this project and encouraged me to apply.



References

Berendt, B. (2019). AI for the common good?! pitfalls, challenges, and ethics pen-testing. *Paladyn, Journal of Behavioral Robotics*, 10(1), 44-65.

Bostrom, N. (2014) Superintelligence: *Paths, Dangers, Strategies*. Oxford University Press.

Cowls, J., & Floridi, L. (2018). Prolegomena to a white paper on an ethical framework for a good AI society. https://ssrn.com/abstract=3198732 or http://dx.doi.org/10.2139/ssrn.3198732

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707.

Gal, D., Perspectives and Approaches in AI Ethics: East Asia (June 7, 2019). Dubber, Markus, Pasquale, Frank, and Das, Sunit, (Eds.) *Oxford Handbook of Ethics of Artificial Intelligence, Oxford University Press*, Forthcoming. https://ssrn.com/abstract=3400816 or http://dx.doi.org/10.2139/ssrn.3400816

Green, B. (2019). "Good" is not good enough. AI for Social Good Workshop, NeurIPS (2019). https://www.benzevgreen.com/wp-content/uploads/2019/11/19-ai4sg.pdf

Hagendorff, T. (2019). The ethics of AI ethics: an evaluation of guidelines. *Arxiv.org*. https://arxiv.org/abs/1903.03425

Hongladarom, S. (2004). Growing science in Thai soil: culture and development of scientific and technological capabilities in Thailand. *Science, Technology and Society*, 9(1), 51-73.

Hongladarom, S. (2020). *The Ethics of AI and Robotics: A Buddhist Viewpoint*. Rowman & Littlefield.

Hughes, J. (2012). Compassionate AI and selfless robots: a Buddhist approach. In Patrick Lin, Keith Abney, and George A. Bekey (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge*, MA: MIT Press.

Kearns, M., & Roth, A. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019.

Mittal, V. (2017). Top 15 Deep Learning applications that will rule the world in 2018 and beyond. *Medium.com*. https://medium.com/breathe-publication/top-15-deep-learning-applications-that-will-rule-the-world-in-2018-and-beyond-7c6130c43b01

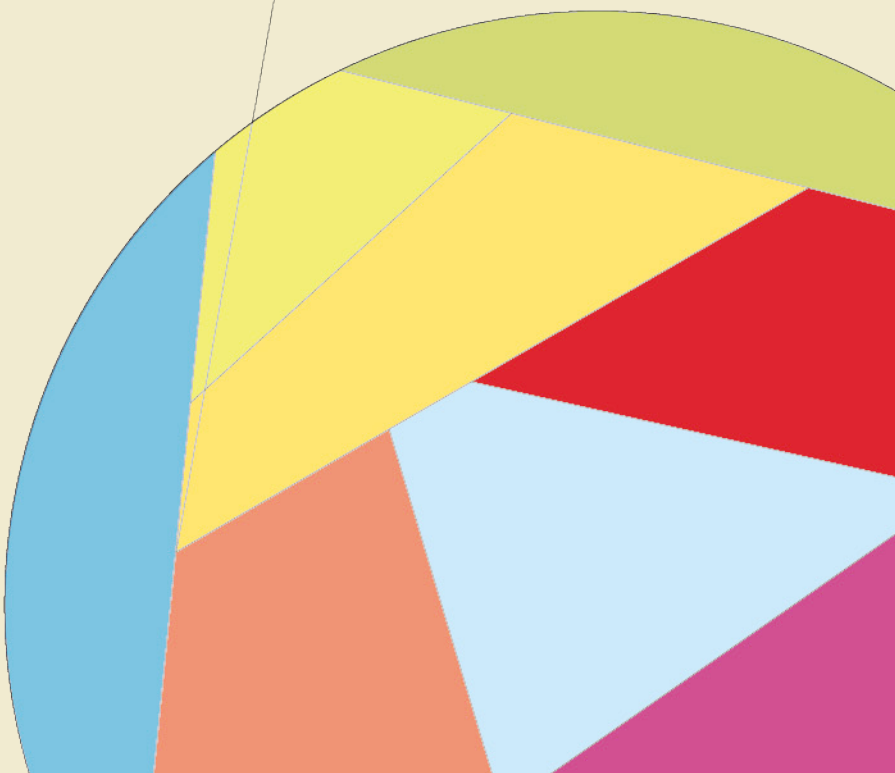
Taddeo, M., & Floridi, L. How AI can be a force for good. *Science* 361 (6404), 751-752. DOI: 10.1126/science.aat5991.

The European Commission's High-Level Expert Group on Artificial Intelligence: Draft Ethics Guidelines for Trustworthy AI. (2018) Working Document for Stakeholders' Consultation. Brussels, 18 December 2018.

Moralizing and Regulating Artificial Intelligence:

Does Technology Uncertainty and Social Risk Tolerance Matter in Shaping Ethical Guidelines and Regulatory Frameworks?

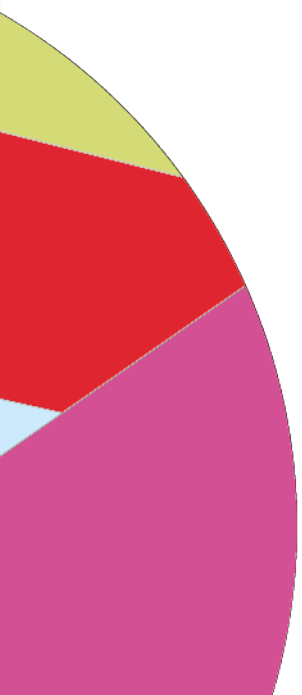
M. Jae Moon
Iljoo Park
Institute for Future Government,
Yonsei University



Introduction

Artificial intelligence (AI) is considered one of the most powerful developments in computer science, which affects every aspect and sector of society. While we are increasingly paying attention to its significance and impact, we do not yet know how and to what extent it affects the replacement and creation of jobs, industrial transformation, and lifestyle changes, which causes uncertainties and risks related to AI. Due to these underlying uncertainties and risks, there has been a growing demand for regulating and moralizing AI in order to minimize AI-caused uncertainties and risks. It is hoped that AI regulation will help to sustain its positive impact on society as a whole. With growing social fears and uncertainties, there has been increasing demand for a specific and proactive approach towards dealing with AI. Responding to these demands, governments and key international actors have attempted to provide regulatory frameworks and ethical guidelines for this rapidly developing technology. This study aims to review the uncertainty and risk issues of disruptive technologies such as AI, and assess their socio-economic and political impacts on society. This study will also discuss how key stakeholders (i.e., governments, industries, international organizations, NGOs, etc.) craft ethical guidelines/principles as well as review how different countries establish AI regulatory frameworks, particularly for autonomous vehicles (AVs).

Tzur (2017) argues that technological advancements fundamentally change the paradigm of regulatory mechanisms, while a conventional regulatory political framework (Wilson, 1980) seems to fail to offer an effective explanation for the nature of emerging disruptive technologies (i.e., AI, gene editing, blockchain, etc.), simply because defining who should benefit and who should bear the costs is quite uncertain and dynamic. Because of uncertainties regarding cost-benefit distributions as well as the opportunities and risks of emerging disruptive technologies, many countries appear to have adopted differing regulatory approaches to these technologies. For instance, national regulatory positions vary widely among different countries regarding the acceptance of cryptocurrencies (i.e., Bitcoins) as legal tender and the banning, regulation, or encouragement of cryptocurrency



exchanges. Notably, some countries such as Japan and the US have relatively light regulatory positions towards cryptocurrencies, while others including China and Korea have very restrictive policies. Likewise, regulations of disruptive technologies also differ in content and intensity from country to country. While some governments are in a strict regulatory position, others remain in an active deregulatory position by introducing regulatory sandboxes. Furthermore, the uncertainty and new forms of risk posed by these technologies (Slovic, 1987) demand social, industrial, and often international agreement, as well as discussion on ethical requirements and technological standards to ensure the maximization of social benefits and the minimization of social risks of these disruptive technologies.

In general, governments enact regulations to correct market failures, pursue collective and public interest goals, and to prevent potential social problems caused by the excessive pursuit of private interests. However, individual regulations do not always meet public expectations or help achieve intended social goals. Regulatory decisions on disruptive technologies are often not timely, primarily because of the lag between the emergence of technology-driven social issues and regulatory policy decision-making. Views regarding the regulation of novel technologies also often vary widely because of country-specific contextual factors—including legal systems, influence of various interest groups, and the ethical perspectives of the general public, which determines the social risk perceptions of the public.

This study uses cross-country comparative case studies by examining the similarities and differences of regulatory actions caused by levels of certainty, as well as the tolerance of social risks for technologies in given countries. As an example, this study will examine regulatory approaches to AVs, which is a product of AI and robotics technology. We will examine the US and three Asian countries, namely China, Japan, and Korea. The aforementioned Asian countries are major economic players in the region, and are all interested in disruptive technologies for the potential implications of economic and social development. The US has been included as a base for comparison since it is more market-oriented than other countries, while the three Asian countries are somehow paternalistic.

Due to the disruptive nature of emerging technologies such as AI and related technologies including robots, AVs, drones, etc., there is no particular consensus regarding how disruptive technologies should be regulated and moralized through social interests in those technologies, as well as research interests in the intertwined relationship between technological advancements and regulations. Despite growing interest in disruptive technologies and related ethical guidelines and regulations, limited research has been conducted in this field. In particular, a comparative analysis of ethical guidelines for AI and different national responses to disruptive technologies have been somewhat lacking, primarily because there is no clear measure of regulatory stringency as the basis for comparative studies of regulation politics (Brunel & Levinson, 2013). In order to fill this research gap, this study aims to look at key ethical elements of AI, and then determine how and why countries develop different regulatory approaches to the same technologies.

Along with the growing interests in AI, governments, research institutes, international organizations, and industries initially began to pay attention to ethical frameworks for AI, as many are puzzled about the potential consequences and ethical dilemmas. For example, an ethical dilemma on this subject is how an autonomous vehicle should deal with an unavoidable accident, where the car must decide whether to kill an innocent bystander or the five passengers inside the vehicle. It is also imperative to question who

should be responsible for an incident involving an autonomous vehicle, among AI programmers, vehicle manufacturers, vehicle sellers, drivers, and others.

As proposed by a group of experts on AI commissioned by the OECD, an ethical guideline specifies and addresses core values in developing, manufacturing, and using AI and AI-loaded machines. In fact, it will not be long before ethical guidelines and principles for AI are offered by governments, international organizations, private companies, and NGOs. Reviewing 84 documents of ethical principles and guidelines, Jobin et.al. (2019) found that most of these documents (88%) were released after 2016 by private companies (22.6%) and government agencies (21.4%).

We will first discuss technology uncertainty and social risk in the context of disruptive technologies. Then, we will review the development of ethical guidelines for AI developed by different actors as a loosely institutional effort to moralize AI technologies. Next, we specifically examine the different regulatory positions of four selected countries to AVs. Finally, policy implications are discussed and policy recommendations are presented.

Determinants of Regulating and Moralizing Disruptive Technologies: Technology Uncertainty and Social Risk Tolerance

Disruptive technologies: benefits and risks

Since being presented by the World Economic Forum in 2016, there has been a growing interest in disruptive technologies which are often proposed as technological engines for the fourth industrial revolution. Figure 1 shows the different levels of expected benefits and costs from each technology. The World Economic Forum (2016) surveyed professionals in each country, asking about their perceptions of the benefits and negative consequences of 12 major emerging disruptive technologies. Participants perceived AI and robotics as the most beneficial and risky technologies, while they perceived blockchain technology as moderately beneficial and risky. Moreover, people tend to perceive both biotechnologies and neuro-technologies to be more beneficial and riskier than blockchain technology.

Despite variations in the perceived benefits and risks of those disruptive technologies, many stakeholders have raised their concerns over the potential risks of such technologies. As such, they have demanded for alternative ways of moderating and minimizing the risks, which often results in informal/unofficial forms of ethical principles and formal/official forms of regulation. While the former is presented as a set of soft, suggestive, and general principles, the latter is a set of hard, legally binding, and specific rules. The former is discussed and manufactured by various stakeholders of different sectors (private, non-profit, and public sectors) at different levels (i.e., local, national, and international), whereas the latter tends to be made by executive or legislative branches through formal rule-making and legislative processes, because each country makes its own regulatory decisions as technological risks and interest conflicts among stakeholders gradually mount. Recently, ethical standards and regulations have been discussed and proposed in the European Union (EU), the OECD, and other economic communities to moralize as well as control (regulate) technologies. While there is a general consensus in the nature and scope of ethical principles for AI, there is no consensus in regulatory frameworks among different countries. Moreover, the governmental regulatory decision can fall even farther behind when the potential costs and benefits of a technology are uncertain.

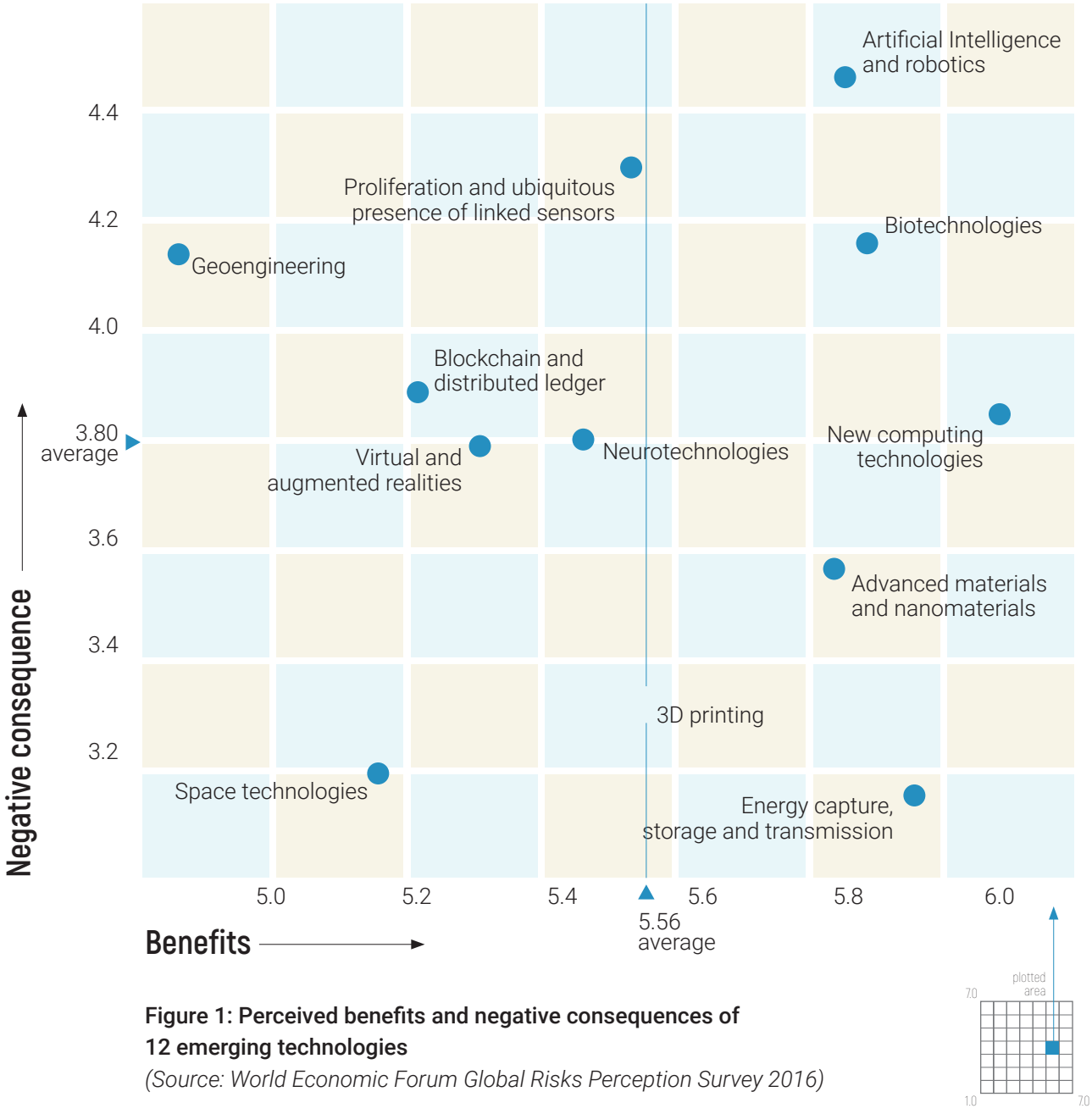


Figure 1: Perceived benefits and negative consequences of 12 emerging technologies

(Source: World Economic Forum Global Risks Perception Survey 2016)

Regulatory lag and regulatory paternalism

Regulators are often uncertain as to whether or how to address the risks (World Bank, 2016). In particular, regulators are uncertain and unclear about assessing the potential benefits and risks of emerging technologies, which makes regulating disruptive technologies even more challenging than conventional technologies (Hunt and Mehta, 2013). Generally, regulations tend to be reactive rather than proactive, which often causes regulatory lag. While regulatory lag is partially a result of market-based and non-interventionistic policy position, it often causes tardy responses to previous problems that could have been addressed in advance.

On the contrary, regulatory paternalism also plays an important role in driving proactive regulations to minimize potential risks. Paternalism originally referred to the ideological belief that governments should intervene to protect people—similar to protecting their children. Thus, regulatory paternalism involves paternalistic regulatory action on the part of governments. Paternalism lies behind many regulatory measures beyond specific instances (e.g., seatbelt and safety helmet laws); it is also the driving force behind the prohibition or control of certain risk-generating products and services. In fact, citizens of contemporary risk-obsessed societies expect their governments to provide them with protection (Ogus, 2005). To overcome excessive regulations formulated by regulatory paternalism, some countries have recently adopted temporary deregulation schemes such as a regulatory sandbox, which is a testing ground that is protected against any possible regulation. This supports a flexible and lenient regulatory position to maximize potential economic and social benefits of various disruptive technologies.

Determinants of cross-country regulation differences

Based on the "psychometric paradigm," Slovic, Fischhoff, and Lichtenstein (1982) conducted a classical study regarding the risk perception of people and offered a solid framework to understand the cross-country regulation difference on disruptive technologies. They suggest two significant factors to distinguish technologies: dreadfulness and unfamiliarity. Dreadfulness refers to the extent to which a technology can be controlled not to be catastrophic, which is understood as a measure for technological risk. Unfamiliarity refers to how much a technological risk is observable, which is considered as a technology uncertainty. It implies that subjective perception is an important factor to the classification of technologies besides objective criteria. It should be noted that these terms are not absolute, and instead used as relative terms. For instance, nuclear power can be a more dreadful and less unknown technology than dynamite, which is a less dreadful but more known technology.

While "uncertainty" and "social risk" are considered to be independent, they are somewhat related since technology uncertainty often causes a higher level of social risk of a particular technology in a society. As a result, the social tolerance of a particular risk would be a significant factor in a country since the response to one technology would be different for other countries, although the objective technological risk would be identical. This leads to specific regulatory positions for different technologies because certain countries may want to control the potential technological risk and take various regulatory measures (e.g., law enactments) to restrict the reckless research, development, and utilization of technology.

1. Technology uncertainty

Technological "unfamiliarity" (Slovic et.al., 1982) is somewhat similar to technology "uncertainty", though the term "uncertainty" may not be used in a strictly defined sense since it is commonly used by many people in different senses (Downey and Slocum, 1975) or often poorly understood (Fleming, 2001). Despite this poor understanding of "uncertainty", it is generally accepted that the degree of technology uncertainty may vary depending on controllability, which is directly related to the level of safety and potential risk of a particular technology. According to Milliken (1987), the three common definitions derived from psychology and economics for "uncertainty" are (1) "an inability to assign probabilities as to the likelihood of future events", (2) "a lack of information about cause-effect relationship", and (3) "an inability to predict accurately what the outcomes of a decision might be". Similarly, we can define technology uncertainty as "the inability to measure the likelihood of a future event and the outcome with probabilistic function and to infer the causal outcome made by a particular disruptive technology".

We argue that *uncertainty about the spillover effects from technologies themselves results in* cross-country variation in regulatory decisions on disruptive technology. For example, the difficulty of predicting the costs and benefits of a technology causes regulatory lag since this can obstruct timely regulations. Governments are likely to identify disruptive technologies based on the extent to which the expected costs and benefits are easily measured. If the costs and benefits derived from a technology can be predicted quickly, the regulatory policies can be developed more promptly. Otherwise, governments may postpone strict regulatory decisions if a technology has the potential to cause harm in ways that cannot be foreseen during the innovation process, preventing them from quickly predicting the costs and benefits the technology could generate. We define such technologies as "uncertain technologies". It should be noted that the regulation of uncertain technologies is also affected by the degree of uncertainty that a particular society should and can tolerate (Kolacz et al., 2019).

In contrast to the uncertainty of expected outcomes from any given technology, *responsiveness to the global consensus* is a significant factor for converging similar regulatory positions. Although it may be challenging to make a public consensus between scientists and the general public (Kahan, Jenkins-Smith, & Braman, 2011), the existing consensus or standards can apply to regulatory decisions regarding emerging technologies. Recently, a global consensus led by international and regional organizations such as the EU, the OECD, and the WHO has also been made, which shapes the nature of regulatory positions of countries that are not necessarily obligated to follow the global standard (Kerwer, 2005).

2. Social risk tolerance

Another reason for differences in regulatory responses between countries is that some countries have different levels of tolerance for social risks. Uncertainty of one technology makes people eager to prepare for potential risks or hazards. We focus on the fact that the preparation level for an uncertain technology can differ depending on the country. Social risk tolerance is closely related to uncertainty avoidance; people who prioritize avoiding uncertainty are likely to control uncertain situations by imposing strong schemes such as regulations. Empirical studies in various areas — e.g., Kanagaretnam et al. (2011) — examine the relationship between high risk perception and low uncertainty avoidance.

Hofstede's 6-D model of national culture is considered one of the major measurements of the general public's uncertainty avoidance. It attempts to measure the degree to which members of a society feel uncomfortable with uncertainty and ambiguity (Hofstede, 2015). According to Hofstede's score out of 100, Japan (92) and Korea (85) have somewhat higher uncertainty avoidance than China (30) and the US (46). Note that the interpretation of this index has been made cautiously because Hofstede originally developed his theory from a management perspective to recognize the difference between diverse cultures. That said, it helps to draw a better understanding of the cultural differences among countries in many aspects, such as uncertainty avoidance. Uncertainty avoidance is different to risk avoidance, but is related

to anxiety and distrust towards the unknown (and vice versa), with the desire to have fixed practices and rituals as well as understanding reality (Hofstede, 2015).

Exploring the determinants of social risk tolerance levels could provide substantial insight into cross-country differences in regulatory decisions regarding disruptive technologies; however, discussion of such an approach in prior research is scarce. We identify the following three main factors that define countries' different tolerance of social risk: (1) *legal traditions and the efficiency of legally challenging regulations*, (2) *competition among interest groups*, and (3) *ethical concerns*.

First, *legal traditions and efficiency of legally challenging regulations* can generate differences in regulatory decisions among countries. Numerous studies, including Beck et al. (2002) and Hail and Leuz (2006), examine the relationship between countries' legal origins and levels of economic development, finding the nations' legal origins significantly impact their financial development. In particular, Beck et al. (2002) suggests that differences in countries' legal origins help explain differences in their levels of financial development.

Furthermore, some empirical studies have identified differences between common law and civil law countries in terms of regulation decisions. For instance, Djankov et al. (2002) finds that, at comparable levels of development, French civil law countries tend to have heavier regulations, less secure property rights, and fewer political freedoms than common law countries. Moreover, Charron et al. (2012) also mention that countries' legal origins could explain cross-country differences in judicial independence and government regulations of economic life, which can be summarized as the quality of institutions, as well as low degrees of corruption and high degrees of the rule of law, which in essence are desirable social and economic outcomes. They suggest that because of stronger legal protections for outside investors and less state intervention, countries with a common law tradition have achieved higher economic prosperity and quality life than civil law countries. La Porta et al. (2008) even summarize their series of articles (La Porta et al. 1997, 1998, 1999) to address the prevalent impact of a wide range of desirable organizations and social outcomes of nations' legal traditions and other related articles to develop a so-called "Legal Origins Theory" (Charron et al. 2012).

Competition among interest groups can also generate differences in countries' regulatory decisions. Gai et al. (2019) explain that regulatory complexity is a consequence of lobbying. They focus on the fact that lobbyists may be able to persuade policymakers or politicians to give their interests to more favorable regulatory treatment, which leads to additional complexity and fragmentation across countries, especially when it comes to financial regulation. In addition to the appeals of individual groups, conflict among many interest groups can significantly affect countries' regulatory decisions. For instance, interest-group politics are heavily involved in cryptocurrency regulation; debates regarding the use of cryptocurrency worldwide is intense, and many stakeholders are involved in this discussion. According to Houben and Snyers (2018), numerous players are involved in the cryptocurrency debate and they all play particular roles: cryptocurrency users, miners, cryptocurrency exchanges, trading platforms, wallet providers, coin inventors, and coin offerors. In addition to these players, policymakers such as the International Monetary Fund (IMF), the Bank for International Settlements, and the World Bank have their own views on cryptocurrency. The groups who utilize cryptocurrency are expected to experience the associated benefits, costs, and discussions, which are still ongoing.

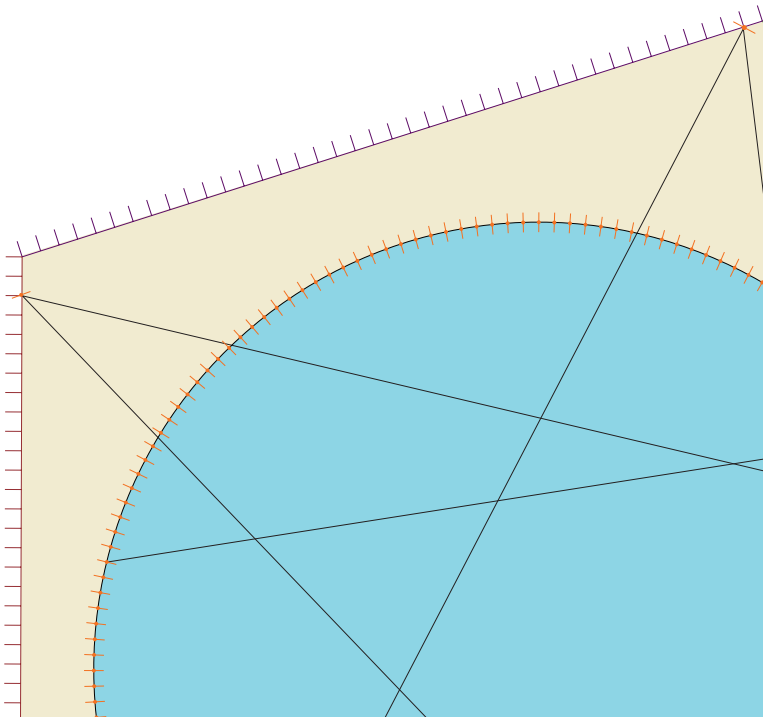
Ethical concerns can also lead to differences in countries' regulatory decisions. Such concerns may be related to general public safety or the religious views of various groups. In particular, regulations regarding genetically modified organisms (GMOs) are affected by the ethical perspectives of countries' citizens. Such perspectives can be affected by religious beliefs or the general views of human morality. Globus and Qimron (2018) investigate the regulations and

cultural perceptions of different countries regarding GMO approval. Their study found that regulatory and supervisory procedures for GM crops and the foods produced from these crops differ because governmental approaches represent the differing responses of citizens and scientific communities. These policies also reflect a variety of cultures, environmental conditions, political pressures, and the interests of different groups such as farmers, agricultural companies, and environmental activists or agencies.

To summarize, we suggest that the regulation of disruptive technology might vary as a result of technology uncertainty and social risk tolerance, and that several socio-economic factors may generate variation in uncertainty and risk tolerance. Two different approaches have been suggested: (1) moralizing technologies based on ethical standards and (2) regulating technologies based on legal mechanisms. The former refers to the efforts of various stakeholders to promote desirable status or conditions through codes of conduct or moral principles, which are often voluntary instead of mandatory. The latter refers to legal actions by governments to mandate and enforce particular actions, or to prohibit illegal actions which in many cases lead to penalty or punishment. In the next section, we examine the evolution of ethical principles for AI and then survey regulatory actions regarding three selected disruptive technologies that pose different degrees of risk in four developed countries.

	Ethical Approach	Legal Approach
Mechanism	Ethical standards	Regulatory laws
Actor(s)	Various stakeholders	Government(s)
Nature	Voluntary; Broadly defined and widely applied	Mandatory; specifically defined and narrowly applied
Consequences	Moral blaming	Punishment or penalty

Table 1: Comparison of ethical approach and legal approach



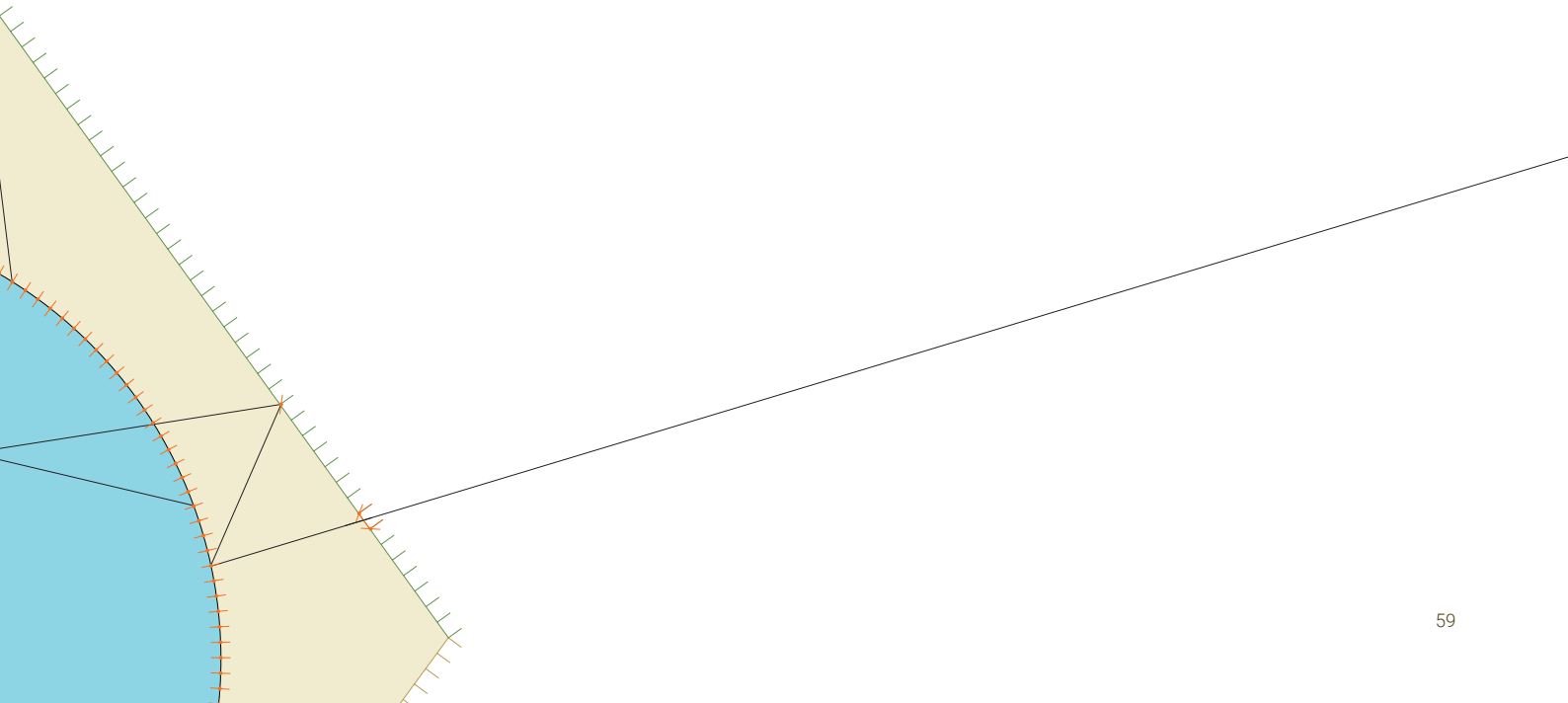
Moralizing Disruptive Technologies: Ethical Guidelines and Principles for AI

Ethical AI (Jobin et.al., 2019), trustworthy AI (European Commission, 2018), and responsible AI (Microsoft, 2018) have been proposed and discussed among various stakeholders (e.g., academics, industries, governments, and international organizations), as AI was presented as a main driver for radical and disruptive changes (Jobin et.al., 2019). Although terms such as "ethical", "trustworthy", and "responsible" are used in documents that cover ethical guidance and principles, they all explain that we must handle AI in a lawful, ethical, and robust way throughout its entire lifecycle. Such guidelines include design, development, deployment, and usage (European Commission, 2018) by recognizing, preparing, and resolving the potential risks and negative impacts of AI in a society.

Ethical AI is often considered as a starting point for moderating any potential negative social and economic impacts of AI and AI-loaded devices, including automation and job replacements, intentional misuses and malevolent consequences, dissemination of social bias and its reinforcement, and an undermining of fairness (Jobin et.al., 2019). Reviewing and scoping 84 documents of ethical guidelines and principles, Jobin and her colleagues (2019) suggest that several key ethical principles are commonly identified including transparency, justice and fairness, non-maleficence, responsibility, and privacy. That said, there is no consensus on how these principles are interpreted and applied in the course of designing, developing, and using AI and AI-loaded devices.

Presenting trustworthy AI, the European Commission (2018) proposed three elements constituting trustworthiness including lawful AI, ethical AI, and robust AI. Lawful AI refers to the fact that AI should be bound by existing legal systems of local, national, regional, and international levels so that they bind any processes and activities involving the entire AI lifecycle. The European Commission (2018) suggests that lawful AI "should not be interpreted with reference to what cannot be done, but also with reference to what should be done and what may be done". In addition to legal compliance as a basic minimal requirement, ethical AI emphasizes the reference of ethical norms in particular because legal systems are often far behind and do not keep up with technological developments. Robust AI is presented to avoid or minimize the possible unintended negative consequences of AI in a society.

As shown in Figure 2, the European Commission (2018) suggests that all stakeholders including developers, deployers, and end-users should meet critical requirements for realizing trustworthy AI. Seven requirements are presented as follows: (1) human agency and oversight (fundamental rights, human agency, and human oversight); (2) technical robustness and safety (resilience to attack and security, fallback plan and general safety, accuracy, and reliability and reproducibility); (3) privacy and data governance (privacy and data protection, quality and integrity of data, and access to data); (4) transparency (traceability, explainability, and



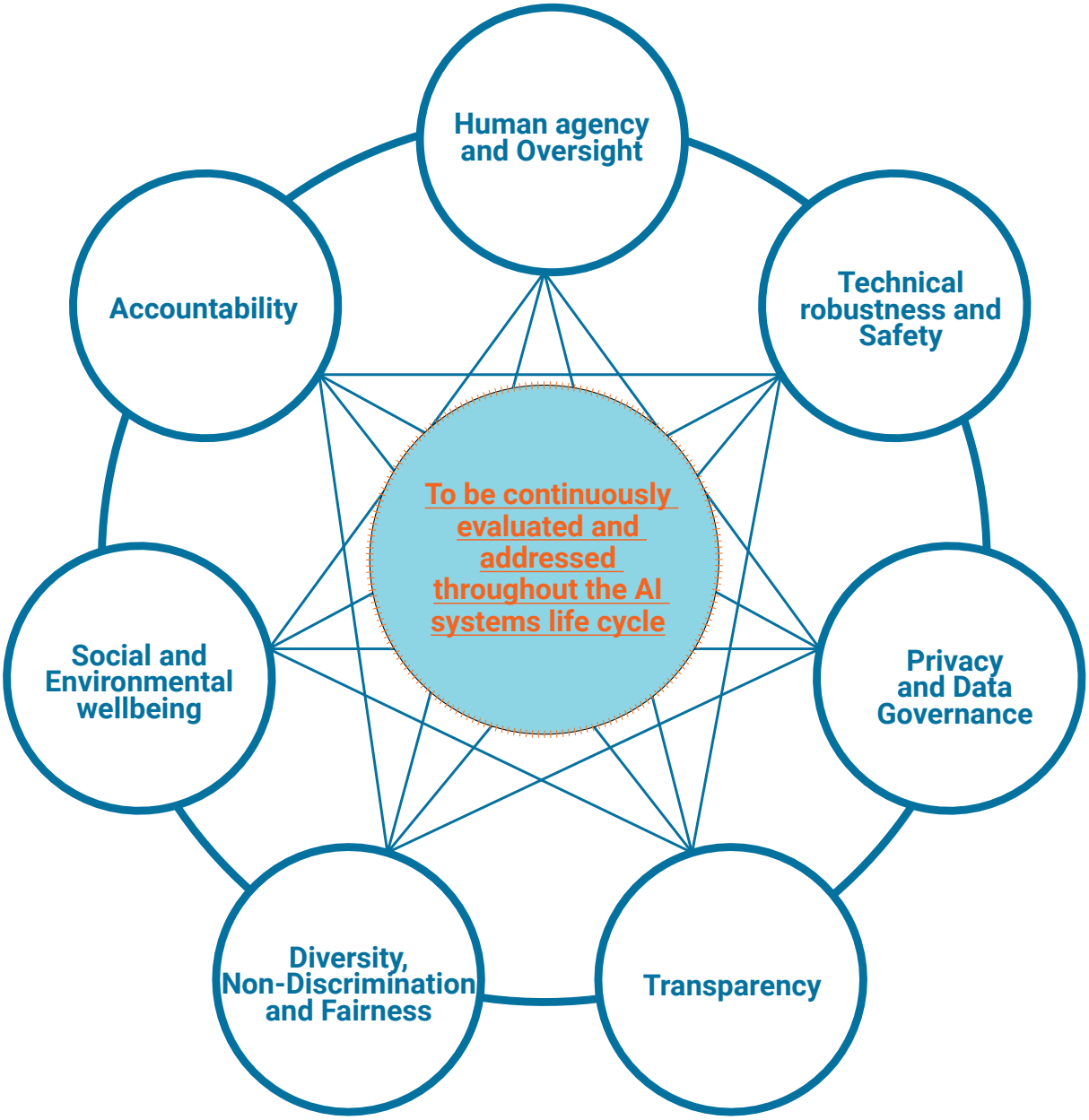


Figure 2: Seven requirements for trustworthy AI and their interrelationship
(Source: European Commission (2018), p. 15.)

communication); (5) diversity, non-discrimination and fairness (avoidance of unfair bias, accessibility and universal design and stakeholder participation); (6) societal and environmental wellbeing (sustainable and environmentally friendly, social impact, and society and democracy); and (7) accountability (auditability, minimization and reporting of negative impacts, trade-offs, and redress) (European Commission, 2018).

Similar to the Ethics Guidelines for Trustworthy AI by the European Commission, many organizations and governments have offered ethics guidelines and principles for AI. As summarized in Table 2, many documents have been formulated by private companies, government agencies, and academic institutions; many of which were formed in the US, UK, and EU institutions. Table 2 shows the breakdown of ethical guidelines and principles for AI by type, geographical location, and target audience.

Type and Geographical Location	Classifications
Type of Issuing Organizations*	19 private companies (22.6%), 18 government agencies (21.4%), 9 academic and research institutions (10.7%), 8 inter-governmental or supra-national organizations (9.5%), 7 non-profit organizations and professional associations (8.3%), 4 private sector alliances (4.8%), 1 research alliance (1.2%), 1 scientific foundation (1.2%), 1 federation of worker unions, 1 political party, 4 others
Geographical Location of Issuing Organizations**	20 USA (23.8%), 16 international organizations, 14 UK (16.7%), 6 EU institutions, 4 Japan, 3 Germany, 3 France, 3 Finland, 2 Netherlands, 1 Iceland, 1 India, 1 Singapore, 1 Norway, 1 South Korea, 1 Spain, 1 UAE, 1 Australia, 1 Canada
Target Audience***	27 for multiple stakeholder groups (32.1%), 24 for own employees of companies (self-directed) (28.6%), 10 for the public sector (11.9%), 5 for the private sector (6.0%), 3 for developers or designers (3.6%), 1 for organizations, 1 for researchers

Source: Compiled by author from Jobin et.al. (2019).
* 4 documents are double counted and 4 are not classified
** 3 are not classified
*** 13 not classified.

Table 2: Ethical guidelines and principles by type and geographical location

Based on content analysis, Jobin and her colleagues identified 11 key ethical principles along with related values. Some key findings on ethical principles from the content analysis by Jobin and her colleagues (2019) are summarized in the following table. As the table indicates, transparency and related values (73/84) appeared the most, followed by justice/fairness (68/84), among 11 key ethical principles including transparency, justice/fairness, non-maleficence, responsibility, privacy, beneficence, freedom/autonomy, trust, sustainability, dignity, and solidarity. Non-maleficence and responsibility are also primary principles which are found in 60 out of 84 documents.

Ethical Principles	No. of Documents	Related Values
Transparency	73	Explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice/fairness	68	Consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60	Security, safety, harm, protection, precaution, prevention, integrity, (bodily or mental), non-subversion
Responsibility	60	Accountability, liability, acting with integrity
Privacy	47	Personal or private information
Beneficence	41	Benefits, well-being, peace, social good, common good
Freedom/autonomy	34	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28	
Sustainability	14	Environment (nature), energy, resources
Dignity	13	
Solidarity	6	Social security, cohesion

Table 3: Ethical principles and related values
(Source: Jobin et.al. (2019), p. 7.)

As noted in the earlier section, international organizations such as the EU have been actively working on formulating ethical guidelines for AI. For example, the European Parliament took an initial action by asking the European Commission to assess AI's social impacts, which led to a set of "recommendations on civil law rules on robotics" in early 2017 (Madiega, 2019). This was followed by the Commission's coordinated plan on AI for EU member countries, which was later endorsed by the EU Council and then became a foundation for the Commission's Ethics Guidelines for Trustworthy AI (Madiega, 2019). The guideline formulated by the High-Level Expert Group on AI of the Commission is considered one of the most comprehensive frameworks for offering critical principles that various stakeholders should consider in designing, developing, and deploying AI. In particular, the guideline emphasizes the core nature of a "human-centric approach", which has been widely accepted beyond the EU. The nature of this human-centric approach to AI is summarized as follows:

> *The human-centric approach to AI strives to ensure that human values are central to the way in which AI systems are developed and deployed, used and monitored, by ensuring respect for fundamental rights, including those set out in the Treaties of the European Union and Charter of Fundamental Rights of the European Union, all of which are united by reference to a common foundation rooted in respect for human dignity, in which the human being enjoys a unique and inalienable moral status. This also entails consideration of the natural environment and of other living beings that are part of the human ecosystem, as well as a sustainable approach enabling the flourishing of future generations to come.*[1]

Emphasizing the lawfulness, ethics, and robustness of a trustworthy AI system from a lifecycle perspective, the guideline essentially promotes ethical principles for ensuring reliable and trustworthy AI. The guideline emphasizes seven key requirements for EU member countries including (1) human agency and oversight, (2) robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental well-being, and (7) accountability (Madiega, 2019).

Regulating AI: The Case of Autonomous Vehicles in Different Countries

As noted earlier, regulatory instruments and levels of regulation vary widely from country to country. We conduct an exploratory comparison of the regulatory approaches of four major countries—China, Japan, Korea, and the US—in terms of the regulatory intensity of AVs. The three Asian countries were selected because they are considered as economic leaders, while also representing countries at different levels of economic development in the region. The US was selected as a basis for comparison, as the country represents market-based and relatively non-interventionist regulation policies.

1. Current status of autonomous vehicle technology development

An autonomous vehicle (AV) is a vehicle that can navigate by itself without human intervention (Taeihagh & Lim, 2019). According to SAE International (originally the Society of Automotive Engineers), automated driving can be divided into six levels, from 0 to 5 (the higher the level, the more automated the vehicle), based on the level of sophistication and automation. As Figure 3 summarizes, AVs are equipped with various autonomous features for driver supporting systems ranging from automatic emergency breaking (Level 0) to lane centering systems (Level 2: partial "hands off" automation), while "automated driving systems" also range from traffic jam "chauffeurs" (Level 3: conditional "eyes off" automation) to the highest level of complete driverless taxis in all conditions (Level 5: full "steering wheel" automation) (QVRTZ, 2019). Several carmakers, including Waymo, are already using level 4 AVs in some areas for ride-sharing or delivery services, but these vehicles have not yet entered the retail market. It has been said that substantive impact of AVs might begin when driverless automobiles are introduced in local areas.

1. Glossary section of the Ethics Guidelines for Trustworthy of AI (2019). Requoted from Madiega (2019), p. 3.

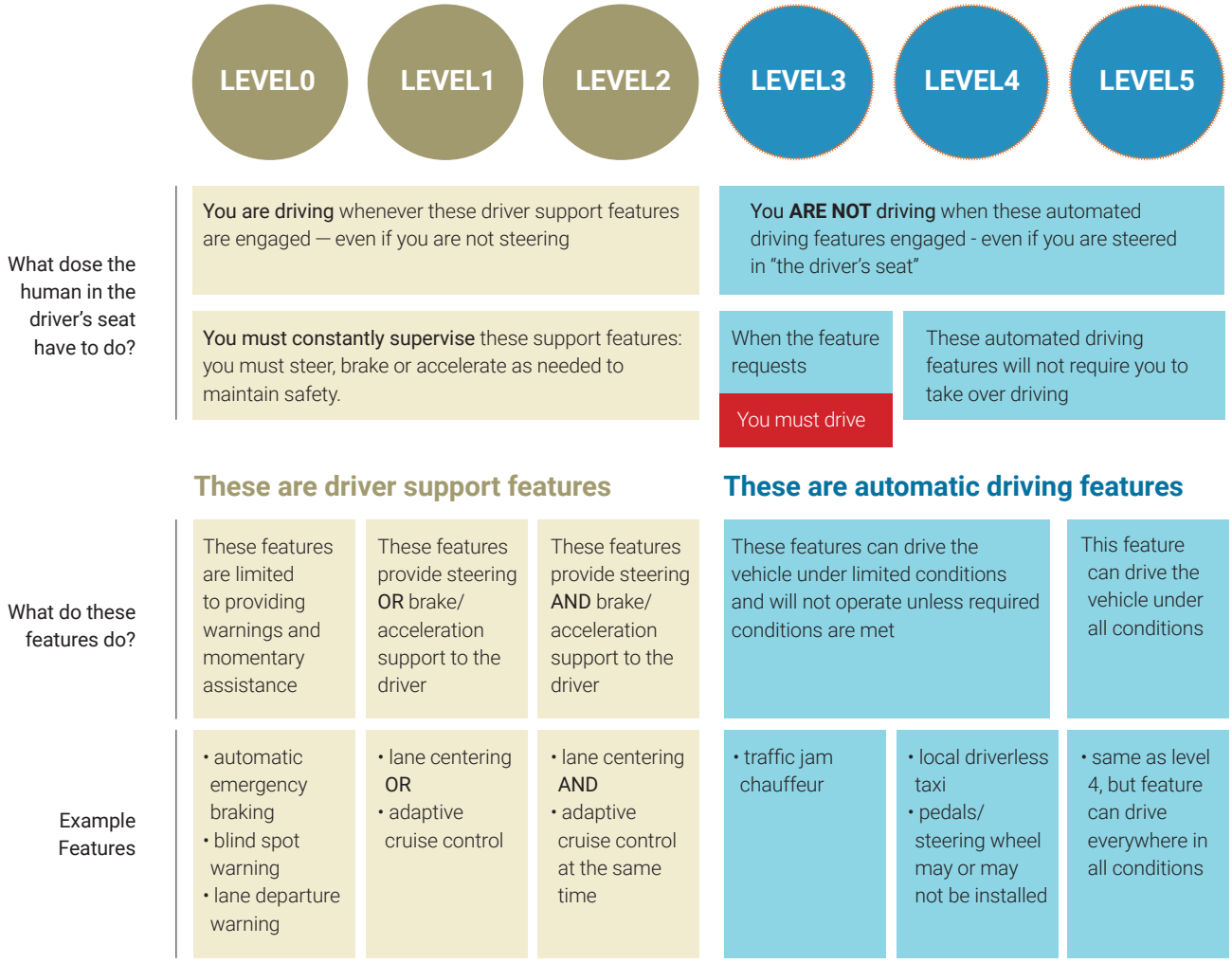


Figure 3: Levels of autonomous vehicles
*(Source: SAE (2018). **https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-standard-for-self-driving-vehicles**)*

2. Regulating autonomous vehicle

Table 4 presents a cross-country comparison of the specific regulations for AVs, particularly focusing on AV driving in China, Japan, Korea, and the US. We consider four regulatory issues: (1) whether the government permits autonomous driving, (2) whether the enforcement is legally binding, (3) whether the government can hold people liable based on laws or guidelines, and (4) whether the government provides any guidelines for users. We will not discuss license issues, since it has been debated at national levels. It should also be noted that no global consensus currently exists and nation states generally have strict requirements for drivers.

The three Asian countries under examination have prohibited autonomous driving when the driving is not for testing, and enforcement is legally binding. The US, however, has placed no strict restraints on autonomous driving; a bill that would establish the federal government's role in ensuring the safety of highly automated vehicles has been referred to a federal committee. All countries except China can hold persons (rather than AVs) liable based on these laws or guidelines; as it stands, China has no official guidelines regarding the issue. Furthermore, people who want to take autonomous driving tests

	Prohibiting free autonomous driving itself	Legally binding enforcement	Holding persons liable based on the laws or guidelines	Offering guidelines for users
China	Yes	Yes	No	Yes
Japan	Yes	Yes	Yes	No
Korea	Yes	Yes	Yes	No
US	No (No strict restraint)	No (Referred the bill to the Committee)	Yes (Those who want to test AVs should obtain the state-designated insurance)	Yes

Table 4: The status of autonomous vehicle driving regulations (as of August 2019)

must obtain state-designated insurance in Korea. Governments' provision of user guidelines for autonomous driving demonstrates their interests in the development of autonomous driving technology and commercialization. China and the US have user guidelines while Japan and Korea do not.

The US Congress passed a bill titled the *Safely Ensuring Lives Future Deployment and Research in Vehicle Evolution Act* (more commonly known as the "SELF DRIVE Act") in 2017. Proponents of the bill claim that by encouraging the testing and deployment of AVs, the bill establishes a federal role in ensuring the safety of highly automated vehicles. It has been received in the Senate, read twice, and referred to the Committee on Commerce, Science, and Transportation (The US Congressional Research Service, 2019). In addition to this bill, the US is the first country to introduce legislation to permit the testing of automated vehicles (UK Department for Transport, 2015). It has also introduced "A Vision for Safety 2.0," federal guidelines for the automobile industry and individual states regarding automated driving systems (ADSs) that builds on the National Highway Traffic Safety Administration's 2016 guidelines. This document has two sections—voluntary guidance and technical assistance for states. The new guidelines focus on Levels 3 to 5 of the SAE International's automation classification, stipulating that entities do not need to wait to test or deploy an ADS, revising the elements of safety self-assessments, aligning federal guidelines with the latest developments and terminology, and clarifying the role of the federal and state governments. The guidelines emphasize their voluntary nature and do not include with compliance requirements or enforcement mechanisms. They represent an attempt to establish best practices for state legislatures, outlining the common safety-related components of ADSs that states should consider incorporating into their legislation. Additionally, they include the US Department of Transportation's view regarding federal and state roles and offers best practices for highway safety officials.

China is also preparing regulations to ensure safe AV testing. Notably, Chinese regulations and policies regarding autonomous driving are seen as relatively moderate compared to their strict control of some other aspects of driving, such as restrictions stating that public maps can only be accurate to a scale of

50 meters at most, and that drivers must keep both hands on the steering wheel at all times (KPMG International, 2018). The road-testing regulation was established in April 2018 and the guidelines for building safe, closed test sites were released in July 2018 (Xinying, 2019). The Chinese do not appear to be very concerned with safety and liability issues; their concerns focus on the technological availability of AVs and economic consideration related to their use (Dickinson, 2018).

Likewise, Japan is preparing the commercialization of level 3 AVs and will enact a new legal amendment for autonomous driving. The National Diet of Japan passed a bill amending the current Road Transport Vehicle Act to include "automatic operating devices" as a vehicle in May 2019. In addition, it passed another bill that allows people to use level 3 AVs in certain conditions and to use cell phones during autonomous driving (Matsuda et al., 2019). Although there has been some progress in AV-related regulations thanks to the May 2019 amendments of Japan's Road Traffic Act, Matsuda and his colleagues (as quoted below) stressed that there are still several issues to be resolved in future.

> *"… One of the main outstanding issues is determination of the rules for criminal and civil liabilities in the event of a traffic accidents involving self-driving vehicles. Because these provisions have not yet been updated, a driver may still be held responsible for criminal or civil liabilities for a traffic accident caused by a vehicle under automated driving even if the driver operated the self-driving vehicle properly. This issue affects not only drivers but also manufactures and insurance companies, and is therefore likely one of the thornier issues remaining to be resolved" (Matsuda et al., 2019).*

In Korea, the Road Traffic Act, Automobile Management Act, and Automobile Damages Guarantee Act currently regulate the use of automobiles, but that will change in 2020 when the Act on the Promotion and Support of the Commercialization of Self-driving Cars comes into force. The Road Traffic Act regulates traffic problems and establishes rules for safe operation. It presumes the presence of a driver who is required to manipulate the steering wheel and braking system. However, the Automobile Control Act defines AVs as cars that can be operated without any driver or passenger input. The Enforcement Rules of the Act, enacted in 2016, specify the requirements for the safe operation and testing of AVs, meaning that the laws are in conflict with each other to some extent regarding whether "a driver" can refer to an automated system. At present, the Ministry of Land, Infrastructure, and Transport requires a temporary operation permit for the testing of AVs, and the "Requirements for Safe Operation of Autonomous Vehicles and Trial Operation Regulations (as of March 31, 2017)," stipulate that a preliminary test of 5,000 km must be conducted (Ministry of Science and ICT and KISTEP, 2018).

The KPMG International's annual reports provide insight into the current state of AV testing. The reports evaluate countries' AV readiness and AV testing restrictions, giving countries scores out of seven based on reviews of media articles, government press releases, and government regulations. A higher score indicates that the country's regulations support AV use and impose fewer restrictions on when, where, and how testing of AVs can occur (KPMG International, 2019). According to the report, among the four countries considered in this study, Japan has the strictest regulations on AV testing with a score of 0.333, while Korea and the US have somewhat fewer restrictions on AV testing, both receiving scores of 0.833; China's score was 0.5 in AV regulation (KPMG International, 2019). The scores of 2018 are largely the same, although a different scale was used (KPMG International, 2018).[2] Similar to AV regulation score, Korea and the US have higher scores than China and Japan in terms of institutional responsibility for AVs (KPMG International, 2019). According to the indicator of the AV-focused government agency by the KPMG International, South Korea's score is 0.857 and the US is 0.714. China's score of consumer AV acceptance is 0.643 and Japan is 0.571, which is the lowest among the four countries (KPMG International, 2019). Considering the fact that regulations are often affected and influenced by the voices of private businesses, the number of AV firms in a country might be a factor which is closely associated with the nature and level of regulations on AV test driving and safety. According

2. According to the 2018 scores on AV regulation, Japan, China, Korea, and the US were scored at 3, 4, 6, and 6, respectively (KPMG International, 2018).

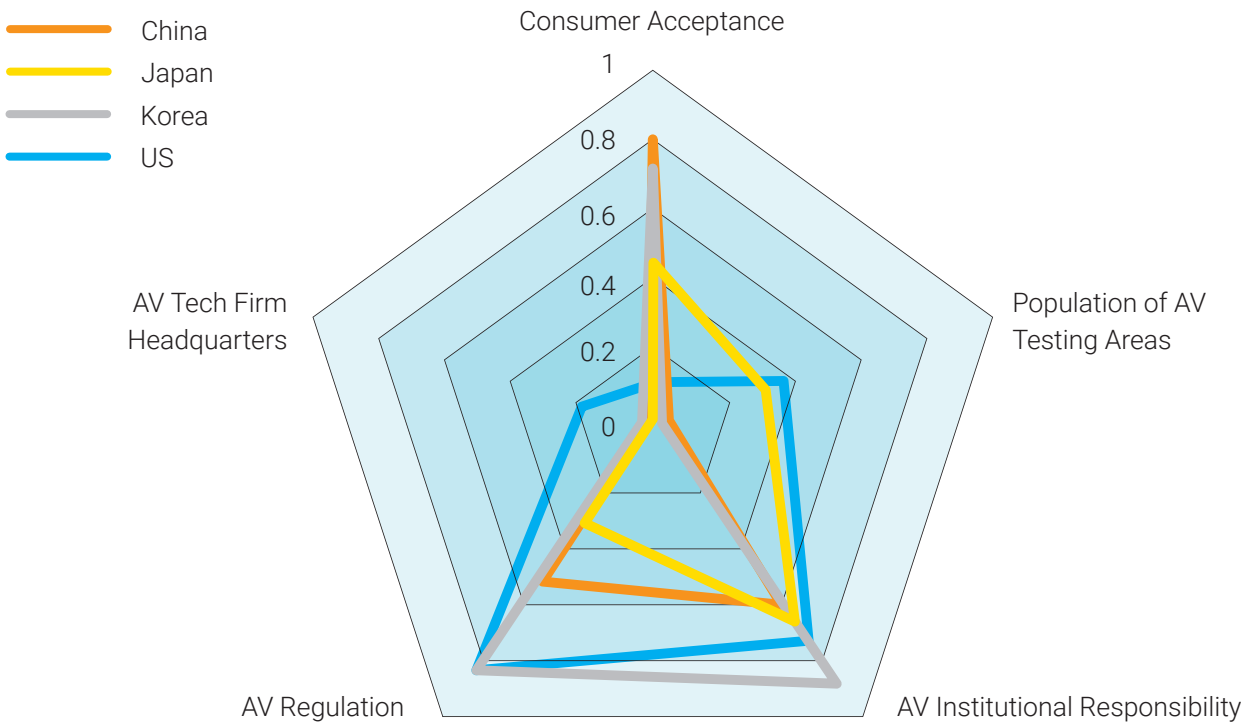


Figure 4: Regulatory and social dimensions for autonomous vehicles
(Source: Made by the author based on the data from KPMG International (2019))

to the index representing the number of AV technology firms' headquarters based on the KPMG International (2019), the US has the highest score of 0.176 followed by Korea (0.043). Japan is 0.029 while China (0.005) scored the lowest among the four countries (KPMG International, 2019).

In addition to AV regulations, social acceptance for AVs appears to be different among countries. As part of the consumer AI acceptance index, a consumer AV acceptance score—based on a branded research online consumer panel survey—shows that China scored the highest with 0.783 followed by South Korea's score of 0.725 (KPMG International, 2019). Japan and the US scored 0.442 and 0.103 respectively (KPMG International, 2019). In addition, the proportion of population living in AV testing areas (cities) vary because the numbers and areas of designated testing sites are different among countries. The US scored 0.355 for the highest percentage of people living in an AV testing area, followed by Japan with a score of 0.301; China and Korea scored 0.043 and 0.020 respectively (KPMG International, 2019).

The regulatory and social dimension scores of AV regulation for these four countries are compared in Figure 4.

The figure suggests that the US and Korea are very proactive and less restrictive about AVs, and have good institutional support for AV test driving. Japan is somewhat passive and cautious, with less institutional arrangement for AVs from the government. However, it is interesting to note that Korean consumers are the least receptive to AVs, and therefore test driving is limited to certain areas (smallest population living in test driving areas). Chinese and American consumers are highly receptive to AVs; particularly the US, as test driving is allowed in more areas than the three other countries, as indicated by the proportion of population in test areas. This suggests that the US is the least strict country when it comes to autonomous driving. It has not enacted specific legislation regarding AVs, but instead established guidelines based on SAE International standards that are used when establishing policies. In the US and Germany, AVs have already been put into operation on public roads.

Meanwhile, Japan has not yet passed legislation, but is preparing for Level-5 autonomous vehicle testing in advance of the Tokyo Olympics (Lee, 2018). Both China and Japan have declared their intentions to boost autonomous vehicle commercialization, and both have already passed related bills to allow test driving in limited areas. Additionally, Japan allows people to use cell phones while engaged in level 3 autonomous driving. Korea has also established a new law that addresses the commercialization of AVs, which is similar to the law for testing AVs. Despite the differences in regulating AVs, countries are similarly moving toward developing regulatory frameworks by introducing restrictions, limiting driving tests, and providing terms of technical standards. That said, there are still differences within these four countries' regulations in terms of technology-supported driving and safety measures.

Conclusions and Policy Recommendations

As governments consider disruptive technologies as a source of future economic competitiveness, many have been shifting their regulatory positions from a regulatory paternalistic position to a somewhat deregulatory position, as seen in sandbox initiatives. While the regulation of disruptive technologies has weakened worldwide due to many people believing that regulation can harm the development of novel technologies, the risks and uncertainties associated with disruptive technologies still remain valid and require some form of regulation. At the same time, ethical guidelines often precede specific and formal regulations due to the uncertain nature of those novel technologies. This study suggests there are two distinctive approaches—an ethical approach and legal/regulatory approach to new disruptive technologies. Examining the ethical guidelines of AI and the regulatory positions of AVs, this study suggests an ethical approach as an informal and unofficial guideline with key principles, which is often introduced before specific and formal regulations are adopted by governments. The ethical approach offers a broad range of key values to be considered for the design, development, deployment, and use of particular disruptive technologies. This study also suggests that regulatory decisions on disruptive technologies are often affected by uncertainties regarding the expected outcomes and social risk tolerance in relation to a specific technology. The regulatory positions of different countries might vary, primarily because of the expected roles of governments and market competition.

Regulatory schemes for novel technologies are not necessarily different from conventional technologies in a society, because regulatory politics are often similarly applied, regardless of the type of technology. However, we believe that disruptive technologies might create new regulatory dynamics in a country because of their novelties as well as their social risks and perceived uncertainty. Considering the implications of ethical and regulatory approaches, as well as their strengths and weaknesses, societies must manage disruptive technologies by carefully adopting and designing both approaches in order to address their uncertainties and perceived social risk. The following recommendations are proposed:

Recommendation 1: Moralizing disruptive technologies should precede, and should be fully discussed and shared among different stakeholder prior to regulating them. Before a society adopts and enacts specific regulatory frameworks for disruptive technologies, ethical guidelines (i.e., AI principles or AI ethical guidelines) must be jointly formulated based upon a thorough deliberation of particular disruptive technologies by different stakeholders representing industries, researchers, consumers, NGOs, international organizations, and policymakers.

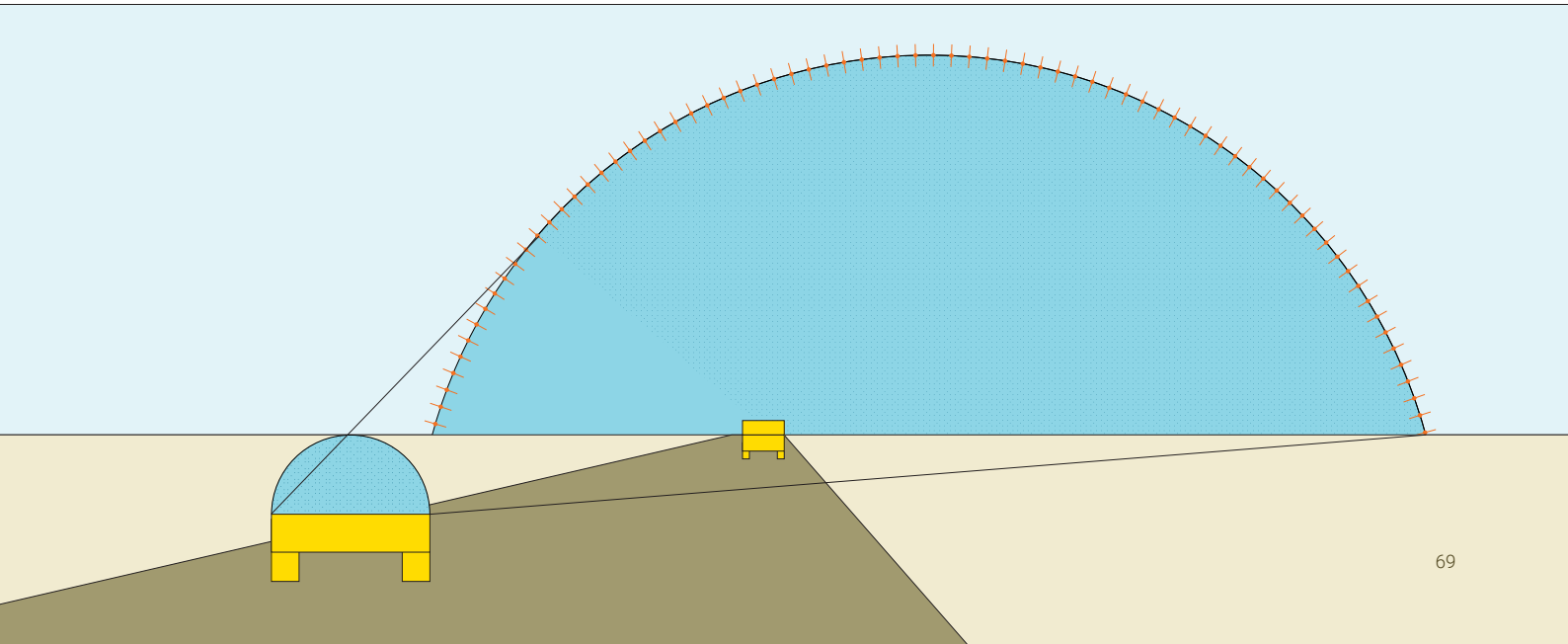
Recommendation 2: AI ethical guidelines should support sustainable and human-centric societies by minimizing the negative socio-economic and international consequences of disruptive technologies (i.e., inequality, unemployment, psychological problems, etc.), while maximizing their potential benefits for environmental sustainability, quality of life among others.

Recommendation 3: Once a general consensus is made on general ethical guidelines, they should be elaborated and specified in details targeting individual stakeholder groups representing different actors and sectors. Specific AI ethical guidelines should be developed and customized for AI designers, developers, adopters, users, etc. based on the AI lifecycle. In addition, industry and sector specific ethical guidelines should be developed and applied to each sector (care industry, manufacturing industry, service industry, etc.).

Recommendation 4: In regulating AI and other disruptive technologies, governments should align regulations with key values and goals embedded in various AI ethical guidelines (transparency, trustworthiness, lawfulness, fairness, security, accountability, robustness, etc.) and aim to minimize the potential social risks and negative consequences of AI by preventing and restricting possible data abuses or misuses, ensuring fair and transparent algorithms, in addition to establishing institutional and financial mechanisms through which the negative consequences of AI are systematically corrected.

Recommendation 5: Governments should ensure the quality of AI ecosystems by increasing government and non-government investment in R&D and human resources for AI by maintaining fair market competition among AI-related private companies, and by promoting AI utilities for social and economic benefits.

Recommendation 6: Governments should carefully design and introduce regulatory sandbox approaches to prevent unnecessarily strict and obstructive regulations that may impede AI industries but also facilitate developing AI and exploring AI-related innovative business models.



References

Aghion, P., Algan, Y., Cahuc, P., & Shleifer, A. (2010). Regulation and distrust. *The Quarterly Journal of Economics,* 125(3), 1015-1049.

Chang, I. (2019). US Legislative Trends and Implications for Gene Editing Technology. *Study on The American Constitution,* 30(1), 213-242.

Choe, Y. S., & Jeong, J. (1993). Charitable Contributions by Low- and Middle-Income Taxpayers: Further Evidence with a New Method. *National Tax Journal,* 46, 33–39.

Beck, T., Levine, R., & Demirgüç-Kunt, A. (2002). *Law and finance: why does legal origin matter?* The World Bank.

Becker, G. S., & Stigler, G. J. (1974). Law enforcement, malfeasance, and compensation of enforcers. *The Journal of Legal Studies,* 3(1), 1-18.

Black, J. (1998). Regulation as Facilitation: Negotiating the Genetic Revolution. *Mod. L. Rev.,* 61, 621.

Berkhout, J., & Lowery, D. (2010). The changing demography of the EU interest system since 1990. *European Union Politics,* 11(3), 447-461.

Borges, B. J. P., Arantes, O. M. N., Fernandes, A., Broach, J. R., Fernandes, P., & Bueno, M. (2018). Genetically Modified Labeling Policies: Moving Forward or Backward? *Frontiers in bioengineering and biotechnology,* 6, 181.

Brodsky, J. S. (2016). Autonomous vehicle regulation: How an uncertain legal landscape may hit the brakes on self-driving cars. *Berkeley Tech., LJ,* 31, 851.

Brunel, C., & Levinson, A. (2013). Measuring Environmental Regulatory Stringency. *OECD Trade and Environment Working Papers,* 2013(5), 0_1.

Castor, A. (11 May 2018). *How Japan Is Creating a Template for Cryptocurrency Regulation. Bitcoin magazine.* https://bitcoinmagazine.com/articles/how-japan-creating-template-cryptocurrency-regulation

Charo, R. A. (2016). The legal and regulatory context for human gene editing. *Issues in Science and Technology,* 32(3), 39.

Charron, N., Dahlström, C., & Lapuente, V. (2012). No law without a state. *Journal of Comparative Economics,* 40(2), 176-193.

Cheon, C. (2018). *Global ICO Regulation Trends and Implications.* Korea Capital Market Institute Issue report, 18-06.

Cohen, J. (19 March 2019). WHO panel proposes new global registry for all CRISPR human experiments, *Science.* https://www.sciencemag.org/news/2019/03/who-panel-proposes-new-global-registry-all-crispr-human-experiments

ComplyAdvantage. (2018). Cryptocurrency Regulations Around The World. https://complyadvantage.com/blog/cryptocurrency-regulations-around-world

Cook, K., Shortell, S. M., Conrad, D. A., & Morrisey, M. A. (1983). A theory of organizational response to regulation: the case of hospitals. *Academy of Management Review,* 8(2), 193-205.

Cyranoski, D. (2016). *CRISPR gene-editing tested in a person for the first time.* Nature. doi:10.1038/nature.2016.20988

Cyranoski, D., & Ledford, H. (2018). *Genome-edited baby claim provokes international outcry.* Nature. https://www.nature.com/articles/d41586-018-07545-0

Cyranoski, D. (2019). *China to tighten rules on gene editing in humans.* Nature. https://www.nature.com/articles/d41586-019-00773-y

Cyranoski, D. (2019). *China announces hefty fines for unauthorized collection of DNA.* Nature. https://www.nature.com/articles/d41586-019-01868-2

Cyranoski, D. (2019). *Japan approves first human-animal embryo experiments.* Nature. https://www.nature.com/articles/d41586-019-02275-3

Das, S. (2017). *China's Central Bank Completes Digital Currency Trial on a Blockchain,* CCN. https://www.ccn.com/chinas-central-bank-completes-digital-currency-trial-blockchain

De Bruycker, I., & Beyers, J. (2015). Balanced or biased? Interest groups and legislative lobbying in the European news media. *Political Communication,* 32(3), 453-474.

Deng, C. (2018, January 11). *China Quietly Orders Closing of Bitcoin Mining Operations,* The Wall Street Journal. https://www.wsj.com/articles/china-quietly-orders-closing-of-bitcoin-mining-operations-1515594021

Dickinson, S. (2018, July 17). *Self Driving Cars in China: The Absence of Non-Technical Barriers,* China Law Blog. https://www.chinalawblog.com/2018/07/self-driving-cars-in-china-the-absence-of-non-technical-barriers.html

Djankov, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2002). The regulation of entry. *The quarterly Journal of economics,* 117(1), 1-37.

Downey, H. K., & Slocum, J. W. (1975). Uncertainty: Measures, research, and sources of variation. *Academy of Management journal,* 18(3), 562-578.

European Central Bank. (2012). *Virtual Currency Schemes.* https://www.ecb.europa.eu/pub/pdf/other/virtualcurrencyschemes201210en.pdf

European Parliament. (2018). Report on three-dimensional printing, a challenge in the fields of intellectual property rights and civil liability (2017/2007(INI)).

Fleming, L. (2001). Recombinant Uncertainty in Technological Search. *Management Science,* 47(1), 117-132.

Fordham, B., & McKeown, T. (2003). Selection and Influence: Interest Groups and Congressional Voting on Trade Policy. *International Organization,* 57(3), 519-549.

Gai, P., Kemp, M., Sánchez Serrano, A., & Schnabel, I. (2019). *Regulatory complexity and the quest for robust regulation* (No. 8). European Systemic Risk Board.

Glaeser, E. L., & Shleifer, A. (2002). Legal origins. *The Quarterly Journal of Economics,* 117(4), 1193-1229.

Globus, R., & Qimron, U. (2018). A technological and regulatory outlook on CRISPR crop editing. *Journal of cellular biochemistry,* 119(2), 1291-1298.

Go, J. (December 2). *[Genetic Editing Baby Controversy] a rekindled debate on human embryo research.* Dong-A Science. http://dongascience.donga.com/news.php?idx=25463

Hacker, P., & Thomale, C. (2018). Crypto-Securities Regulation: ICOs, Token Sales and Cryptocurrencies under EU Financial Law. *European Company and Financial Law Review,* 15(4), 645-696.

Hail, L., & Leuz, C. (2006). International differences in the cost of equity capital: Do legal institutions and securities regulation matter?. *Journal of accounting research,* 44(3), 485-531.

Hofstede, G. (2015). *The 6-D model of national culture.* https://geerthofstede.com/culture-geert-hofstede-gert-jan-hofstede/6d-model-of-national-culture/

Houben, R., & Snyers, A. (2018). Cryptocurrencies and blockchain: legal context and implications for financial crime, money laundering and tax evasion. Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament.

Hunt, G., & Mehta, M. (Eds.) (2013). *Nanotechnology: "Risk, Ethics and Law".* Routledge.

Hwang, Y. (2018, August 29). *Deregulation of human embryo research and other...There's going to be a controversy over bioethics.* http://www.hani.co.kr/arti/society/health/859834.html

Jeung, T. (2018, October 10). *Japan Is Drafting a Rulebook for Ethically Editing the Genes of Human Embryos: Which country will be first to create a CRISPR baby?* https://www.inverse.com/article/49725-governments-regulate-human-embryo-gene-editing

Kahan, D. M., Jenkins-Smith, H., & Braman, D (2011). Cultural cognition of scientific consensus. *Journal of risk research,* 14(2), 147-174.

Kerwer, D. (2005). Rules that many use: Standards and global regulation. *Governance, 18*(4), 611-632.

Kim, B. (2015). A Study on Uber Taxi and the Fit of It. *Chonbuk Law Review, 46, 99*~134.

Kisiel, D. (2018). Legal concept of internet currencies. *Financial Law Review,* 11(3), 81-91.

Kolacz, M. K., Quintavalla A., & Yalnazov. O. (2019). Who Should Regulate Disruptive Technology? European *Journal of Risk Regulation,* 10(1), 4-22.

KPMG International. (2018). *Autonomous Vehicles Readiness Index - Assessing countries' openness and preparedness for autonomous vehicles.* https://assets.kpmg/content/dam/kpmg/xx/pdf/2018/01/avri.pdf

KPMG International. (2019). *2019 Autonomous Vehicles Readiness Index – Assessing countries' preparedness for autonomous vehicles.* https://assets.kpmg/content/dam/kpmg/xx/pdf/2019/02/2019-autonomous-vehicles-readiness-index.pdf

Kun, L., & Xiaodong, W. (2019, March 1). *Rules to be revised on organ donations.* China Daily. http://www.chinadaily.com.cn/a/201903/01/WS5c78936aa3106c65c34ec237.html

Lander, E. S., Baylis, F., Zhang, F., Charpentier, E., Berg, P., Bourgain, C., Friedrich, B., Joung, J. K., Li, J., Liu, D., Naldini, L., Nie, J., Qiu, R., Schoene-Seifert, B., Shao, F., Terry, S., Wei, W., & Winnacker, E. (2019, March 19). *Adopt a moratorium on heritable genome editing,* Nature. https://www.sciencemag.org/news/2019/03/who-panel-proposes-new-global-registry-all-crispr-human-experiments

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Robert, V. (1997). Legal determinants of external finance. *Journal of Finance* 52, 1131–1150.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Robert, V. (1998). Law and finance. *Journal of Political Economy* 106, 1113–1155.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Robert, V. (1999). The quality of government. *Journal of Law, Economics, and Organization* 15, 222–279.

La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2008). The economic consequences of legal origins. *Journal of economic literature* 46(2), 285-332.

Lee, S. (2018). *Issues on Regulatory Reform for Industrial Revitalization of Self-driving Cars.* ICT Spot issue, IITP, S18-06.

Lee, S. & Kim, H. (2018). International Regulatory Trends on Genome Editing Research Using Human Embryo and Its Implication. *Korean Journal of Medicine and Law* 26(2), 71-96.

Marchant, G., Meyer, A., & Scanlon, M. (2010). Integrating social and ethical concerns into regulatory decision-making for emerging technologies. *Minn. JL Sci. & Tech.,* 11, 345.

Marris, C., Langford, I., Saunderson, T., & O'Riordan, T. (1997). Exploring the "psychometric paradigm": comparisons between aggregate and individual analyses. *Risk analysis, 17*(3), 303-312.

Marshall, A. (2018). *New York City Goes After Uber and Lyft.* Wired. https://www.wired.com/story/new-york-city-cap-uber-lyft

Martin-Laffon, J., Kuntz, M., & Ricroch, A. E. (2019). Worldwide CRISPR patent landscape shows strong geographical biases. *Nature biotechnology.* 37(6), 613-620.

Matsuda, D., Mears, E., & Shimada, Y. (2019). *Legalization of Self-Driving Vehicles in Japan: Progress Made, but Obstacles Remain.* DLA Piper. https://www.dlapiper.com/en/global/insights/publications/2019/06/legalization-of-self-driving-vehicles-in-japan

Milliken, F. J. (1987). Three types of perceived uncertainty about the environment: State, effect, and response uncertainty. *Academy of Management review,* 12(1), 133-143.

Ministry of Science and ICT. (2017). *Korea Provides Gene Scissors, U.S. Corrects Human Embryo Gene Mutation.* Press Releases.

Ministry of Science and ICT and KISTEP. (2018). Comparative analysis of domestic and foreign legislation on autonomous vehicles and policy alternatives, *In Science, ICT Policy and Technology Trends, 128*.

Molteni, M. (2019, July 30). The World Health Organization Says No More Gene-Edited Babies, *Wired*. https://www.wired.com/story/the-world-health-organization-says-no-more-gene-edited-babies

Nature. (2019, March 13) *Hybrid embryos, ketamine drug and dark photons.* https://www.nature.com/articles/d41586-019-00790-x

Normile, D. (2019). *China tightens its regulation of some human gene editing, labeling it 'high-risk'.* Science. https://www.sciencemag.org/news/2019/02/china-tightens-its-regulation-some-human-gene-editing-labeling-it-high-risk

Normile, D. (2019) *Gene-edited foods are safe, Japanese panel concludes,* Science. https://www.sciencemag.org/news/2019/03/gene-edited-foods-are-safe-japanese-panel-concludes

OECD. (2018). Blockchain Technology and Corporate Governance.

Ogus, A. (2005). Regulatory paternalism: when is it justified?. *Corporate governance in context: Corporations, states, and markets in Europe, Japan, and the US*, 303-320.

Ormond, K. E., Mortlock, D. P., Scholes, D. T., Bombard, Y., Brody, L. C., Faucett, W. A., Garrison, N. A., Hercher, L., Isasi, R., Middleton, A., Musunuru, K., Shriner, D., Virani, A., & Young, C. E. (2017). Human germline genome editing. *The American Journal of Human Genetics. 101*(2), 167-176.

Oshiro, Y., & Ohkohchi, N. (2017). Three-dimensional liver surgery simulation: computer-assisted surgical planning with three-dimensional simulation software and three-dimensional printing. *Tissue Engineering Part A, 23*(11-12), 474-480.

Park, T. (2019). *Does Uber want to tap the Korean market again? Foreign Taxi Call Service Initiated.* Hankyoreh. http://www.hani.co.kr/arti/economy/it/879525.html#csidx7f70c05e63c5236a29bda7b854ff47f

Pinto, C. (2012). How autonomous vehicle policy in California and Nevada addresses technological and non-technological liabilities. Intersect: *The Stanford Journal of Science, Technology, and Society,* 5.

Pollock, D. (2018, March 21). *G20 and Cryptocurrencies: Baby Steps Towards Regulatory Recommendations.* https://cointelegraph.com/news/g20-and-cryptocurrencies-baby-steps-towards-regulatory-recommendations

Herskind, N., Lim, C.K., & Hoist, S. (2019). How China will shape the future of autonomous vehicles. QVARTZ. https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-standard-for-self-driving-vehicles

Roca, J. B., Vaishnav, P., Morgan, M. G., Mendonça, J., & Fuchs, E. (2017). When risks cannot be seen: Regulating uncertainty in emerging technologies. *Research Policy, 46*(7), 1215-1233.

Sabel, C., Herrigel, G., & Kristensen, P. H. (2018). Regulation under uncertainty: The coevolution of industry and regulation. *Regulation & Governance, 12*(3), 371-394.

Schwinger, A. (2018, March 14). *Federal court holds that CFTC can regulate virtual currencies as commodities,* Norton Rose Fulbright website. https://www.nortonrosefulbright.com/en/knowledge/publications/6c7bcc30/federal-court-holds-that-cftc-can-regulate-virtual-currencies-as-commodities

Shim, M. (2019, June 13). Legal Issues Related to Genetics Patent. *Korea Institute of Intellectual Property.* https://www.kiip.re.kr/board/report/view.do?bd_gb=data&bd_cd=4&bd_item=0&po_item_gb=5&po_item_cd=&po_no=12504

Shukla-Jones, A., Friedrichs, S., & Winickoff, D. E. (2018). Gene editing in an international context: Scientific, economic and social issues across sectors. *OECD Science, Technology and Industry Working Papers,* 2018(4), 0_1-51.

Siegrist, M. (2010). Psychometric paradigm. *Encyclopedia of science and technology communication, Volume 2,* pp. 600-601. SAGE Publications.

Slovic, P. (1987). Perception of risk. *Science, 236*(4799), 280-285.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1982). Why study risk perception?. *Risk analysis, 2*(2), 83-93.

Starr, C. (1969). Social benefit versus technological risk. *Science,* 1232-1238.

Taeihagh, A., & Lim, H. S. M. (2019). Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transport Reviews, 39*(1), 103-128.

The Francis Crick Institute. (2019). Kathy Niakan: Human embryo genome editing licence. https://www.crick.ac.uk/research/labs/kathy-niakan/human-embryo-genome-editing-licence

The Law Library of Congress. (2014). Restrictions on Genetically Modified Organisms. Global Legal Research Center.

The Library of Congress. (2018, August 16). *Regulation of Cryptocurrency Around the World.* https://www.loc.gov/law/help/cryptocurrency/world-survey.php

The Library of Congress. (2018, August 16). *Regulation of Cryptocurrency: China.* https://www.loc.gov/law/help/cryptocurrency/china.php

The United Nations Economic and Social Council. (2017). Consolidated Resolution on the Construction of Vehicles (R E.3).

The US Congressional Research Service. (2019). H.R.3388 - SELF DRIVE Act. https://www.congress.gov/bill/115th-congress/house-bill/3388

Tomlinson, T. (2018). A crispr future for gene-editing regulation: a proposal for an updated biotechnology regulatory system in an era of human genomic editing. *Fordham L. Rev., 87,* 437.

Tzur, A. (2017). Uber Über regulation? Regulatory change following the emergence of new technologies in the taxi market. *Regulation & Governance.* https://doi.org/10.1111/rego.12170

UK Department for Transport. (2015). *The Pathway to Driverless Cars: A detailed review of regulations for automated vehicle technologies.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/401565/pathway-driverless-cars-main.pdf

Van Rijssen, W. J., & Morris, E. J. (2018). Safety and Risk Assessment of Food From Genetically Engineered Crops and Animals: The Challenges. *In Genetically Engineered Foods,* pp. 335-368. Academic Press.

Vienna Convention on Road Traffic. (2009). 1968 Vienna Convention on Road Traffic: Consolidated Resolution on Road Traffic. Revised on 14 August, 2009.

Wilson, J. (1980). *Politics of Regulation.* New York: Basic Books.

World Bank. (2016). *World Development Report 2016: Digital Dividends.* Washington, DC: World Bank. DOI:10.1596/978-1-4648-0671-1

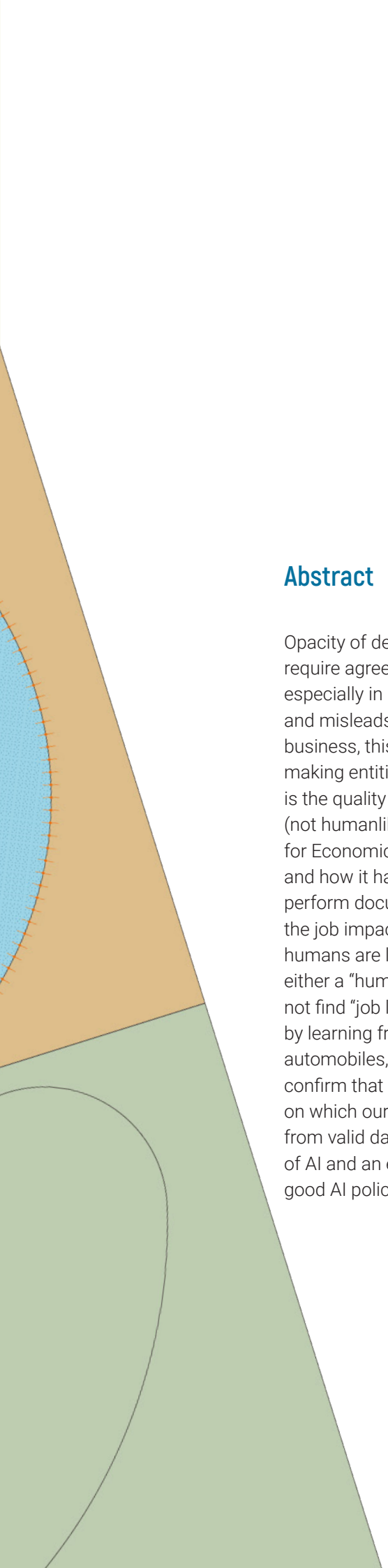
World Economic Forum. (2017). *Global Competitiveness Index 2017-2018.* whttp://reports.weforum.org/global-competitiveness-index-2017-2018/?doing_wp_cron=1565516422.9761869907379150390625

Xinying, Z. (2019, March 1). Ministry to speed development of self-driving vehicles. http://www.chinadaily.com.cn/a/201903/01/WS5c78992ca3106c65c34ec27d.html

Definition and Recognition of AI and its Influence on the Policy: Critical Review, Document Analysis and Learning from History

Kyoung Jun Lee
School of Management,
Kyung Hee University

Yujeong Hwangbo
Dept. of Social Network Science,
Kyung Hee University



Abstract

Opacity of definitions hinders policy consensus; and while legal and policy measures require agreed definitions, to what artificial intelligence (AI) refers has not been made clear, especially in policy discussions. Incorrect or unscientific recognition of AI is still pervasive and misleads policymakers. Based on a critical review of AI definitions in research and business, this paper suggests a scientific definition of AI. AI is a discipline devoted to making entities (i.e., agents and principals) and infrastructures intelligent. That intelligence is the quality which enables entities and infrastructures to function (not think) appropriately (not humanlike) as an agent, principal, or infrastructure. We report that the Organization for Economic Co-operation and Development (OECD) changed its definition of AI in 2017 and how it has since improved from humanlike to rational and from thinking to action. We perform document analysis of numerous AI-related policy materials, especially dealing with the job impacts of AI, and find that many documents which view AI as a system that mimics humans are likely to overemphasize the job loss incurred by AI. Most job loss reports have either a "humanlike" definition, a "human-comparable" definition, or "no definition". We do not find "job loss" reports that rationally define AI, except for Russell (2019). Furthermore, by learning from history, we show that automation technology such as photography, automobiles, ATMs, and Internet intermediation did not reduce human jobs. Instead, we confirm that automation technologies, as well as AI, creates numerous jobs and industries, on which our future AI policies should focus. Similar to how machine learning systems learn from valid data, AI policy makers should learn from history to gain a scientific understanding of AI and an exact understanding of the effects of automation technologies. Ultimately, good AI policy comes from a good understanding of AI.

1. Scientific understanding of AI

How one recognizes something influences their attitude when dealing with it. With AI being a very new concept compared with traditional subjects such as physics, economics, and sociology, there have been numerous misunderstandings; and while these have been overcome by the AI communities themselves, there is still incorrect and unscientific recognition of AI. Definitional ambiguity hampers the possibility of conversation; and although legal and regulatory intervention requires agreed-upon definitions, consensus surrounding the definition of AI has been elusive, especially in policy conversations (Krafft et al., 2020). In the following sections, we attempt to correct this misconception, thereby redefining AI.

1.1. AI is a discipline not an entity

Although AI is a discipline, some view it as a physical thing, in other words, a machine or entity. For example, the physicist Stephen Hawking told the BBC that "[the] development of full artificial intelligence could spell the end of the human race" (Cellan-Jones, 2014). This statement highlights Stephen Hawking's misunderstanding of AI, which, in turn, can mislead mass media and people. Just as he regarded AI as an entity and not a discipline, the non-AI community and non-professional community sometimes show their misunderstanding of AI by defining it as "machines performing humanlike cognitive functions" (OECD, 2017) or "intellectual machines and systems… that could automatically sense people's situations or expectations, and offer necessary information before it is required" (Ema et al., 2016). That said, mainstream AI research communities have known AI is an activity devoted to making machines intelligent (Nilsson, 2010),[1] is the science of making machines smart (Hassabis, 2015), and is a discipline. The most frequently used textbook in AI, "Artificial Intelligence: A Modern Approach" (Russell & Norvig, 1995), says that AI is "one of the newest fields in science and engineering". Textbooks older than this also explain that AI is the study of how to make computers do things which, at the moment, people do better (Rich, Knight & Nair, 1985); the study of mental faculties through the use of computational models (Charniak & McDermott, 1985); and the study of the computations that make it possible to perceive, reason, and act (Winston, 1992).

1.2. AI is not about humans, it should be based on rationalism

The definition of AI should not include the word "human". Physics is not about humans, chemistry is not about humans; both are natural science. History is about humans, sociology is about humans; these are humanities and social science, respectively. AI is the science of the artificial (Simon, 1969), it is not a science about humans. A natural science similar to AI is brain science, which is concerned with how human and animal brains work. AI, however, is not about how the human brain works, since even animals can be intelligent. As such, AI should not deal solely with human intelligence. Including the word "human" in the definition of AI confines the scope of the discipline and misleads academic and practitioner communities. AI is simply an activity that makes certain entities intelligent. It is not about making machines humanlike in intelligence; Nor is it about making machines more intelligent than humans, despite numerous non-professionals explaining AI as trying to making something more intelligent than a human (Bostrom, 2014; Cellan-Jones, 2014; Clifford, 2017; Manyika et al., 2017; Niyazov, 2019; John, 2019; Adel, 2019).

We found evidence that even AI researchers such as Rich and Knight (1991), incorrectly define AI as about making humanlike intelligence or human-comparable intelligence. Defining AI as human-related is a very common mistake in the non-AI and non-professional communities, such as with the aforementioned OECD (2017) and Ema et al. (2016). Merriam-Webster also shows an incorrect understanding of AI by defining it as "the capability of a machine to imitate intelligent human behavior".

1. AI is the activity devoted to making machines intelligent, and intelligence is that quality which enables an entity to function appropriately and with foresight in its environment (Nilsson, 2010).

This misconception of AI as "imitating humans" comes from the misunderstanding of Alan Turing's imitation game, the so-called Turing Test. Alan Turing, the father of computer science, suggested using the test as an operational definition of a "machine that can think". If a machine can pass test, then he suggested we can say the machine can think. However, different from his original intention, early AI scholars considered passing the imitation game as the goal of AI. Many AI researchers began to think that the goal of AI was to make a machine that is indiscernible from a human.

However, this outdated belief began to change after Hayes and Ford's speech at the International Joint Conference on Artificial Intelligence (IJCAI) in Montreal, Canada in 1995. Hayes and Ford asserted that the Turing Test has harmed AI development. They explained how, to be able to fly, it is not necessary for us to construct a bird-like flying machine or a machine that is indiscernible from a bird. Just as aeronautics is based on Bernoulli equation (Bernoulli, 1738) and not ornithology, AI does not have to be based on brain science. Russell and Norvig (1995) also referred to Hayes and Ford (1995) in their famous book, "Artificial Intelligence: A Modern Approach".

They propose two dimensions on the view of AI: humanlike or rational and thinking or acting. In choosing rationality over humanlike and acting over thinking, theirs is the first really "modern" approach to AI in comparison with traditional textbooks. As will be discussed in the following sections, the AI community has evolved by overcoming the Turing Test and not emphasizing AI cognition. Gershman et al. (2015), also proposes computational rationality as a potential unifying paradigm for intelligence in brains, minds, and machines.

1.3. AI is not only about cognition

Certain explanations of AI emphasize the cognitive aspect (Drum, 2017; Miller-Merrell, 2019; Frey & Osborne, 2017; Manyika et al., 2017). For example, we see plenty of examples of using the word "cognitive" or "cognition" when defining AI, such as Eysenck et al.'s (1990) definition of AI as the "attempt to develop complex computer programs that will be capable of performing difficult cognitive tasks". OECD (2017) also defines AI as "machines performing humanlike cognitive functions". Sometimes this emphasis on cognition stems from attempting to differentiate AI from robotics. However, robotics also deals with cognition. Bostrom's (2014) definition of superintelligence, as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest", also mistakenly emphasizes cognition. This emphasis on cognition is not only wrong but is also misleading, in that it implies the AI system can think. As Turing tried to explain, we cannot determine when a thing thinks or not. Instead, he simply suggested a proxy test for the decision. Emphasis on cognition runs the risk of neglecting the action aspect of AI, which is a more important aspect of intelligence.

The traditional explanation of intelligent systems says an intelligent system has three processes: perception, cognitive, and motor. The perceptual system consists of sensors and associated memories. The cognitive system receives information from the stores in its working memory and uses previously stored information in long-term memory to make decisions about how to respond. The motor system carries out the response (Card et al., 1983). However, this traditional sandwich (perception-cognitive-motor) model has been criticized, for example, by Hurley (1998), and has now evolved into "enactivism". This is defined as the manner in which a subject of perception creatively matches its actions to the requirements of its situation (Protevi, 2006). Similar to the relatively new enactivism, traditional behaviorism also excludes or doubts the central role of cognition in intelligent systems. As such, the view regarding cognition as the center of intelligence is now being challenged, such as in Auer-Welsbach (2019).[2] As explained above, there still exists a disagreement over the central role of cognition; hence, the definition of AI should not only include the word "cognitive".

2. The fundamental composition of the most advanced intelligent system, the Homo Sapiens system, is not comprised of independent information processing units which interface with each other via representations. Instead, the system is comprised of independent and parallel producers of activity which all interface directly with the world through perception and action, rather than interface with each other exclusively. From this perspective, the notions of central and peripheral systems evaporate, as everything is both central and peripheral.

1.4. AI should be extended to not just agents

To date, AI applications have been confined to making agents intelligent from the principal-agent perspective. Meaning that the agents in AI disciplines only refer to machines, software, and robots that are owned and controlled by human principals. For example, Nilsson's (2010) definition of AI, as explained earlier, satisfies all three conditions: (1) it is referred to as a discipline, (2) it is not humanlike, and (3) there is not only an emphasis on cognition. This definition is the most accepted and up-to-date, and is therefore referred to by the comprehensive review and prospect report, "Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence" (Stone et al., 2016).

However, Nilsson's (2010) definition has one limitation which confines the intelligent entity to only a machine. This similar to Hassabis' (2015) definition in its limitation. In this paper, we extend Nilsson's definition since AI now plays a wide role in society. It is important to remember that AI is a discipline which makes entities and infrastructures intelligent, whereby the entities not only refer to agents such as machines, but also include principals such as humans, organizations, businesses, and nations. Infrastructures include computing elements, which can be imbedded into the natural world such as forest, lakes, and seas, as well as artificial infrastructures such as roads, cities, buildings, and homes. The extension to infrastructures from entities in the definition of AI removes the humanlike feature, since it is nonsense to imagine humanlike roads or buildings. We assume that the agent orientation in defining AI could lead to humanlike orientation, which we can avoid by extending the scope of AI in its definition.

At the time, Russell and Norvig's (1995) approach which defined AI as making rational agents was the most pioneering and scientific at the time, hence why their book has been the most widely used at top AI schools around the world for more than 20 years since its publication. That said, it is necessary to extend Nilsson's (2010) and Russel and Norvig's (1995) definition and approach from making agents rational to making entities and infrastructures rational. Until now, AI research has concentrated only on optimizing the behavior of agents under a given condition. However, sensors and their networking technologies, such as Internet of things (IoT) technology, and automatic recognition technologies, such as convolution neural networks (CNNs), enable making infrastructures intelligent. Nowadays, AI needs to deal with the intelligence of not only single entities but also of infrastructures. This enlarged perspective encompasses the efforts for and contributions to human intelligence augmentation. In other words, augmented intelligence and intelligence amplification (Licklider, 1960; Engelbart 1962).[3] Jordan (2018) suggests a new term called intelligent infrastructure (II). Our new AI definition encompasses intelligence amplification (IA) and II, as well as traditional agent-oriented AI.

3. By "augmenting human intellect" we mean increasing the capability of someone to approach a complex problem, to gain comprehension to suit their particular needs, and to derive solutions to the problem. In this respect, increased capability is taken to mean a mixture of the following: more rapid comprehension, better comprehension, the possibility of gaining a useful degree of comprehension in a situation that was previously too complex, speedier solutions, better solutions, and the possibility of finding solutions to problems that before seemed insoluble (Engelbart 1962).

2. Scientific definition of AI

The simplest definition of AI is *a discipline that makes entities and infrastructures intelligent.* If we refine that definition, AI is *a discipline devoted to making entities and infrastructures intelligent, with intelligence being that quality which enables entities and infrastructures to function appropriately.*

2.1. The meaning of functioning appropriately

"To function appropriately" is derived from Nilsson's (2010) definition. It also means "acting rationally", as per Russell and Norvig's (1995) two-by-two matrix. This paper will dispense with a detailed explanation of each quadrant of the matrix because we have already criticized humanlike and cognition emphases when defining AI in an earlier section. Appropriate functioning is necessary for an entity to survive and prosper. Intelligence is evolved for the process of survival and, simultaneously, becomes the result of the prospering of entities. Thus, appropriate functioning is developed through evolution for natural entities and through optimization by a designer for artificial agents and infrastructures. We found that Nilsson's (2010) "functioning appropriately" comes from Albus's (1991) definition of intelligence as "the ability of a system to act appropriately in an uncertain environment, where appropriate action is that which increases the probability of success, and success is the achievement of behavioral subgoals that support the system's ultimate goal". According to Albus (1991), the criteria of success and the system's ultimate goal are defined externally to the intelligent system. For an intelligent machine system, the goals and success criteria are typically defined by designers, programmers, and operators. For intelligent biological creatures, the ultimate goal is gene propagation, with success criteria being defined by the processes of natural selection.

Albus (1991) deals with the intelligence of both artificial intelligent systems and intelligent nature. His notion of intelligence corresponds with Anastasi's (1992) explanation that intelligence is the combination of abilities required for survival and advancement within a particular culture, and with Roth and Dicke's (2005) definition of intelligence.[4] In the definition of AI, "appropriate action" is also found in Kubacki (2009).[5] The recognition of intelligence as an instrument for survival and prosperity has not been popular in AI communities, though the idea was prevalent in evolutionary biology and psychology. However, we can find attempts by AI communities who view AI for the survival and prosperity of entities. Weng (2002) regards the performance of an intelligent entity as keeping the norm defined by social groups,[6] which can be called "institutional intelligence". This approach can be called an institutional approach to AI. Since institutional economics is a relatively new discipline in economics, the institutional approach to AI is a novel area to investigate.

4. Intelligence may be defined and measured by the speed and success of how animals, including humans, solve problems to survive in their natural and social environments (Roth & Dicke 2005).

5. Artificial, "embodied" intelligence refers to the capability of an embodied "agent" to select an appropriate action based on the current, perceived situation (Kubacki 2009).

6. Different age groups of developmental robots have corresponding norms. If a developmental robot has reached the norm of a human group of age k, we can say that it has reached the equivalent human mental age k (Weng 2002).

2.2. Optimization as the science of functioning appropriately

AI traditionally focuses on optimizing the behaviors of an agent under the conditions and goals given by its principal. Intelligent agents fundamentally seek to form beliefs and plan actions in support of maximizing expected utility (Gershman et al., 2015). Our new definition of AI emphasizes approaches to enabling the appropriate actions of agents, principals, and infrastructures. Hence, AI can be divided into: (1) making agents rational – finding a method of optimizing the behavior of an agent with the goals given by the principal (i.e., the owner of the agent), and (2) making entities and infrastructures function appropriately – finding the optimization method in which the entities survive and prosper while interacting with other entities and the infrastructures in their environment by making the rational entities and infrastructures learn, adapt, and improve the institutions of the world or society. In either case, it is important to recognize that optimization is the main problem when creating such AIs.

Optimizing a behavior of an agent under a principal has been covered by many studies on optimization systems. It is important to note that there is an intractable problem in which the optimal solution cannot be obtained, no matter how good the computer's performance. Stuart Russell's recent book, "Human Compatible: Artificial Intelligence and Problem Control", also confirms that the existence of intractable problems gives us reason to think that computers cannot be as intelligent as humans. There is also no reason to assume that humans can solve intractable problems either (Russell, 2019).

Gershman et al. (2015) emphasizes that ideal maximizing expected utility (MEU) calculations may be intractable for real-world problems. That is, finding optimal solutions can be intractable, even though optimization can be effectively approximated by rational algorithms which maximize a more generally expected utility incorporating the costs of computation. Thus, even though AI methodology improves, there are still certain optimization problems which cannot be solved under limited time and resources.

Judd (1990) proved learning in neural networks is NP-complete, and thus demonstrated that it has no efficient general solution. Goodfellow et al. (2015) also confirmed neural networks cannot avoid local minima.[7] Google-developed quantum computers solved a problem in three minutes, while the IBM Summit, the most powerful supercomputer in existence, requires a calculation time of 10,000 years (Arute et al., 2019). If quantum computing, which is 1 billion times faster than current supercomputing, is well developed and widely used for optimizing problems, it may become possible to solve problems considered intractable. If so, the range of problems that mankind could solve would be drastically expanded. Russell (2019) confirms that quantum computation helps slightly in solving intractable problems, but not enough to change the basic conclusion that there is no reason to suppose that humans can solve intractable problems.

On the other hand, if such developments are not realized, AI will still be forced to incompletely solve numerous problems and create a system for making occasional mistakes. Such incomplete systems should be used safely under human control. Although the performance of deep learning algorithms has improved, mistakes (i.e., local optima) have not gone away, which is the main problem of deep learning. Since deep learning is simply a neural network, it inherits the characteristics of a neural network, such as inexplainability and error inevitability. Research into increasing explanatory possibilities continues, and automatic recognition by deep learning is evolving, however, there is still a danger due to recognition error. Therefore, it is only suitable for use in areas where mistakes are not fatal and statistically good results are achieved. Current AI methodology is essentially a system that is able to make mistakes (Szegedy et al., 2014; Nguyen et al., 2016). Thus, Facebook researchers (Bordes et al., 2015) emphasize research and development through artificial tasks, just as an artificial task, such as XOR (exclusive OR) (Minsky & Papert, 1969), led to the birth of a multi-layer perceptron (Rumelhart et al., 1986).

7. Do neural networks enter and escape a series of local minima? Do they move at varying speed as they approach and then pass a variety of saddle points? [...] we present evidence strongly suggesting that the answer to all of these questions is no (Goodfellow et al., 2015).

2.3. An AI approach defined as an optimization problem

An AI algorithm is an algorithm which can find an optimal path to a preferred goal node, provided that the heuristic function satisfies certain conditions (Hart et al., 1968). Genetic or evolutionary algorithms are a type of optimization algorithm, meaning they are used to find the maximum or minimum of a function (Carr, 2014) called a "fitness function" – often a black-box in real-world applications. Automated theorem proving also finds proofs via application of optimization methods (Yang et al., 2016).

Most machine learning problems, once formulated, can be solved as optimization problems, with the essence of most machine learning algorithms being to build an optimization model and learn the parameters in the objective function from the given data (Sun et al., 2019). Sun et al. (2019) formulates supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning as optimization problems. For example, with supervised learning, the goal is to find an optimal mapping function to minimize the loss function of the training samples. Deep learning, if without nonlinearity in the hidden layer, would reduce to a generalized linear model. As such, minimizing the nonlinear and nonconvex loss functions is difficult, and at best we seek good local optima (Efron and Hastie, 2016). Reinforcement learning is a branch of machine learning, whereby an agent interacts with the environment through a trial and error mechanism, and learns an optimal policy by maximizing cumulative rewards (Sutton and Barto, 1998). Dialogue can also be considered as optimal decision making (Gao et al., 2018). The goal of dialogue learning for realizing conversational AI is to find optimal policies to maximize expected rewards in a reinforcement learning framework.

2.4. Successful AI applications in the pursuit of optimization

Successful AI applications and developments include the optimization perspective in their explanations. Libratus (Brown and Sandholm, 2017), the first AI system to defeat top humans in heads-up no-limit Texas hold 'em poker, formulates itself by finding the optimal strategy for solving subgames. While Libratus may not be able to arrive at an equilibrium by independently analyzing subtrees, it may be possible to improve the strategies in those subtrees when the original base strategy is suboptimal, as is typically the case when abstraction is applied. DeepMind's AlphaGo is also based on the optimization perspective, claiming that all games of perfect information have an optimal value function, which determines the outcome of the game from every board position or state, under perfect play by all players (David et al, 1986).

On the other hand, IBM's Watson is not based on the optimization perspective. Watson is a knowledge-based decision support tool that suffers from the requirement to manually craft and encode formal logical models of the target domain. This should be evolved into an interactive decision support capability that strikes a balance between a search system and a formal knowledge-based system (Ferrucci, 2012). IBM's Watson has not been successfully deployed, experiencing only failures, particularly in the medical field (Brown, 2017; Herper, 2017; Bloomberg, 2017; Strickland, 2019).

Softbank's Pepper is not formulated as an optimized machine either. As a result, Pepper is rather limited in how it can help customers and its answers do not seem that helpful (Mogg, 2018). Pepper's failure was predicted (Lee, 2014) and widely reported on (Alpeyev & Amano, 2016; Bivens, 2016; Boxall, 2017; Nichols, 2018). Hanson Robotics' robot, Sophia, is a typical example of AI being based on the incorrect humanlike perspective, rather than the rational optimization perspective. As such, it only makes jokes and cannot have meaningful conversations (Campanella, 2016). Similarly, Honda's ASIMO business operation has also been stopped (Ulanoff, 2018). Humanoids such as Pepper, Sophia, and ASIMO all failed because they were based on a humanlike paradigm and not on an optimization framework.

3. OECD's redefinition of AI

Of the aforementioned perspectives, the OECD (2017) definition of AI is the most inaccurate, as it includes all three misconceptions. OECD (2017) defined AI as "Machines performing humanlike cognitive functions", thereby mistaking AI as an entity and not a discipline and incorrectly believing that AI should be humanlike. When defining AI, OECD (2017) also only emphasized cognition – a common misconception. This critical mistake in the definition of AI by the world-leading policy organization could have resulted in misguided policy decisions. In 2017, OECD was advised by one of this paper's authors to revise its definition. Interestingly, OECD (2018) changed it to: "Equipping systems with cognitive functions that allow them to function appropriately and with foresight in their environment". From this, it is apparent that OECD (2018) adopted Nilsson's (2010) definition. In the new definition, OECD (2018) avoided the humanlike criterion, stating that AI is an activity, rather than simply objects such as machines. Unfortunately, OECD (2018) unnecessarily added the word "cognitive", meaning that even this definition was inaccurate. In 2019, the definition was revised again, removing the word "cognitive", to read: "An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy".

In the OECD (2019) definition, it is worth noting the phrase "given set of human-defined objectives". Since rationalization refers to optimization under human-defined objectives, the OECD (2019) definition can be seen as taking the "rational" perspective. It is also explained that AI technologies can only deliver value if they are part of the organization's strategy and are used in the right way (Hippold, 2019). This also corresponds to the phrase "given set of human-defined objectives" in OECD (2019). Gartner's criticism of AI misconceptions shows its "rational" approach to AI. It also criticizes humanlike AI, explaining that while some forms of AI might give the impression of being clever, it is unrealistic to think that current AI is similar or equivalent to human intelligence (Hippold, 2019).

4. Identifying the definition of AI's influence on policy: Document analysis

Through our document analysis we were able to find research that was very close to ours. Krafft et al. (2020) compares AI researchers' recognition of AI with policy reports' perspective of AI. Similar to our claim in this paper, Krafft et al. (2020) criticizes the human emphasis in the definition of AI in most AI policy reports, while noting that AI researchers' recognition is more inclined to rational emphasis. Krafft et al. (2020) found that 28% of definitions by AI researchers and 62% from published policy documents use the word "human". There was more disagreement over whether existential threats are relevant (42% agreed) – an issue more relevant to (hypothetical) humanlike AI. In our paper, we analyze AI policy-related reports and classified resources according to their definition or perspective on AI. We particularly focus on resources which define AI as humanlike (thinking or action) entities.

For the analysis, we had planned to perform document analysis to investigate their position on: (1) the concern, fear, peril, threat, and danger of AI; (2) the fairness of AI (discrimination, oppression, discrimination, and inequality); and (3) unemployment and job loss. However, it was difficult to obtain systematic results, since it is very time consuming to analyze the perspectives of reports only by human reading. At first, we considered automatic document analysis using AI techniques. However, it is still difficult to automate document analysis to replace human reading; although there is research on the subject, such as Hermann et al. (2015). In near future, AI-based document analysis software will help human researchers perform this kind of research. With such AI discipline-based software, human researchers will be able to improve their performance and reduce the necessary research time. During our research, because we could not find such software for our purposes, we narrowed our focus to only job related reports, then analyzed them by keyword search and human reading. Krafft et al.'s (2020) study also seems to be based on this method.

4.1. Relationship between the perception of AI and the expectation of job loss

We investigate the relationship between the perception of AI and the expectation of job loss incurred by AI. We conject that a policymaker who believes or defines AI as something that thinks or acts in humanlike manner will be likely to overemphasize AI's negative impact on job creation. We were able to find numerous reports using humanlike AI definitions, such as Miller-Merrell (2019), Molla (2019), and Hawksworth et al. (2018). For example, Miller-Merrell (2019) describes AI as a branch of computer science that uses machine learning algorithms which "mimic" cognitive functions, making machines more humanlike. While Molla (2019) explains machine learning as something that can make humanlike decisions.

4.2. AI-induced job loss expectation defining AI as humanlike

Policy reports

The report "Australia's Future Workforce?" by the Committee for Economic Development of Australia (CEDA, 2015) recognizes the ability of computers to emulate human thought patterns, claiming that AI is able to take over intellectual tasks, as well as routine ones. Hindi (2017) argues that the real issue facing governments today is the failure to transition to a sustainable AI society, which will lead to massive job loss and economic downturn. Hindi (2017) defines AI as the ability for a machine to reproduce human behavior. Daniel (2020) asserts that the pace at which AI is replacing the way humans work, forecasts that the future to be fully automated, even to the extent that jobs for humans will no longer exist. She explains that intelligent AI-models are trained to enable them to "act like a human" in real-world situations and that machines "think like human minds".

Business websites

Many business web sites also make similar mistakes. For example, John (2019) defines AI as computers or devices that mimic humanlike movements, and expects that with automation – the real essence of the AI revolution – robots will takeover of several jobs, although not all careers will be destroyed. Balatayan (2018) claims even white-collar jobs are being cut due to technological advancements, defining an AI system as any software that can mimic a rudimentary form of thinking.

McClelland (2020) explains that the impact of AI and automation will be profound, and that we need to prepare for a future where job loss reaches 99%. His definition of AI is based on the following two assumptions, that (1) we will continue making progress in building more intelligent machines, and (2) human intelligence arises from physical processes. With this in mind, McClelland (2020) concludes that we will build machines which have human-level or higher intelligence. However, these assumptions were criticized by George Zarkadakis in his seminal book, In Our Own Image. In it, he describes six metaphors that people have used over the past 2,000 years to try and explain human intelligence. Zarkadakis (2015) shows that each metaphor simply reflected the most advanced thinking of the time.

Consulting and research institute reports

Bughin et al. (2017) at McKinsey define AI as the ability of machines to exhibit humanlike intelligence, and explains that AI-powered automation could have a profound impact on jobs and wages. The Digital Marketing Institute (2019) raises the question, of whether AI will really steal our jobs in the future, and characterizes AI systems as being able to do things that humans can do and imitate the way we think. Wisskirchen et al. (2017) of the IBA Global Employment Institute describes AI as the work processes of machines that would require intelligence if performed by humans, asserting that both blue-collar and white-collar sectors will be affected.

Media reports

Dai and Jing (2018) of the South China Morning Post refers to Oxford-Yale AI impact research – based on a survey of 352 machine learning experts – which estimates that there is a 50% chance of AI outperforming humans in all tasks in just 45 years, and which could take over every job in the next

century. The research explains that AI is the science of "simulating" intelligent behavior in computers, enabling the latter to exhibit humanlike behavioral traits such as knowledge, reasoning, common sense, learning, and decision making. Knapton (2016) of the Telegraph reports that the rise of robots could lead to unemployment rates greater than 50%, and that many middle-class professionals' jobs would be outsourced to machines within the next few decades, leaving workers with more leisure time than ever. Such comments are common misconceptions of people who see AI as being humanlike. The report itself also uses the term humanlike robots. Kelly (2019) of Forbes maintains that AI, robotics, and technology will displace millions of workers, and defines AI as the ability of a machine to mimic human behavior.

Adel (2019) of Medium states that AI's effect on work will be disruptive, and predicts a future in which robots take jobs from human workers. Adel (2019) also defines AI as the act of "simulating the human brain" in a machine, i.e., creating an artificial human mind far more powerful than an actual human one. Wadhwa (2016) of FactorDaily argues that we are facing a jobless future because AI systems emulate the functioning of the human brain's neural networks. Xu (2017) of Northeastern's J-school's Ruggle Media reports that computers have become substitutes for various types of jobs for numerous reasons, such as recent developments in AI machine learning. Machine learning will not only reduce the huge demand for labor input with tasks since it can be routinized depending on pattern recognition, it will also increase the demand for labor-performing tasks that are not subject to computerization. Xu (2017) recognizes that every aspect of learning or any other feature of intelligence can, in principle, be so precisely described that a machine can be made to simulate it.

4.3. AI-induced job loss expectation regarding AI as a super-intelligent entity

Through the document analysis, we found a number of reports that regard AI as a competitor to humans, i.e., a superhuman entity. Although the reports do not explicitly describe AI as being humanlike, they also belong to the humanlike category. Cellan-Jones (2014) refers to Stephen Hawking's fears on the consequences of creating something that can match or surpass humans (who are limited by slow biological evolution), as well as the concerns that clever machines, capable of undertaking tasks performed by humans up until now, will swiftly destroy millions of jobs. Clifford (2017) refers to Elon Musk's belief that a machine could be far smarter than a human, that robots will be able to do jobs better than humans, and that there will certainly be job disruption. Manyika et al. (2017) of McKinsey is of a similar opinion, saying that "machines already exceed human performance". Finally, Niyazov (2019) assumes that AI algorithms and automated manufacturing are much better at performing tasks.

4.4. AI-induced job loss expectation without a specific definition of AI

There are also claims of job loss by AI without a specific definition of AI (Brynjolfsson & McAfee, 2011; Kurzweil Network, 2012; Frey & Osborne, 2013; World Economic Forum, 2016; Acemoglu & Restrepo, 2017; Frey & Osborne, 2017; Rieley, 2018; Lambert & Cone, 2019; Ambika, 2019; The Week, 2019; Muro et al., 2019). For example, Krafft et al. (2020) mentions that over 40% of policy reports do not have a definition of AI. Frey and Osborne (2013) of Oxford Martin School reports that 47% of total US employment is in the high-risk category, and that associated occupations are

potentially automatable over an unspecified number of years – perhaps a decade or two. The World Economic Forum (2016) holds that current trends could lead to a net employment impact of more than 5.1 million jobs lost to disruptive labor market changes from 2015–2020; with a total loss of 7.1 million jobs, two thirds of which are concentrated in the office and administrative job family, and a total gain of 2 million jobs in several smaller job families.

Using a model in which robots compete against human labor in various tasks, Acemoglu and Restrepo (2017) of the Massachusetts Institute of Technology (MIT) and Brown University show that robots may reduce employment and wages, and that the local labor market effects of robots can be estimated by regressing the change in employment and wages on the exposure to robots in each local labor market – defined from the national penetration of robots into each industry and the local distribution of employment across industries. Frey and Osborne (2017) of Oxford Martin School claim that recent developments in machine learning will put a substantial share of employment at risk across a wide range of occupations in the near future, and that nearly half of all US jobs were at risk from AI-powered automation. Rieley (2018) of the US Bureau of Labor Statistics also asserts that employment of bookkeepers is projected to decline 1.5% from 2016–2026, representing a loss of 25,200 jobs.

Ambika (2019) also maintains that AI technologies being adopted around the globe will replace numerous jobs currently being done by humans. The Week (2019) reports that over the next decade, automation and AI could put 54 million Americans out of work. Muro et al. (2019) of Brookings Institute reports that although robots are not replacing everyone, a quarter of US jobs will be severely disrupted as AI accelerates the automation of existing work. Lambert and Cone (2019) of OxfordEconomics.com claim that with the rise of robots in business models, many sectors will be seriously disrupted and millions of existing jobs will be lost, with 20 million manufacturing jobs set to be lost to robots by 2030.

Most job loss reports have either a “humanlike” definition, a “human-comparable” definition, or “no definition”. According to our definition of AI, we claim that job loss reports make mistakes due to the incorrect recognition and understanding of the characteristics of AI. We were unable to find job loss reports that define AI as rational, except for Russell (2019). Russell is a very respectable AI pioneer who wrote an innovative textbook on AI (Russell & Norvig, 1995). However, even though he makes an attempt, he confesses not to be qualified to opine on the job issue. Other AI experts, such as Lee (2018a), also make similar mistakes when defining AI by incorrectly emphasizing “humanlike” and “cognitive”. AI policies are too important to leave entirely to technical AI experts. As Russell (2019) asserts, the job issue is too important to leave entirely to economists. For example, Martin Ford, a journalist who is not an AI expert, wrote a book exaggerating job loss from AI (Ford, 2015). However, he seems to have changed his mind after interviewing numerous world-renowned AI experts (Ford, 2018). It is therefore necessary for us to explain AI to policy experts, as well as promote collaboration among AI and policy experts.

5. Automation creates more jobs than it eliminates: Learning from history

5.1. AI creates more jobs than it eliminates

Throughout the research, we found numerous reports claiming that AI will not eliminate jobs. Shrive (2018) claims that AI cannot replace humans in performing all tasks, especially in the property management domain. AI has been specifically developed to simplify repetitive and time-consuming processes, thereby freeing up time for property managers, letting agents, and contractors to deal with more pressing problems. Lokitz (2018) asserts that with every job taken over by a machine, there will be an equal number of opportunities for jobs to be done by people. Furthermore, in many cases, humans and machines will find themselves in symbiotic relationships, helping each other to do what they do best.

The World Economic Forum (2018) asserts that 38% of businesses surveyed expect to extend their workforce to new productivity-enhancing roles, more than a quarter expect automation to lead to the creation of new roles in their enterprise, about half of today's core jobs – making up the bulk of employment across industries – will remain stable up to 2022, and current estimates suggest a decline of 0.98 million jobs and a gain of 1.74 million jobs. Atkinson (2018) asserts that there is no reason to believe that this coming technology wave will be any different in pace and magnitude than previous waves. Each past wave has led to improved technology in a few key areas (e.g., steam engines, railroads, steel, electricity, chemical processing, and information technology), and these were then used by many sectors and processes. Within manufacturing, for example, each wave has led to important improvements, however, there have always been many other processes that have required human labor. The British Academy (2018) maintains that while there is now a consensus that AI does not spell the end of work, neither will the transition be painless for all. Although human-level intelligence ('general AI') receives significant media attention, it is still some time away from being delivered, and it is unclear when it might be possible. Krafft et al. (2020) points out that hype surrounding general AI centers on humanlike AI, and that it is a problem that many policy analysts think of it in this way.

AdextAI (2019) explains that, as the technology has evolved, unemployment rates have decreased as a result of the new jobs created. Naudé (2019) holds that, in the foreseeable future, AI is unlikely to cause huge job losses (or job creation), at least in advanced economies. The main reasons for this conclusion are based on: (1) the fact that the methods used to calculate potential job losses are sensitive to assumptions; (2) automation may affect tasks more significantly, rather than the jobs within which they are performed; (3) net job creation can be positive because automation stimulates the creation of new jobs or jobs elsewhere; (4) diffusion of AI may be much slower than is thought or assumed; and (5) the tempo of innovation in AI is slowing down. Thomas (2019) explains that AI is poised to eliminate millions of current jobs and create millions of new ones – some of which have yet to be invented. Liang (2019) describes that recent advances in AI, while seemingly impressive, are very narrow in scope and require a lot of human supervision and input to work in real applications. While as many as 47% of current jobs contain tasks that may be automatable, less than 5% of jobs will be fully automatable by 2030. As with many new technologies that came before, AI tools will augment and not replace workers by automating subtasks of a job.

5.2. Automation proved more of a blessing than a threat

Garry Kasparov says that he is the first knowledge worker whose job was threatened by a machine (Knight, 2020). Referring to Kasparov, Knight (2020) claims that technology destroys jobs before creating new ones. This story has been repeated since the Industrial Revolution in the 19th century. For example, with the emergence and popularity of machines in 19th century Britain, many workers lost their jobs. Luddism centered around the defense of hand trades in the textile industry in the face of innovation which threatened jobs (Beckett, 2012). Led by artisans who felt their jobs were being threatened by the increased use of machines in the production process, Luddites began destroying machines as a form of protest. An agricultural manifestation of Luddism occurred during the Swing Riots of 1830, which saw the destruction of threshing machines. Although automation freed people from mundane and repetitive tasks, it caused some people to lose their jobs.

William Lee was an English clergyman and inventor who, in 1589, devised the first stocking frame knitting machine, the design of which was used for centuries. Having perfected his design and desiring to secure Queen Elizabeth I's patronage, whose partiality for knitted silk stockings was well known, Lee went to London to exhibit the loom before the Queen. However, her reaction was not what he had expected. She is said to have opposed the invention on the grounds that it would deprive a large number of poor people of their employment of hand knitting (Smiles, 2005).

Although people have always been afraid of new automation technologies, they always proved more of a blessing than a threat. As machine learning systems learn from data, intelligent human beings should learn from history. In 1790, 90% of Americans were farmers. Nowadays that number is less than 2% (Dimitri et al., 2005). So, has American agriculture disappeared? The answer is no, it has simply become more automated. The US has transformed from an agricultural economy to an industrial economy, then to a service economy, and now to an information economy. Dimitri et al. (2005) concludes that automation creates far more jobs than it eliminates. Even if automation takes on a variety of professional roles, it does not always take away people's jobs.

5.3. The camera created more jobs and industries than it eliminated

Invented roughly 200 years ago, cameras began to be distributed about 100 years ago. At the time, many people thought that there would be no more need for artists as a result. However, cameras allowed for the development of modern art, and many painters used cameras in the studios. Even early contributors to the invention of photography and the camera were painters themselves, such as Leonardo da Vinci, who used the camera obscura for his painting, and Louis-Jacques-Mandé Daguerre, who was a theatre set painter and inventor of the daguerreotype process of photography (Daval, 1982). With cameras, i.e., the new automation technology of the time, painters were able to dramatically reduce the time needed for painting and sell photos of their works to more customers. The existing skills needed for drawing portraits, simply became the basis for becoming a better photographer. In other words, the new technology became an opportunity to expand the existing portrait market into the photography market (Benjamin, 1969).

In addition, the invention of camera allowed related industries to develop. New industries emerged, such as film manufacturing, camera manufacturing, film sales, photo album production, photo studios, photographic development, photo distribution,

newspapers, magazines, advertising, and publishing industries, etc. Cameras also contributed to the development of other industries. For example, as more people began to take cameras with them when they travelled, the photos being taken encouraged more people to travel. Cameras also had an impact on the movie industry (Jeong, 2015), while the influence of celebrities such as Marilyn Monroe and John F. Kennedy was greater as a result of photography. Today, not only do people take pictures with their smartphones, but the continued development of photography has created new businesses such as Facebook and Instagram.

5.4. Automobiles created jobs and industries

A photograph taken on 5th Avenue in New York in 1900 shows the horse and cart to be the predominant mode of transport. By 1913, in little more than a decade, the automobile had replaced the horse as the main form of transport. In turn, this led to the development of related industries, such as automobile manufacturers, mechanics, and automobile salesmen. In addition to the development of personal automobiles, the city bus, intercity bus, express bus, taxi, and trucking industries all developed. At the same time, the construction of roads and car parks resulted in an increase in jobs (Lee, 2018). Not only did automobiles spark a desire for long-distance travel, but by shortening travel times, the travel industry and related transportation, lodging, and restaurant industries also developed alongside one another.

5.5. Digital typesetting created more jobs by promoting publishing

Physical typesetting is the composition of text through the arranging of metal “types” and is most well-known in the production of newspapers in the late 19th century. Being a typesetter was a highly skilled position, so much so that when the Hankyoreh newspaper in Korea was founded in 1988, it was unable to find a skilled typesetter. To solve the problem, the newspaper introduced an innovative technology called the Computerized Typesetting System (CTS). Starting with the Hankyoreh newspaper, many newspapers in Korea soon adopted this system, leading to a lot of typesetters losing their jobs. At the same time, however, demand for digital typesetters increased, which the traditional typesetters quickly learned, becoming desktop publishing professionals (Lee et al., 2012).

5.6. ATMs created jobs by contributing to bank expansion

When Automated Teller Machines (ATMs) were first invented in the 1970s, there were serious concerns about the layoffs of tellers. In the 1980s, US banks introduced ATMs to improve work efficiency, with the number of employees per branch decreasing to one third as a result. Between 1995 and 2010, the number of ATMs in the US surged from 100,000 to 400,000. However, there was no massive unemployment, since the number of bank branches increased by more than 40%. Furthermore, by 2015, the number of bank employees had increased from 250,000 to 500,000. As the introduction of ATMs reduced the cost of creating new branches, banks were able to expand and hire more employees than in the past. In addition, with ATMs replacing simple deposit and withdrawal

services, banks were able to focus on developing profitable financial products such as loan counselling and insurance. As a result, bankers were freed up to perform more important tasks than ever before. Not only were new jobs created when ATMs took over performing simple and repetitive tasks, bankers were able to take charge of tasks requiring high-level capabilities (James, 2015; Deloitte, 2018).

5.7. Internet intermediaries created jobs by reintermediation

Baen and Guttery (1997) predicted that increased use of the Internet and information technology would have a dramatic and negative impact on the real estate industry in terms of both income and employment levels. They argued that buyers and sellers with access to information available via the Internet would have no need for traditional "infomediaries", and that several other players in real estate support positions would also be disintermediated by the Internet. The authors predicted job losses in sectors directly related to real estate, including sales agents and developers, as well as sectors involved in the support of real estate transactions, such as legal services and banking. Muhanna and Wolf (2002) revisited Baen and Guttery's (1997) examination of technology's effect on the real estate industry and found that, in general, their most ominous predictions of income and employment loss have not materialized. In the years since their 1997 article, according to the Bureau of Labor's statistics, the real estate industry, like most sectors in the US, has experienced steady growth. Specifically, more workers were employed as real estate agents, developers, and legal service providers.

It is often argued that as electronic markets lower the cost of market transactions, traditional roles for intermediaries will be eliminated, leading to "disintermediation". Bailey and Bakos (1997) discuss the findings of an exploratory study of intermediaries in electronic markets which suggests that markets do not necessarily become disintermediated as they become facilitated by information technology. Middle businesses, functions, or people need to move up the food chain to create new value or face being disintermediated. However, the "reintermediation" opportunities are greater than the disintermediation perils (Tapscott, 1997). Yoon (2015) also explains that attention should be paid to reintermediation, where the value of brokerage functions has been recently created. There will be an opportunity to create new value for middlemen connecting consumers and suppliers.

These aforementioned examples show that new technology does not threaten the existence of someone's job. Just as a painter adapted to the invention of the camera and found a new job in a related field, so will it be the same in the case of AI. People currently engaged in fields such as health care, architecture, and law, where AI is expected to be applied, will acquire AI-related skills and take on new jobs.

6. Summary and Conclusion

Incorrect or unscientific understanding of AI is still pervasive and misleads policymakers. While ambiguity in definition has hampered conversation, legal and regulatory intervention requires agreed-upon definitions. However, consensus over the definition of AI has been elusive thus far, especially in policy conversations (Krafft et al., 2020). In this study, we reviewed numerous definitions of AI, and based on our critical review, we suggest a scientific definition of AI. Namely, that AI is a discipline devoted to making entities and infrastructures intelligent, with the intelligence being that quality which enables agents, principals, and infrastructures to function appropriately. We have observed how, since 2017, OECD has continued to update its definition of AI; and have noted how OECD has improved its definition from humanlike to rational and from thinking to action.

We investigated numerous AI-related policy documents, particularly those dealing with the impact of AI on jobs, and found that those which view AI as a system that mimics humans are likely to overemphasize job loss incurred by AI as an automation technology. In addition, most job loss reports have either a "humanlike" definition, a "human-comparable" definition, or "no definition". We were unable to find job loss reports that defined AI as rational. Through our historical review, we showed that automation technology, such as photography, automobiles, ATMs, and the Internet as an automatic intermediation technology, did not reduce human jobs. Instead, they created numerous jobs and industries. AI will also create a wide range of jobs and industries, on which our future AI policies should instead focus. Similar to how machine learning systems learn from valid data, AI policy makers should learn from history to gain a scientific understanding of AI and an exact understanding of the effects of automation technologies. Ultimately, good AI policy comes from a good understanding of AI.

We suggest four policy recommendations as follows:

Recommendation 1: Policy experts should be well educated about what AI is and what is really going on in the AI researches and businesses. Especially, AI should be considered as a discipline making entities and infrastructures intelligent, and the intelligence is that quality that enables agents, principals, and infrastructure to function appropriately. AI should not be considered as human-like or super-human system. Past AI policies based on the old paradigm should be rewritten.

Recommendation 2: Government should make program for educating the administrative officials, policy experts in public-owned research institute, and lawmakers in the national assembly.

Recommendation 3: Just as machine learning systems learn from data, policymakers should also learn from history and data. The positive impacts of automation technology should be recognized by policy makers and the new AI policy should be established based on the new recognition.

Recommendation 4: Government and society should recognize the characteristics of AI, as an optimization system, to have more public benefit, faster business outcomes and less risks from AI adoption.

Acknowledgements

We would like to thank the Association of Pacific Rim Universities (APRU) for initiating the "AI for Social Good" project, of which this study is a part. We would like to thank Prof. Jiro Kokuryo of Keio University, Japan, the Principal Investigator of this project, for giving us the opportunity to be involved in such an exciting project. Our thanks must also go to Christina Schönleber, Director for Policy and Programs, APRU, as well as all of my colleagues on the project, from whom I have learned a great deal.

References

Acemoglu, D., & Restrepo, P. (2017). Robots and Jobs: Evidence from US Labor Markets. *NBER Working Paper No. 23285*.

Adel, K. (2019). The Future of Jobs in Artificial Intelligence Era. *medium.com*. Retrieved from https://medium.com/analytics-vidhya/the-future-of-jobs-in-artificial-intelligence-era-93e34c33c25f

Adext AI. (2019). "How Many Jobs Will Be Lost Because of Artificial Intelligence?" Is the Wrong Question. *Adext AI*. Retrieved from https://blog.adext.com/jobs-lost-artificial-intelligence/

Albus, J. (1991). Outline for a theory of intelligence. *IEEE Transactions on Systems, Man, and Cybernetics, 21*(3).

Alpeyev, P., & Amano, T. (2016). A Japanese Billionaire's Robot Dreams Are on Hold. *Bloomberg*. Retrieved from https://www.bloomberg.com/news/articles/2016-10-27/a-japanese-billionaire-s-robot-dreams-are-on-hold

Anastasi, A. (1992). What Counselors Should Know About the Use and Interpretation of Psychological Tests. *Journal of Counseling & Development, 70*(5).

Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J. C., Barends, R., . . . Collins, R. (2019). Quantum supremacy using a programmable superconducting processor. *Nature, 574*(7779), 505-510.

Atkinson, R. (2018). Shaping structural change in an era of new technology. *policynetwork.org*.

Auer-Welsbach, C. (2019). Interview with cognitive scientist Newton Howard on AI. *Medium*.

Baen, J. S., & Guttery, R. S. (1997). The Coming Downsizing of Real Estate: Implications of Technology. *The Journal of Real Estate Portfolio Management, 3*(1), 1-18.

Bailey, J. P., & Bakos, Y. (1997). An exploratory study of the emerging role of electronic intermediaries. I*nternational Journal of Electronic Commerce, 1*(3), 7-20.

Baltayan, A. (2018). Robots, Automation & Technology Taking Over – Is Your Job at Risk? *Money Crashers*. Retrieved from https://www.moneycrashers.com/robots-automation-technology-replacing-jobs/

Beckett, J. (2012). *Luddites*. Retrieved from The Nottinghamshire Heritage Gateway: http://www.nottsheritagegateway.org.uk/people/luddites.htm

Bellman, R. E. (1978). *An Introduction to Artificial Intelligence: Can Computers Think?* San Francisco: Boyd & Fraser Pub. Co.

Benjamin, W. (1969). *Das kunstwerk im zeitalter seiner technischen reproduzierbarkeit.* Frankfurt am Main: Suhrkamp.

Bernoulli, D. (1738). *Hydrodynamica.* France: The University of Strasbourg: Johann Reinhold Dulsecker.

Bessen, J. (2015). Toil and Technology. *Finance and Development, 52*(1).

Biven, M. (2016). Pepper Salé: Lessons from the bitter Aldebaran / SoftBank project. Retrieved from https://markbivens.com/m/archives/pepper-sale-lessons-from-the-bitter-aldebaran-softbank-project

Bloomberg, J. (2017). https://www.forbes.com/sites/jasonbloomberg/2017/07/02/is-ibm-watson-a-joke/. *Forbes.*

Bordes, A., Weston, J., Chopra, S., Mikolov, T., Joulin, A., Rush, S., & Bottou, L. (2015). Artificial Tasks for Artificial Intelligence. *Facebook AI Research ICLR*.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* UK: Oxford University Press.

Boxall, A. (2017). Pepper is everywhere in Japan, and nobody cares. Should we feel bad for robots?

British Academy. (2018). The impact of artificial intelligence on work. *royalsociety.org*.

Brown, J. (2017). Why Everyone Is Hating on IBM Watson—Including the People Who Helped Make It. *GIZMODO*. Retrieved from https://gizmodo.com/why-everyone-is-hating-on-watson-including-the-people-w-1797510888

Brown, N., & Sandholm, T. (2017). Safe and Nested Endgame Solving for Imperfect-Information Games. *In Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Brynjolfsson, E., & McAfee, A. (2011). *Race against the machine: how the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy.* Digital Frontier Press.

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlstrom, P., . . . Trench, M. (2017). Artificial intelligence: the next digital frontier? *McKinsey and Company Global Institute,* 1-80.

Choudhury, A. (2019). AI May Kill These 5 Jobs By 2030, Say Experts. *Analytics India Magazine.* Retrieved from https://analyticsindiamag.com/ai-may-kill-these-5-jobs-by-2030-say-experts/

Campanella, E. (2016). *Meet Sophia, the human-like robot that wants to be your friend and 'destroy humans'.* Retrieved from Global News: https://globalnews.ca/news/2888337/meet-sophia-the- human-like-robot-that-wants-to-be-your-friend-and-destroy-humans/

Card, S., Moran, T., & Newell, A. (1983). *The Psychology of Human-Computer Interaction.* Hillsdale.

Carr, J. A. (2014). An Introduction to Genetic Algorithms. *Senior Project, 1*(40), 7.

CEDA. (2015). Australia's future workforce? *CEDA.*

Cellan-Jones, R. (2014). *Stephen Hawking warns artificial intelligence could end mankind.* Retrieved from BBC: https://www.bbc.com/news/technology-30290540

Charniak, E., & McDermott, D. (1985). *Introduction to Artificial Intelligence.* Unitted States: Addison-Wesley Longman Publishing Co., Inc.

Clifford, C. (2017). *Elon Musk: 'Robots will be able to do everything better than us'.*

Dai, S., & Jing, M. (2018). *Worried AI will replace your job?Here's an explainer to prepare for that day.* Retrieved from SCMP: https://www.scmp.com/tech/innovation/article/2131339/worried-ai-will-replace-your-jobheres-explainer-prepare-day#:~:text=Tech%20%2F%20Innovation-,Worried%20AI%20will%20replace%20your%20job%3FHere's%20an,to%20prepare%20for%20that%20day&text=The%20Oxford%2D

Daniel, E. (2020). *Role of Artificial Intelligence in Human Revolution.* Retrieved from Thrive Global: https://thriveglobal.com/stories/role-of-artificial-intelligence-in-human-revolution/

Daval, J.-L. (1982). *Photography History of an Art.* First American Edition.

Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). MIT Press.

Deloitte. (2018). *The Future of Work.* One and One Books.

Digital Marketing Institute. (2019). *The Rise of AI: Will It Take or Create Digital Jobs?* Retrieved from DMI Blog: https://digitalmarketinginstitute.com/blog/the-rise-of-ai-will-it-take-or-create-digital-jobs

Dimitri, C., Effland, A., & Conklin, N. (2005). The 20th Century Transformation of U.S. Agriculture and Farm Policy. *Economic Information Bulletin,* 17.

Drum, K. (2017). *You Will Lose Your Job to a Robot—and Sooner Than You Think.* Retrieved from Mother Jones: https://www.motherjones.com/politics/2017/10/you-will-lose-your-job-to-a-robot-and-sooner-than-you-think/

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science.* UK: Cambridge University.

Ema, A., Akiya, N., Osawa, H., Hattori, H., Oie, S., Ichise, R., & Kanzaki, N. (2016). Future Relations between Humans and Artificial Intelligence: A Stakeholder Opinion Survey in Japan. *IEEE Technology and Society Magazine, 35*(4), 68-75.

Engelbart, D. C. (1962). *Augmenting Human Intellect: A Conceptual Framework.* Stanford Research Institute.

Eysenck, M. W., Hunt, E., Ellis, A., & Johnson-Laird, P. N. (1991). T*he Blackwell Dictionary of Cognitive Psychology.* UK: Wiley-Blackwell.

Ferrucci, D. (2012). ntroduction to "This is Watson". *IBM Journal of Research and Development.*

Ford, M. (2015). *Rise of the Robots: Technology and the Threat of a Jobless Future.* Basic Books.

Ford, M. (2018). *Architects of Intelligence: The Truth about AI from the People Building it.* Packt Publishing Ltd.

Frey, C. B., & Osborne, M. A. (2013). The Future of Employment: How Susceptible are Jobs to Computerisation? *The Oxford Martin Programme on Technology and Employment.*

Frey, C. B., & Osborne, M. A. (2017). The Future of Employment: How Susceptible are Jobs to Computerisation? *Technological Forecasting and Social Change, 114,* 254-280.

Gao, J., Galley, M., & Li, L. (2018). *Neural Approaches to Conversational AI.* Retrieved from https://arxiv.org/pdf/1809.08267.pdf

Gershman, S. J., Horvitz2, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349*(6245), 273-278.

Goodfellow, I. J., Vinyals, O., & Saxe, A. M. (2015). Qualitatively characterizing neural network optimization problems. *arXiv*.

Haugeland, J. (1985). *Artificial intelligence: The very idea.* MIT Press.

Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics, 4*(2), 100-107.

Hassabis, D. (2015). *DeepMind Technologies - The Theory of Everything.* Retrieved from Google Zeitgeist.

Hawksworth, J., Berriman, R., & Cameron, E. (2018). *Will robots really steal our jobs? An international analysis of the potential long term impact of automation.* PwC.

Hayes, P., & Ford, K. M. (1995). *Turing Test Considered Harmful.* International Joint Conference on Artificial Intelligence.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2014). Teaching Machines to Read and Comprehend. *In Advances in neural information processing systems,* 1693-1701.

Herper, M. (2017). *MD Anderson Benches IBM Watson In Setback For Artificial Intelligence In Medicine.* Retrieved from Forbes: https://www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine/#395dbe613774

Hindi, R. (2017). *How my research in AI put my dad out of a job.* Retrieved from Medium: https://medium.com/snips-ai/how-my-research-in-ai-put-my-dad-out-of-a-job-1a4c80ede1b0

Hippold, S. (2019). *Gartner Debunks Five Artificial Intelligence Misconceptions.* Retrieved from Gartner: https://www.gartner.com/en/newsroom/press-releases/2019-02-14-gartner-debunks-five-artificial-intelligence-misconce

Hurley, S. (1998). *Consciousness in Action.* United States: Harvard University Press.

Jeong, M. S. (2015). *Humanities travel through film.* Kyungsung University.
12 Jobs that Will Be Soon Replaced by AI. (2019). Retrieved from apruve: https://blog.apruve.com/12-jobs-that-will-be-soon-replaced-by-ai

Jordan, M. (2019). Artificial Intelligence—The Revolution Hasn't Happened Yet. *Harvard Data Science Review.*

Judd, J. S. (1990). *Neural Network Design and the Complexity of Learning.* MIT Press.

Kelly, J. (2019). *Unbridled Adoption of Artificial Intelligence May Result in Millions of Job Losses and Require Massive Retraining for Those Impacted.* Retrieved from Forbes: https://www.forbes.com/sites/jackkelly/2019/09/30/unbridled-adoption-of-artificial-intelligence-may-result-in-millions-of-job-losses-and-require-massive-retraining-for-those-impacted/#6cd5dac51de7

Knapton, S. (2016). *Robots will take over most jobs within 30 years, experts warn.* Retrieved from The Telegraph: https://www.telegraph.co.uk/news/science/science-news/12155808/Robots-will-take-over- most-jobs-within-30-years-experts-warn.html

Knight, W. (2020). *Defeated Chess Champ Garry Kasparov Has Made Peace with AI.* Retrieved from Wired: https://www.wired.com/story/defeated-chess-champ-garry-kasparov-made-peace-ai/

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2019). Defining AI in Policy versus Practice. *arXiv*.

Kubacki, J. (2009). Artificial intelligence. *SpringerLink.* Retrieved from SpringerLink.

Kurzweil, R. (1990). *The Age of Intelligent Machines.* MIT Press.

Kurzweil Network. (2012). *2 Billion Jobs to Disappear by 2030.* Retrieved from Kurzweil Accelerating Intelligence: https://www.kurzweilai.net/2-billion-jobs-to-disappear-by-2030#!prettyPhoto

Lambert, J., & Cone, E. (2019). How Robots Change the World. *Oxford Economics.*

Lee, K. (2014). Human Robot Era is Far yet. *Money Today.*

Lee, K.-F. (2018a). *AI Superpowers: China, Silicon Valley, and the New World Order.* Houghton Mifflin Harcourt.

Lee, Y. J., Jang, S. L., & Kim, W. J. (2012). *Gutenberg's Return* . Idambooks (Korean).

Lee, S. (2018b). *The Future of the 4th Industrial Revolution.* One and One Books (Korean).

Liang, J., Ramanauskas, B., & Kurenkov, A. (2019). *Job Loss Due To AI — How Bad Is It Going To Be?* Retrieved from Skynet Today: https://www.skynettoday.com/editorials/ai-automation-job-loss

Licklider, J. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics.*

Lokitz, J. (2018). *The future of work: How humans and machines are evolving to work together.* Retrieved from businessmodelsinc.com: https://www.businessmodelsinc.com/machines/

Luger, G. F., & Stubblefield, W. A. (1993). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving (2nd ed.).* Benjamin-Cummings Publishing Co., Inc.

Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., . . . Sanghvi, S. (2017). Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages. *McKinsey Global Institute,* 1-160.

McClelland, C. (2020). *The Impact of Artificial Intelligence - Widespread Job Losses.* Retrieved from iotforall.com: https://www.iotforall.com/impact-of-artificial-intelligence-job-losses/

Minsky, M., & Papert, S. (1969). *Perceptrons: an introduction to computational geometry.* MIT Press.

Miller-Merrell, J. (2019). *Resources, How Artificial Intelligence (AI) is Changing Human.* Retrieved from Randstad RiseSmart: https://www.randstadrisesmart.com/blog/how-artificial-intelligence-ai-changing-human-resources

Mogg, T. (2018). *Pepper the robot fired from grocery store for not being up to the job.* Retrieved from Digital Trends: https://www.digitaltrends.com/cool-tech/pepper-robot-grocery-store/

Molla, R. (2019). *"Knowledge workers" could be the most impacted by future automation.* Retrieved from Vox: https://www.vox.com/recode/2019/11/20/20964487/white-collar-automation-risk-stanford- brookings

Muro, M., Maxim, R., & Whiton, J. (2019). *Automation and Artificial Intelligence: How machines are affecting people and places.* Retrieved from Brookings Metropolitan Policy Program: https://www.brookings.edu/research/automation-and-artificial-intelligence-how-machines-affect-people-and-places/

Naudé, W. (2019). The Race against the Robots and the Fallacy of the Giant Cheesecake: Immediate and Imagined Impacts of Artificial Intelligence. I*ZA Discussion Paper no. 12218*.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *arXiv*.

Muhanna, W. A. (2002). The Impact of E-Commerce on the Real Estate Industry: Baen and Guttery Revisited. *Journal of Real Estate Portfolio Management*(2), 141-152.

Nichols, G. (2018). *Robot fired from grocery store for utter incompetence.* Retrieved from ZDNet: https://www.zdnet.com/article/robot-fired-from-grocery-store-for-utter-incompetence/

Nilsson, N. J. (2010). *The Quest for Artificial Intelligence: A History of Ideas and Achievements.* UK: Cambridge University Press.

Niyazov, S. (2019). *How the Replacement of Blue-Collar Jobs by AI Will Impact the Economy.* Retrieved from iotforall.com: https://www.iotforall.com/how-ai-replacing-blue-collar-jobs-impact-economy/#:~:text=AI%20Is%20Emerging&text=The%20research%20conducted%20by%20Accenture,%25%20to%204.6%25%20by%202035.&text=This%20will%20boost%20exports%2C%20encourage,the%20US%20a%20manufac

OECD. (2017). OECD Science, Technology and Industry Scoreboard 2017. *OECD.*

OECD. (2018). AI: Intelligent machines, smart policies. *OECD Digital Economy Papers,* 0-33.

Protevi, J. (2006). *A Dictionary of Continental Philosophy.* United States: Yale University Press.

OECD. (2019). *Recommendation of the Council on Artificial Intelligence.* Retrieved from OECD: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Rich, E., Knight, K., & Nair, S. (1985). *Artificial intelligence.* New York: McGraw-Hill.

Rich, E., Knight, K., & Nair, S. (1991). *Artificial Intelligence.* New York: McGraw-Hill.

Rieley, M. (2018). In the money: occupational projections for the financial industry. *Beyond the Numbers, 7*(16).

Roth, G., & Dicke, U. (2005). Evolution of the brain and intelligence. *TRENDS in Cognitive Sciences, 9*(5), 250-257.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning Internal Representations by Error Propagation.* MIT Press.

Norvig, P., & Russell, S. J. (1995). *Artificial Intelligence: A Modern Approach.* United States: Prentice Hall.

Russell, S. J. (2019). *Human Compatible: Artificial Intelligence and Problem Control*. Viking.

Schalkoff, R. J. (1990). *Artificial Intelligence Engine.* United States: McGraw-Hill, Inc.

Page, S. E. (2018). *The Model Thinker: What You Need to Know to Make Data Work for You.* United States: Basic Books.

Shrive, T. (2018). *AI will never replace jobs in the property market.* Retrieved from Finance Digest: https://www.financedigest.com/ai-will-never-replace-jobs-in-the-property-market.html#:~:text=AI%20mimics%20human%20behaviour%20and,%2C%20in%20theory%2C%20be%20automated.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. v., . . . Sutsk. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587).

Simon, H. A. (1969). T*he Sciences of the Artificial.* MIT Press.

Smiles, S. (2005). *Rev. William Lee, inventor of the Stocking Frame.* Retrieved from victorianweb.org: http://www.victorianweb.org/technology/inventors/lee.html

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., . . . Tambe, M. (2016). *Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence.* Report of the 2014 study panel, Stanford University.

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum, 56*(4), 24-31.

Sun, S., Cao, Z., Zhu, H., & Zhao, J. (2019). A Survey of Optimization Methods from a Machine Learning Perspective. *arXiv*.

Barto, A., & Sutton, R. S. (1998). *Reinforcement Learning: An Introduction.* MIT Press.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *arXiv*.

Tapscott, D. (1997). Strategy in the new economy. *Strategy & leadership, 25*(6), 8-15.

The Week. (2019). *Will you lose your job to a robot?* Retrieved from The Week: https://theweek.com/articles/866339/lose-job-robot

Thomas, M. (2019). *Artificial Intelligence's Impact on the Future of Jobs.* Retrieved from builtin.com: https://builtin.com/artificial-intelligence/ai-replacing-jobs-creating-jobs

Turing, A. (1950). *Computing Machinery and Intelligence.* Mind.

Ulanoff, L. (2018). *Say Hello to Our Disappointing Robot Future.* Retrieved from Medium: https://medium.com/@LanceUlanoff/say-hello-to-our-disappointing-robot-future-e6f7d1d42e24

Wadhwa, V. (2016). *We are heading towards a jobless future: is it good or bad?* Retrieved from Factor Daily: https://factordaily.com/artificial-intellgence-automation-india-jobless-future/

Weng, J. (2002). A theory for mentally developing robots. *Proceedings 2nd International Conference on Development and Learning,* 131-140.

Winston, P. H. (1992). *Artificial Intelligence (Third edition).* United States: Addison-Wesley Longman Publishing.

Wisskirchen, G., Biacabe, B. T., Bormann, U., Muntz, A., Niehaus, G., Soler, G., & Brauchitsch, B. v. (2017). *Artificial Intelligence and Robotics and Their Impact on the Workplace.* IBA Global Employment Institute.

World Economic Forum. (2016). *The Future of Jobs Report 2016.* World Economic Forum.

World Economic Forum. (2018). *The Future of Jobs Report 2018.* World Economic Forum.

Xu, J. (2017). Will human beings lose jobs due to AI? *Ruggles Media.*

Yang, L.-A., Liu, J.-P., Chen, C.-H., & Chen, Y.-p. (2016). Automatically Proving Mathematical Theorems with Evolutionary Algorithms and Proof Assistants. *arXiv*.

Yoon, S. (2015). *Digital Economy Leadership by Don Tapscott.* Retrieved from https://www.mk.co.kr/news/business/view/2015/11/1081476/

Zarkadakis, G. (2016). *In Our Own Image: Savior or Destroyer? The History and Future of Artificial Intelligence.* Pegasus Books.