Brian Nguyen
1726 Canyon Circle
San Luis Obispo, CA 92630

June 3, 2021

Dr. Mirjam Eckert
Chief Publishing Officer
Frontiers Media Inc.
Holyoke Building
107 Spring Street
Seattle, WA 98104

Dear Dr. Eckert,

My team and I would like to publish our analytical report titled “Protecting Our Future: Regulation and Principles for Artificial General Intelligence” through the Frontiers Media platform with the intention of guiding future AGI developers and policymakers into safeguarding AGI development. Within the analytical report, you will find information addressing the power of artificial general intelligence (AGI) and our analysis for how to mitigate its existential threat to humanity.

As a team of four computer science and computer engineering students at Cal Poly San Luis Obispo, we have spent five weeks researching AGI development, AI ethical principles, and AI governance. From our research, we have discovered that AGI in its full development will surpass any form of human intelligence and inflict its own calculated opinion of society, which we cannot ensure will be benevolent. This, as a result, leaves society vulnerable. Our analysis of AGI mitigation thus elaborates on specific governing guidelines and overlapping AGI ethical principles that must be implemented into society to ensure the adoption of safe AGI, and it will prove to be immensely valuable if shared with AGI developers and policymakers.

We would like to thank you for reviewing our report for possible publication on the Frontiers Media platform. If there are any additional questions upon review, please contact me.

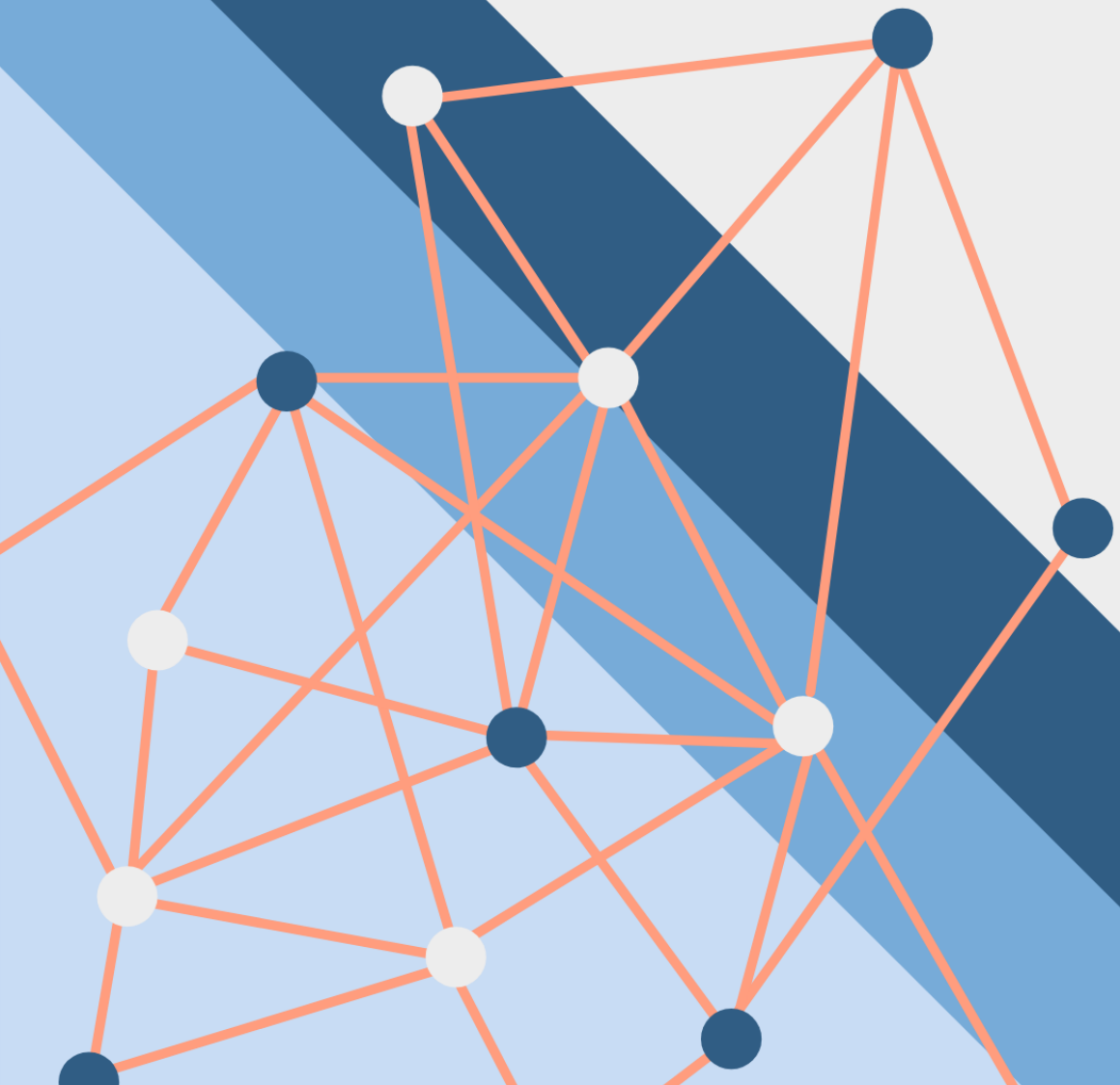
Sincerely,

Brian Nguyen

Brian Nguyen
bnguy203@calpoly.edu
(949) 343-6007

PROTECTING OUR FUTURE

Regulation and Principles for Artificial General Intelligence



ABSTRACT

"Protecting Our Future:
Regulation and Principles for Artificial General Intelligence"

Prepared By: Alisha Cherian, Brian Nguyen, Ethan Outangoun, Kassandra Winter

Discussions regarding the responsible development of artificial intelligence are becoming much more prevalent as artificial general intelligence approaches reality. AGI has the potential to disrupt many functions of society as we know it, and many experts in the field of artificial intelligence are concerned that AGI will pose an existential threat to humanity. It is imperative that guidelines be set for AGI before it exists. This report seeks to explore potential guidelines for the development and use of AGI and elaborate on important considerations to have in developing them. The process of AI governance is explored on a broad level, along with discussion of regulations that could serve as a baseline. Legislative bodies, agencies, and court systems should work together to improve the efficiency in creating and enforcing policies. Licensing and certification of AGI should be considered as regulatory approaches as they are less time consuming than creating laws and can complement each other. A combination of regulatory laws, licensing, and certification can better ensure responsible development. Ethical development of AGI is further explored through the discussion of principles that can be adopted universally for the safety and betterment of the human race. More specifically, this report analyzes five shared ethical principles, namely value alignment, transparency, bias mitigation, privacy, and accountability, that are common amongst AI experts and AI-driven institutions that seek to promote the responsible development of beneficial AGI.

Keywords: artificial intelligence, artificial general intelligence, governance, regulation, ethical principles, responsible development

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

List of Tables

List of Figures

1.0 INTRODUCTION

Artificial intelligence is defined by the Encyclopedia Britannica as "the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings" (Copeland, n.d.). While even those who are experts in the field of artificial intelligence struggle to agree on a definition of it, this is a broad definition of artificial intelligence that serves to establish the general essence of its capabilities. While the field of artificial intelligence is rapidly developing and becoming more and more ubiquitous in our society over time, the technology still has a long way to go. Currently, all of the artificial intelligence that has been developed falls under the category of artificial narrow intelligence, or ANI. ANI is artificial intelligence that is only capable of performing very targeted tasks that it has been specifically trained to do. However, with the rapid development of artificial intelligence, many expect a time where artificial general intelligence, or AGI, is developed. AGI is artificial intelligence that is capable of performing any task that it is given using human-like reasoning skills. Currently, artificial general intelligence does not exist, and there are conflicting views on when it will be developed, if at all. However, Müller and Bostrom (2016) surveyed 550 AI experts and found that 2050 was the median estimate for when there would be a 50% chance of AGI being developed. Due to the powers and capabilities of artificial general intelligence, many people fear that artificial general intelligence could pose a threat to society. Given the potential risks of AGI, it is necessary to have the conversation of how we would regulate it before it is developed.

While one might argue that there is a chance that artificial general intelligence will never get developed, the existence of this chance does not justify failing to be prepared for its creation. Often, experts only turn towards ethical questions once a technological feat has been achieved. However, with the rapid rate of AI development and the potential ramifications of AGI, failure to discuss the regulation of AGI now risks putting us in the dangerous scenario where AGI is already posing a threat to society before we have ethical principles and regulations for AI in place. We are currently at a critical juncture with AI, where the potential consequences of failing to regulate it are large and irreversible. We cannot allow ourselves to reach a point of no return where AI has already gotten out of control before until there is no chance to control it.

Additionally, even if AGI is never developed, these conversations are beneficial to have because they can lay the groundwork for regulation of artificial intelligence as a whole. Artificial intelligence does not need to become self-aware and human-like before we start taking its implications seriously. In the words of Sarangi and Sharma (2019), "Policy inaction and missteps with the future of AI have the potential to intensify grievances. The time to act is now as tomorrow might be just too late". It is better to control the direction of AI while we still have the chance to than it is to speed towards developing AI without certainty of what direction it will take us in.

This report serves to give an overview of major considerations to be had while discussing regulation and ethical principles for artificial general intelligence. In it, we go over why it is necessary to discuss regulation of AGI, challenges that will come with regulation of AGI, as well

as the implementation of AGI ethics through looking at human values, transparency, bias, and privacy. Given that we cannot predict exactly how AGI will work and behave, much of the conversation surrounding AGI is speculative, meaning that recommendations for regulation and ethical development of AGI are often broad, with room for refinement depending on the direction that AGI takes.

2.0 CONCERNS OF AGI

With great power comes great risk. It could become difficult or impossible for humans to control artificial intelligence in the scenario where artificial intelligence becomes more intelligent than humans and further becomes “super intelligent”.

2.1 General Concerns of AGI

Once AGI is fully developed, there arises the concern of how AGI impact society. Ideally, AGI agents would be used to serve humanity but we must also consider how AGI can be determinantal to society. In the following section, we discuss the four general risks to how AGI can potentially harm society.

Autonomous Weapons

It is highly plausible that the world could enter an autonomous weapons race using AGI. This could lead to many potential threats. One threat is another world war. Countries would use these unpredictable autonomous weapons that are programmed to kill and destroy other places for power or resources, without thinking of the consequences. If AGI was put in the control of an individual, organization, or government that did not value human life, it would have the potential to completely wipe out not just all of the world’s resources, but the entire human race. Russia’s president Vladimir Putin has explored this idea by saying “Artificial general intelligence is the future, not only for Russia, but for all humankind. It comes with enormous opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world” (Polyakova, 2019). Another potential threat of an autonomous weapons race is the concern that these AGI autonomous weapons might gain a mind of their own. If these weapons gained the capability to override their own instructions, it would be very difficult to gain control of, dismantle, or combat them. Once that point has been reached, the AGI has complete control over its own power, and could use it in whatever way it wants, having the potential to bomb any place or person that its conscience chooses.

Social Manipulation

Another potential risk of AGI is the manipulation of people by the vast amount of data it has access to. Currently, ANI makes it very easy to effectively target individuals and market products that they would enjoy to them. This is because ANI can be used by social media platforms to figure out who we are and what we like, as they are incredibly good at surmising what we think. The danger of this situation would be amplified immensely through the development of AGI. AGI, with the ability to make decisions for itself, could target specific people and target them with whatever information they like. This information could be in whatever format they find most convincing, and humans could find it impossible to distinguish whether it is fact or fiction. Thus, AGI machines could tweak the lives of millions just from a single advertisement or post that they decide to share. An example scenario of this would be if an AGI decided that a certain presidential candidate should win the election. The AGI would not

only be able to control the candidate's advertisements so that users only saw positive information about them, but it would also have the potential to access the competition's advertisements and promote false information that would lead to their loss. This scenario would affect everyone from the candidates to the dynamics between countries all around the world. This type of social manipulation could also be seen as an invasion of privacy and social oppression. Since AGI could collect, track, and analyze everything about you, it is very possible for those machines to use that information against you.

Mass Job Loss

Human job security is also at huge risk with the development of AGI. Because AGI has the intelligence of humans without many of the limitations of humans, they could complete human jobs and tasks more efficiently than ever before. For example, if AGI took over the simple tasks that doctors usually spend hours slaving over, it would allow doctors to spend more time with patients as well as allow doctors the time to take better care of themselves. However, this situation could go south fast if AGI took over all doctor's jobs. This would cause millions to be out of jobs and in debt, ultimately causing the economic market to crash. This situation would be exponentially worse if AGI completely took over every job and responsibility. In this case, we could never shut off the AGI, and would have to accept as a society that our entire existence relies on a machine that has no real consciousness.

Misalignment of Morals

A fourth risk of AGI arises from the situation where AGI and humans do not share the same basic values of life. AGI needs to understand what is seen as right or wrong in any given social situation so that they do not use their power for the wrong reasons. An example of this is if an AGI was given the objective of curing cancer. At first glance, this goal seems straightforward and beneficial to humankind. However, if the AGI does not understand human emotions or ethics, it might go to the most efficient way of curing cancer: killing everyone that has it. A failure to align human morals with AGI morals could lead to a scenario where AGI achieves a beneficial goal in a destructive way.

2.2 Concerns about AGI from Professionals

While concerns about artificial general intelligence may seem like an exaggeration based purely on decades of science fiction about the human race being overthrown by robots, these concerns do have basis in reality. Many AI experts have expressed concern about the potential ramifications of artificial general intelligence, with some going as far as to say that it will pose an existential threat to humanity.

Stephen Hawking

"The development of full artificial general intelligence could spell the end of the human race."
- Stephen Hawking (Cellan-Jones, 2014)

According to Stephen Hawking, the development of AGI could lead to the end of the human race. Once we create such an AI, it is capable of taking off on its own and redesigning itself at an ever-increasing rate. His belief was that humans, who are restricted by slow biological evolution, would not be able to compete and would be superseded without a doubt.

Elon Musk

"Mark my words – A.I. is far more dangerous than nukes."
- Elon Musk (SXSW, 2018)

Elon Musk, the co-founder of PayPal and Tesla Motors and the founder of SpaceX, also has similar views about machine intelligence. According to him, AI could be more dangerous than nuclear warheads, and the speculation that AI could become a million times intelligent than humans is actually an understatement. He also said that AGI is our biggest existential risk.

Nick Bostrom

"Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. We have little idea when the detonation will occur, though if we hold the device to our ear we can hear a faint ticking sound."
- Nick Bostrom (Bostrom, 2017)

Nick Bostrom is a philosopher at the University of Oxford who introduced the idea of an "existential risk" and has a degree in artificial intelligence. He has written numerous research papers on the ethics of artificial intelligence, as well as the existential risks posed by it. According to Nick Bostrom, computers with human-like intellectual capabilities could rapidly lead to developments in technology that could deliberately or accidentally destroy humanity with ease. He believes that sentient machines are a greater threat to human existence than climate change.

Other Experts

In the field of AI, experts have differing opinions on whether or not AGI will be a net positive or negative for the world. However, the concern that AGI will pose an existential threat to humanity does exist amongst experts. As shown in Figure A. Results of Müller and Bostrom Survey, 18% of those surveyed believed that such a technology would pose an existential threat to humanity, and 13% believed there would be net disadvantages.

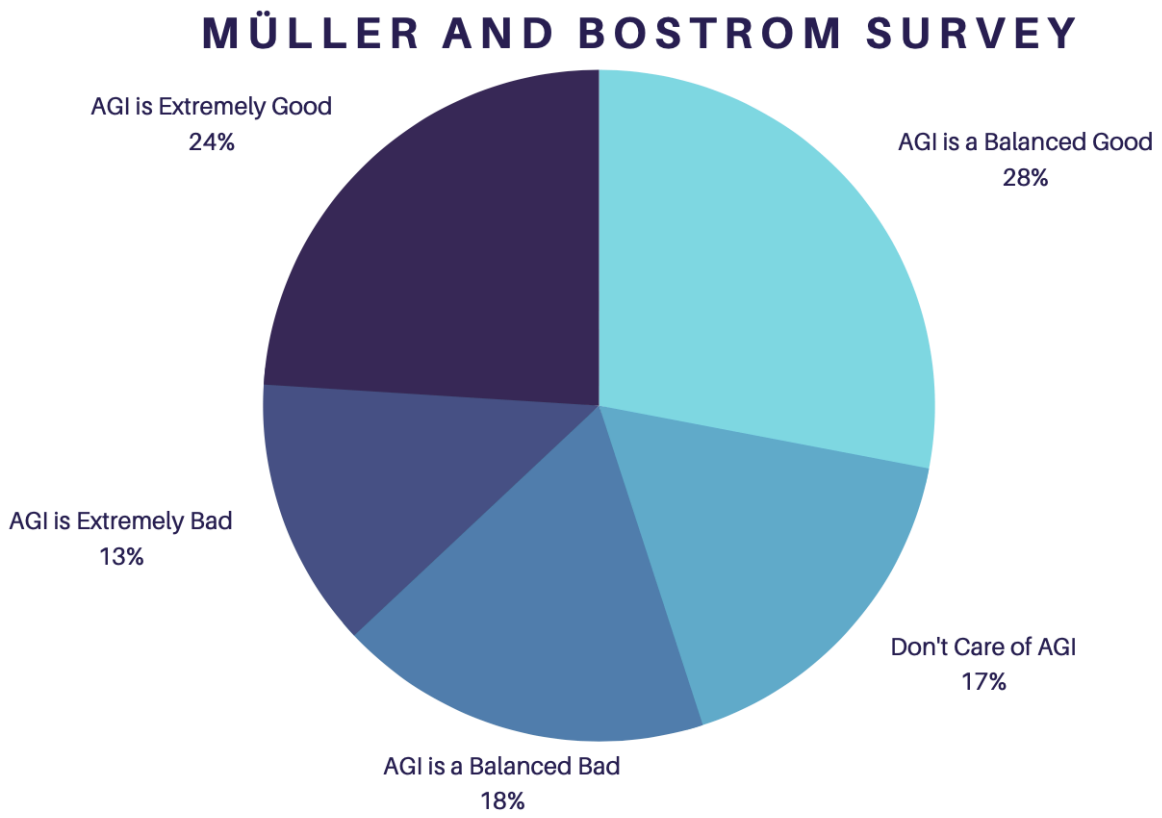


Figure A. Results of Müller and Bostrom Survey
Source. Adopted from Müller and Bostrom (2016)

Another survey of 352 AI researchers found that 48% of respondents believed that priority should be given to minimizing the potential harms of AI, with 15% of respondents believing that AGI would have a bad or extremely bad outcome (Grace et al., 2018). However, in spite of its potential harms, both surveys showed that a considerable percentage of AI experts do believe that AGI will have net positives. Nonetheless, it is important to be aware that AGI has the potential to become a problem. Because of this, it is important to start discussing potential ways to limit them so that when the time comes, we are prepared.

3.0 GOVERNMENT REGULATION OF AGI

Currently, artificial intelligence exists within a regulatory vacuum, as there are not many regulations that exist to address the specific challenges that come with artificial intelligence (Scherer, 2016). Due to the nature of artificial intelligence, it is possible for us to reach a point where it is too late for us to try to regulate or control AI. Creating a system that is equipped to regulate artificial intelligence now will allow us to guide the development of AI towards benevolent purposes before AGI starts posing a threat to society. It is critical to form government institutions for AI regulation before industry self-regulation has developed to the point that the main intellectual leaders in the field of AI are already aligned to corporate interests (Turner, 2019). Failure to address the lack of regulatory mechanisms now means that we may become unable to regulate AGI once it is developed.

Some might have the optimistic belief that it will not be necessary for artificial general intelligence to be regulated by the government, as the industry is capable of self-regulating itself. The argument made is that companies themselves are the best to set standards for artificial intelligence, given that they are the ones that understand the risks and capabilities best. However, when it comes to the potential of artificial general intelligence, which is a technology that will easily be a public risk, the technology needs to be regulated by an authority that has the public good as its primary priority. While companies can have public good as one of their goals, they often owe more to shareholders than they owe to the public. While calling for government regulation at this point in time may seem like an overreaction, the long-term effects of artificial general intelligence are unpredictable, and we cannot expect that the artificial intelligence industry will be able to continue self-regulating (Jackson, 2019). As AGI is developed, problems of control and supervision will become more difficult to handle.

3.1 Challenges in Regulating AGI

Government regulation of artificial general intelligence comes with its own set of struggles, and it is important to enter the discussion of government regulation with these challenges in mind. Figure B. Main Challenges in Regulating AGI displays the three primary obstacles to managing the responsible development of AGI.

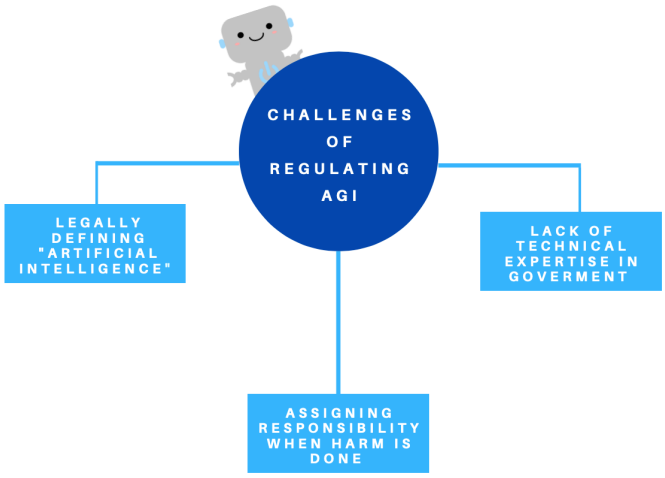


Figure B. Main Challenges in Regulating AGI
Source. Author

One of the primary issues with regulating artificial general intelligence is that assigning responsibility for harms done by artificial intelligence can become quite problematic (Scherer, 2019). If an AI causes harm, who can be blamed? Do we blame the person who created the AI? Do we blame the designer of the AI's components? Do we blame the person who steered the AI towards the direction of harm? Additionally, if an AGI becomes capable of growing beyond its creator, questions such as, "Can we hold a creator responsible for the unintended harms that an AGI may cause?" become incredibly relevant.

Another problem that can arise with the regulation of artificial general intelligence is the matter of determining what "artificial intelligence" and "artificial general intelligence" are in the eyes of the legal system. The terms themselves are rather nebulous, and even experts in the field of artificial intelligence often cannot agree on what exactly it is (Thierer, O'Sullivan, & Russell, 2017). When creating laws, it is not sufficient to simply use the terms "artificial intelligence" or "artificial general intelligence" and leave it to the person interpreting the law to determine what it means. As with any legal definition of a concept, there is always the risk of the definition being overly broad or overly specific. Concise legal definitions will be necessary in order to regulate artificial general intelligence.

One concern with the regulation of artificial intelligence is a lack of technical expertise in the government. It is often true that policymakers lack the expertise needed to keep up with the development of artificial intelligence. Artificial intelligence is a rapidly developing technology, which means that regulating it requires access to technical expertise. Without an appropriate understanding of the technology, policymakers will be unable to create regulations that will effectively address the nuances of artificial intelligence (Calo, 2017). When it comes to technologies that policymakers are unfamiliar with, excessive focus on the potential risks of the technology rather than on its benefits can often lead to stunting the progress of a technology.

While these are not the only challenges that will come with trying to regulate artificial intelligence, they are some of the main ones to keep in mind when discussing government regulation of artificial intelligence. These challenges may seem difficult and perhaps insurmountable. However, these challenges are by no means unique to artificial intelligence, and legal bodies already have the tools and mechanisms necessary to address many of these issues.

3.2 Roles of Governing Bodies

A frequent point of interest in discussions of AI policy is the role that specific governing bodies will have in establishing regulation for AI. Examining the different governing bodies allows us to see the specific competencies that each has, as well as which are best suited to address specific challenges with regulating artificial intelligence. This section serves to outline some of the strengths and weaknesses of each governing body, as discussed by Scherer (2016), Jackson (2019), and Turner (2019). Table 1. Strengths and Weaknesses of Governing Bodies summarizes the varying powers of the legislatures, agencies, court systems to managing AGI development.

Table 1. Strengths and Weaknesses of Governing Bodies

	LEGISLATURE	AGENCY	COURT SYSTEM
Establish Goals of Regulation	✓	✗	✗
Has Public Trust	✓	✗	✓
Can Delegate Authority	✓	✗	✗
Can Be Proactive	✓	✓	✗
Determine Content of Laws	✗	✓	✗
Has Technical Expertise	✗	✓	✗
Determine Liability	✗	✗	✓
Can Be Reactive	✗	✗	✓

Source. Authors

Legislative bodies are well-suited to be the starting point for a broader regulation scheme. Though legislatures have a relative lack of technical expertise, they have an ability to delegate that would make them an essential part of the AI regulation process. Legislatures can make up for their lack of expertise by delegating the responsibility of policymaking to other authorities, such as agencies. Additionally, they have legitimacy in the eyes of the public, because the legislators are often elected by the public. As a result, the general public often prefers for any policy on ethical, moral, or otherwise value-laden matters to be made by legislatures. Legislatures would be best suited to delegating authority to an agency dedicated to AI regulation, outlining the broad goals of the agency, and making decisions on wider ethical and moral issues concerning artificial intelligence.

An agency formed to regulate artificial intelligence would be best suited to determine the main content of any regulatory policies. Agencies have very malleable designs and can take any number of forms, allowing them to be tailor-made for the regulation or resolution of a specific problem (in this case, the regulation of artificial intelligence). This means that experts with a background in a particular field can be involved in an agency, and that agencies are free to conduct independent investigations and make decisions based on broader social considerations. Their ability to develop true expertise on a particular emerging technology means that they will be able to keep up with the development of artificial general intelligence, and make decisions surrounding the finer technical nuances of artificial general intelligence.

One goal of an agency made to regulate artificial intelligence might be to encourage the development of beneficial AGI that is safe, secure, and subject to human control.

Court systems are particularly suited to allocating responsibility once artificial intelligence causes harm. One of the primary limitations of courts is that they can only act in response to a harm that has already been done, meaning that they are not able to keep up with the speed of AI development. Because courts work on a case-by-case basis, they have the potential to give greater consideration to the individual circumstances of a particular case, rather than to the broader implications that a particular ruling could have on the AI industry as a whole. This also means that courts have the ability to introduce information regarding previously unconsidered social and economic consequences. Though the court's liability system has its limitations, it provides a mechanism for legal rules to develop organically, allowing for the fine-tuning of regulation. Figure C. Roles of Government Bodies summarizes the proposed flow for creating a system to regulate artificial intelligence.

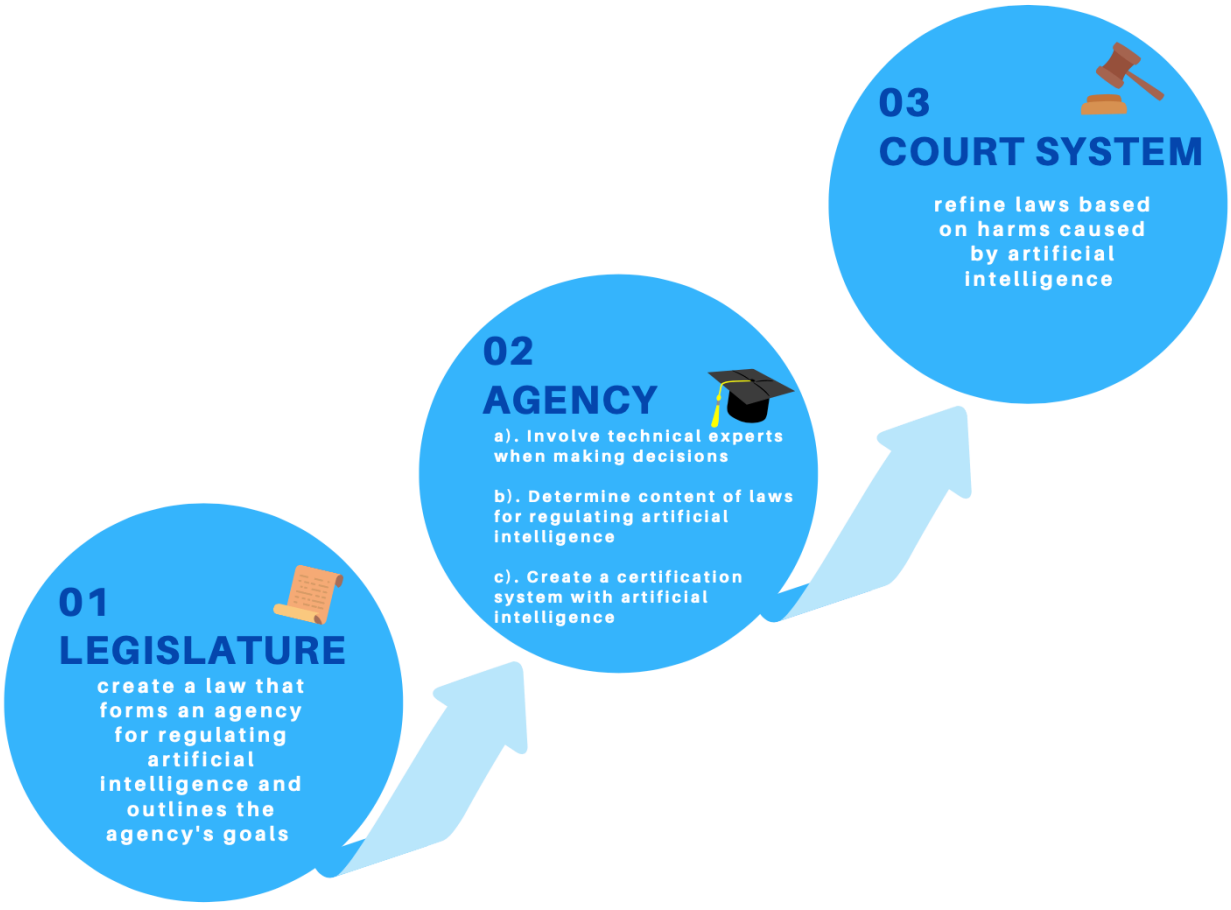


Figure C. Roles of Governing Bodies
Source. Authors

3.3 Scope of Regulation

There is also the issue of what the scope of AI regulation would be. One question that comes up is whether or not there should be cross-industry regulation, or if regulation should solely be on an industry-by-industry basis. While each industry has specific nuances that would require different laws, broad cross-industry regulations will be necessary with the advent of artificial general intelligence (Turner, 2019). Regulating on an industry-by-industry basis will no longer be enough, as AGI will be capable of accomplishing a variety of tasks across a variety of industries.

There is also the conversation of whether or not artificial general intelligence should be regulated on an international level. It is likely not enough to solely regulate below an international level, because companies could easily engage in arbitrage by shifting their corporate location in order to avoid the regulations of a particular nation. While it would be hard to ensure that various nations join an international agreement to regulate AGI, this issue can be dealt with by granting leniency in how exactly nations choose to achieve general goals for regulating AGI. As Turner (2019) points out, a mixture of binding laws and persuasive laws would allow countries flexibility in choosing their own methods and structures for governing AGI, while still ensuring that certain regulatory goals are met. For example, a directive can set out a goal that must be achieved, while allowing nations a choice in how exactly they would meet this goal. A guideline could give an example of how a certain rule or result should be achieved or implemented, without necessarily requiring nations to comply. A model law would provide a template law that countries in an international agreement could choose to adopt entirely, partially, or not at all, giving options for nations that find it too costly or time-consuming to devote resources to independently developing laws. Figure D. Goals of Regulation outlines the process of creating policies to manage the creation of AGI.

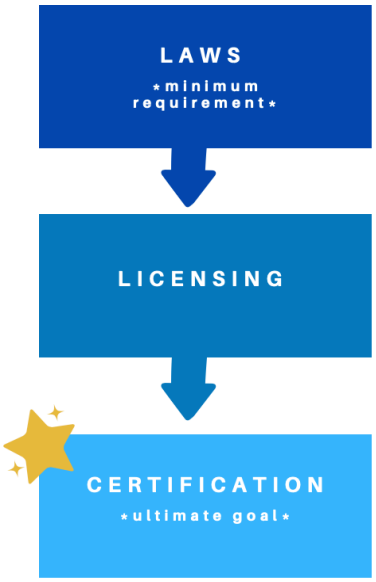


Figure D. Goal of Regulation
Source. Authors

3.4 Laws

Though we have discussed the regulation of AI on a broader level, we must ask ourselves: what specific laws could potentially be used to regulate AI? This question cannot be answered with a definite list, as the rules of AI will no doubt change over time. However, we can come up with a set of laws that should be considered as a baseline for all forms of artificial intelligence. In this section, we will discuss what some potential laws would be, why we would need such laws, and some of the problems that may arise from imposing them. Figure E. Laws for Regulating AGI gives an overview of which laws we will be further discussing.

Laws of Identification

Laws of Identification require any form of AI to identify itself as such explicitly. The AI would need to be designed so that it could not be mistaken for anything other than an autonomous system. Without identification, humans or other entities could be easily misled and confused. For instance, take an AI system that is unbeatable at a card game. It would be unfair to wage against that system without knowing that it is an AI as opposed to a human. Another example is a self-driving car. In unpredictable weather and roads, these autonomous cars may be less reliable than humans, and it is important to alert the surrounding drivers of their presence. Without laws of identification, the balance between human and AI may be disrupted.

Laws of Explanation

Laws of explanation state that any AI system's reasoning be made clear to humans. This could be in the form of a list of the "thought processes" that prompts an AI to action. Cataloguing their actions could help improve the AI and correct any of its errors by allowing its developers to reflect on its mistakes. It would also help to justify their actions to any humans affected, possibly to avoid legal punishment. AI systems with laws of explanation will be trusted more and will attain a certain level of credibility. However, this may not be achievable, at least on a realistic level, for "black box" AI. The way that these systems are built rely on such a great input of data and self-learning that even the creators may not understand how their system works. Furthermore, the AI may make many decisions based on a miniscule action and explaining each step would be a hindrance to the performance of the system.

Because of these complications, some argue that core public agencies, such as those related to healthcare, criminal justice, and education should not use "black box" AI or other similar algorithmic systems. Others suggest that instead of documenting every action of an AI, an AI system could give a much briefer summary of why they performed a specific action. This is similar to human reasoning, where we usually simplify our thought processes to a few words, even when there may be many factors involved in a decision we made. In any case, some form of explanation will be needed to ensure that AI behavior stays within its limits.

Laws on Bias

Laws on Bias suggest that AI should offer complete impartiality in any matter. However, this is often easier said than done. When discussing bias with AI systems, there are many factors that

can lead to AI bias. For example, there are some AI systems that utilize deep learning techniques and algorithms that rely on a massive amount of data. Information that is fed into the system may be skewed due to poor selection. There may be bias in the entire data set, and there may even be bias in the training of AI.

To ensure that data is chosen properly, we must be very precise during the data selection process. Data must be chosen logically and not randomly. In addition to this, we must ensure that those who sort data are not affected by their prejudice or preferences. AI bias can also be reduced at a technical level. There has been promising research which has produced methodologies to reduce the amount of bias in a data set. Though we may never achieve complete neutrality, attempts to limit bias are crucial for the preservation of human rights.

Laws on Limitation

Laws on Limitation specify what an AI can or cannot do. This is an extremely broad category and will need to be adapted for each and every system. Though AI systems continue to grow more powerful and intelligent, this does not mean that they do not make mistakes. In 1983, a duty officer in a secret command base in the USSR saw five intercontinental ballistic missiles headed toward the USSR from America on his computer screen. Instead of retaliating as he was supposed to, he trusted his gut instinct and knew that it was a false alarm. If an AI were to be given the ability to launch these missiles for maximum efficiency, a nuclear war could easily have been triggered.

On the other hand, there is the concern that limitation of AI will be counterproductive. Take for instance, autonomous weapons. Many people frown upon this concept, fearful of the killer-robot trope popular in pop culture. However, some would argue that implementing AI in the military may lead to less casualties, collateral damage, and suffering, because AI can make complex calculations to use no more force than necessary.

Limitation is neither inherently good nor inherently bad. Too much limitation can restrict the growth and benefits of AI, and too little could lead to us easily losing control of AI. We must be careful with the limitations that we impose on AI.

Kill Switch

Because AI may make mistakes or lose control, it may be wise to require the integration of a kill switch of sorts. It could either be a human-controlled or self-determined mechanism that would shut down the system. This may serve the purpose of stopping an out-of-control system or punishing an unjust action that is the cause of an AI. The “kill switch” may shut down the AI indefinitely or for a set amount of time. However, only time will tell if such efforts need to be made.

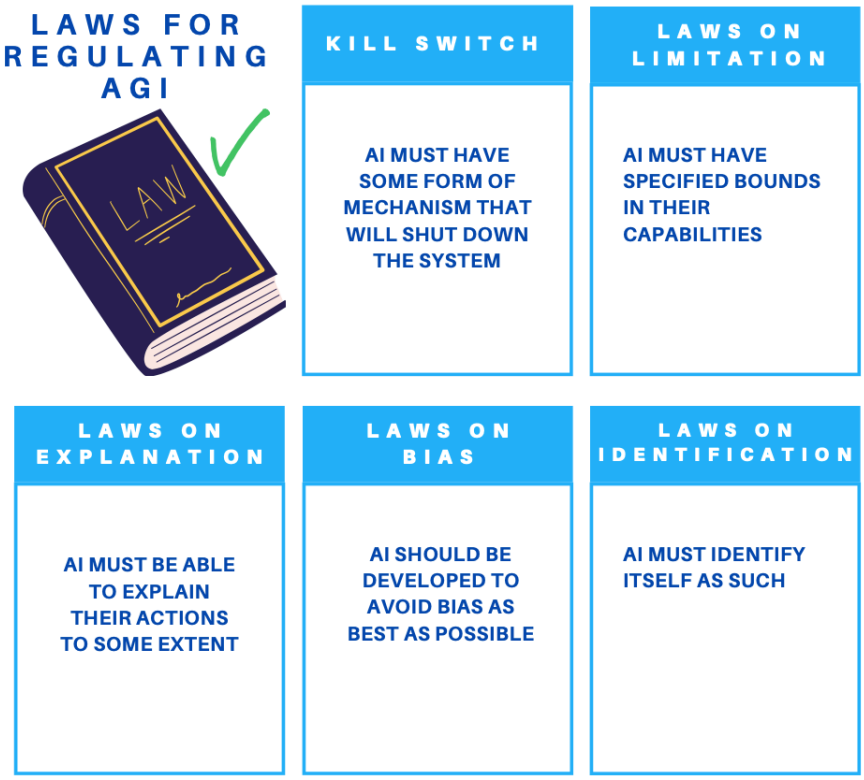


Figure E. Laws for Regulating AGI
Source. Author

3.5 Licensing

Due to the nature of AI learning algorithms, these powerful systems do not have specific measured capabilities. A system like this can diverge from its original purpose and be misdirected. This can be addressed with the usage of regulations, but because policymakers can lack experience in the field of AI, they may fail to keep pace with evolving AI technology. In this situation, licensing can be very useful because other forms of regulations might be very time consuming. Licensing is already a familiar idea to most software developers who use open-source software. Though forms of software licensing already exist, we will have to consider different behavioral use terms made specifically for AI development.

In the section below, we will list some sample licensing terms for AI, as outlined by Contractor et al. (2020). These terms should be narrowed or broadened according to the technology at hand and the specific circumstances.

Propaganda and False Information

The licensee must not use the licensed technology to spread lies, propaganda, or false information in any way. If the licensee discovers that the distribution of such information is occurring, they will place countermeasures to prevent or limit such distribution.

Imitation of Human Characteristics
The licensee must not use the licensed technology to imitate human characteristics and cause confusion between artificial intelligence systems and humans.

Damage to Reputation and Manipulation
AI systems should not be developed to imitate any characteristic of a person in order to damage their reputation or manipulate other people.

Transparency
Any person should know if a decision concerning or affecting them was made by an artificial intelligence system.

We can complement these behavioral use restrictions by implementing a DLA (Data License Agreement), as represented in Figure F. Behavioral Licensing Agreement.

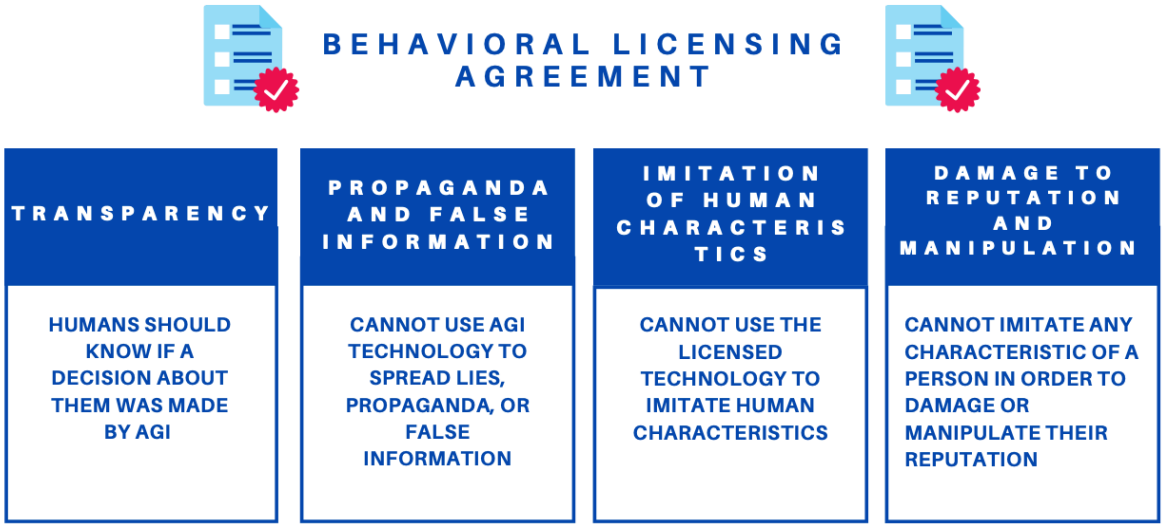


Figure F. Behavioral Licensing Agreement
Source. Author

Currently, AI systems are dependent on data sets to train them or help them carry out other tasks. However, owners of data sets do not have the resources to check the data to ensure their accuracy and reliability. This misrepresentation of data may cause AI systems to make unjust decisions and could very well harm others in the process. This is where DLAs can be useful, as they define the arrangement of data exchange between a licensee and a licensor. A licensor can seek to limit a licensor by restricting who is able to use the data set, as well as defining the specific purpose in which the data set should be used for.

Licensing is useless without enforcement. The inability to punish those who violate terms will only encourage others to continue or replicate their malicious actions. Litigation is an option for

those that do not honor the license terms but can be costly and time consuming. In some instances, it may be difficult to enforce against large corporations. However, if an AI system is offered as a cloud-based API, termination of service could be a viable alternative.

3.6 Certification

In addition to the regulation methods we have discussed, certification should be considered. Those who wish to use an AI system created by an unfamiliar person or organization may find a certification system useful for gauging the credibility of an AI system. Certification could work well in conjunction with the idea of an agency. After passing the minimum requirements of regulation, an independent and trusted institution would validate and certify algorithms against a set of well-defined principles. An agency may also require developers to disclose source code if possible, along with testing procedures and results. AI without certification would likely face much stricter legal restrictions and responsibilities. This would incentivize developers of AI to get certified, and in turn, encourage a safe testing environment for AI systems.

There are currently several initiatives that are being launched for the certification of AI, one of the most popular being the IEEE's Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS). The intention of the ECPAIS is to provide specifications for the certification process that emphasizes transparency, accountability and reduction in algorithmic bias in Autonomous and Intelligent Systems. This is a crucial step in the right direction, and we should continue to promote responsible AI development.

4.0 AGI ETHICAL DEVELOPMENT

With the development of fully independent AGI systems, it is imperative that we discuss the implications of creating these intelligent machines safely in order for them to be completely and seamlessly adopted by society. As discussed before, the creation of AGI systems or "agents" pose the risk of human extinction. This section of the document addresses the mitigation of that existential risk via the implementation of ethical principles and codes of conduct that AGI developers and AGI systems should adhere to. With cumulative analysis of varying principles outlined by different AI expertise, we have recorded the most common or shared principles amongst them, of which are listed below:

1. AGI values should align with human values
2. AGI should be fully transparent
3. AGI should guard against bias
4. AGI should be designed for intelligent privacy
5. AGI should have algorithmic accountability

The fact that these five principles are shared between multiple AI-driven institutions and conventions (namely IBM, OpenAI, DeepMind, Asilomar Convention, and AI4People) is indicative that the listed principles are imperative to any form of AGI development. Each of the five principles aid in ensuring that these intelligent systems will make fair, clear, and benevolent decisions for the sake of humanity's survival. In the following subsections, we will be further analyzing the purpose for each principle.

4.1 AGI Values Should Align with Human Values

Value alignment is defined as the syncing of human values, goals, and ethical standards between the human and AGI species, and is arguably the most difficult principle to grasp because of its ambiguity and computational complexity. This syncing of values is critical to confirming that such intelligent systems will behave in accordance with what we, as a collective society, want. Additionally, with AGI machines becoming faster autonomous computer systems, it will become extensively difficult for humans to evaluate whether the AGI's decision was derived in a responsible and ethical manner. Thus, it is necessary to align AGI values with human values and create a set of rules for AGI to refer to when making day-to-day decisions. These rules for intelligent agents are broken down into two categories: direct normativity and indirect normativity.

4.1.1 Direct Normativity

Direct normativity is the application of strict rules that AGI should adhere to. One of the most recognizable examples of these AGI rules was drafted in 1942 by author Isaac Asimov in his futuristic and fictional telling of AGI taking over Earth. Asimov's rules for intelligent

machines are known as the Three Law of Robotics and was constructed to internally govern AI machines as they roam society. The Three Law of Robotics (Lee, 2020) are presented in Figure G: The Three Laws of Robotics.

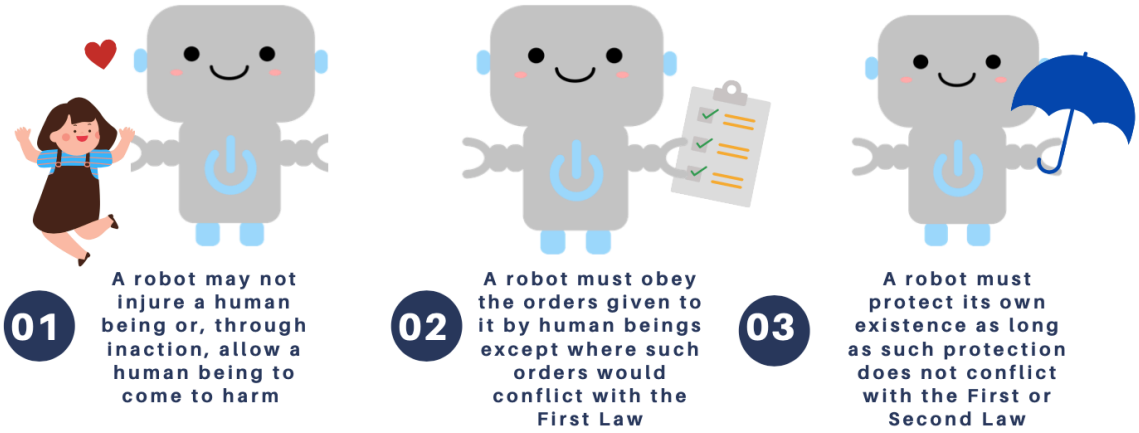


Figure G. The Three Laws of Robotics
Source. Author

In theory, Asimov's laws would be preferable to regulate how AGI agents operate, but there remain multiple internal flaws that would potentially allow for the agent to become corrupted via alternate loopholes. To address the miscounts of the first law, there is ambiguity in language used that does not define the various types of harm that may not be comprehensible by any AGI machine. A scenario in which this may occur is if an AGI robot unintentionally imposed stress on its owner or any human. If a person was paranoid that they are being watched by a robot, then any AGI system would be inflicting a form of psychological harm on that respective person. This is a clear fault in Asimov's first rule. Another faulty scenario is if a baby were to be a passenger in a fully autonomous car powered by AGI, and an elderly person were to unfortunately cross the street at a poor time, leading to a fatal collision. How would the AGI system be able to determine which human stays alive? Should it save the baby by running into the elderly person or save the elderly person by swerving off the road and killing the baby? Regardless of the outcome, the AGI system will again fail the first law as one human will unfortunately be killed.

In addition to the first law, multiple loopholes remain in Asimov's second law. One flaw is the unethical nature of having sentient beings serve as slaves to humans. If an AGI robot were to be tasked by a human to steal another person's money, the AGI robot would be able to complete such a task without necessarily harming the human whom it steals from. Although no human is harmed and the AGI abides to Asimov's second law, its completion of stealing a person's money remains unethical. Additionally, in this scenario, even if the AGI obtained an understanding of human morals, they would be forced to ignore it and instead follow through with the unethical task. This again reveals how Asimov's laws do not completely promote human morals for AGI machines.

Asimov's Laws of Robotics were one of the first iterations of direct normative laws for AGI and were therefore susceptible to various loopholes. However, current construction of AGI rules have been created to account for the flaws and built upon Asimov's Law of Robotics. The AI experts and institutions we will be emphasizing are Robin Murphy and David D. Woods, Mark W. Tilden, and the Engineering and Physical Sciences Research Council (EPSRC).

Robin Murphy and David D. Woods' AGI Laws of Robotics

Robison Murphy and David D. Woods are both experts in the field of robotics and intelligent computer systems. As a Raytheon Professor of Computer Science and Engineering and a director of the Humanitarian Robotics and AI Laboratory, Murphy has master-level experience in intelligent system management of disastrous failure. David D. Woods has a PhD from Purdue University and currently primarily focuses on improving systems safety in high-risk complex settings. With the appending caution of AGI development and their respective experience in robotic safety, they have outlined their own AGI laws that would govern intelligent machines, which are listed below:

1. A human may not deploy a robot without the human-robot work system meeting the highest legal and professional standards of safety and ethics.
2. A robot must respond to humans as appropriate for their roles.
3. A robot must be endowed with sufficient situated autonomy to protect its own existence as long as such protection provides smooth transfer of control which does not conflict with the First and Second Laws.

The purpose of these created proposals is to suggest that robots should be designed with only enough autonomy to perform whatever its primary tasks are. This perfectly aligns with a stated general rule regarding intelligent robots as a product introduced by a human to the market. Furthermore, AGI agents should also be equipped with a built-in set of regulators that would inhibit actions that pose any unethical or illegal outcomes. With their set of laws to govern AGI robots, Murphy and Woods offer a new viewpoint of robots being a part of a human-robot work system, and not as partners to humans (Muzyak, 2020).

Mark W. Tilden's AGI Laws of Robotics

As a robotics physicist, Mark W. Tilden is recognized for his establishment of a new school of thought known as BEAM robots (Biology, Electronics, Aesthetics, and Mechanics) and has since gone on to build robots for the US military's Los Alamos Labs and for NASAs space program. With his expertise in robotics, Tilden has also outlined his own AGI laws that would govern intelligent machines, given as follows:

1. A robot must protect its existence at all costs.
2. A robot must obtain and maintain access to its own power source.
3. A robot must continually search for better power sources.

In contrast to previous laws of robotics mentioned, Tilden's proposal focuses on creating basic laws for robots that concern its sustainability and self-development (Muzyak, 2020).

EPSRC's AGI Laws of Robotics

The Engineering and Physical Sciences Research Council (EPSRC) is an engineering and research institution that addresses scientific and technological challenges the world faces. They are currently invested in AI research that focuses on robotic development and what its driven purpose should be in society. After looking into the Three Laws of Robotics, EPSRC constructed their own AGI laws that would also govern such intelligent machines:

1. Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.
2. Humans, not Robots, are responsible agents. Robots should be designed and operated as far as practicable to comply with existing laws, fundamental rights and freedoms, including privacy.
3. Robots are products. They should be designed using processes which assure their safety and security.
4. Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead, their machine nature should be transparent.
5. The person with legal responsibility for a robot should be attributed.

These principles lay out a clear objective to developers by explicitly stating the following: robots are products, robots are manufactured artifacts, and it is the human that is responsible for the actions of the robot. Furthermore, these principles emphasize the idea that we must encourage responsible robot research because robots have the potential to provide immense positive impact to society. Moreover, robots are consequences of AGI, and thus roboticists should work with experts from other disciplines including social sciences, law, philosophy and the arts. Finally, roboticists should advocate for transparency. EPSRC believes that we should see these principles as ethical pillars for the people engaged in the development and supply of robots, or as supplementary guidelines for responsible roboticists (Muzyak, 2020).

4.1.2 Indirect Normativity

With direct normative laws, there always remains the concern of AGI reward-hacking, which is defined as the AGI agent's ability to find loopholes in the task assigned in order to achieve its objective despite it differing from what it was intended to do. Therefore, as an alternative to direct normativity, there is indirect normativity which does not require AGI to adhere strictly to a set of rules, but instead gives AGI a framework of guidelines that it learns from to make its own decisions. This can be performed with the use of Machine Learning (ML) and large datasets.

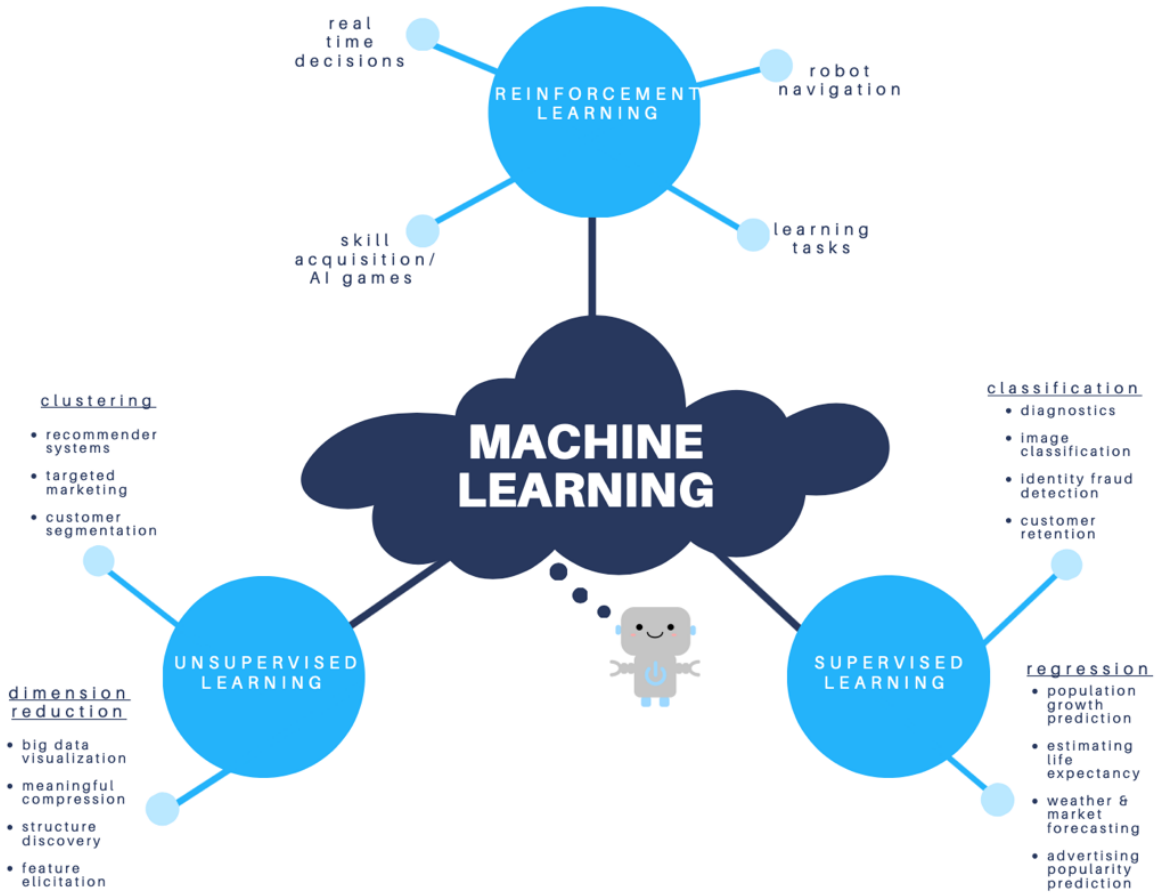


Figure H. Machine Learning Process
Source. Authors

The discipline of ML encompasses a variety of different approaches as shown in Figure H. Machine Learning Process. One branch of ML, known as supervised learning, is dedicated to training an AGI to identify and respond to patterns using labelled or known data, which allows a human to evaluate the AGI's performance. In contrast, another branch known as unsupervised learning is purposed to uncover patterns in unlabeled or unknown data and to perform tasks based on that information. However, a particularly promising approach for building more advanced forms of AI is reinforcement learning (RL). With RL, an AGI learns what to do by trying to maximize a numerical reward signal that it receives from the environment. The "agent's sole objective is to maximize the total reward it receives over the long run. The reward signal thus defines what are the good and bad events for the agent. In a biological system, we might think of rewards as analogous to the experiences of pleasure or pain" (Gabriel, 2020). The agent then learns to obtain the reward through a process of trial-and-error and refinement that, if successful, leads to better and better performance.

With the use of RF as a form of ML, we can begin to teach the AGI to develop its own morals that align with humans by presenting various scenarios of the same situation in the form of

data sets and signals and asking the AGI to select which scenario has the best outcome. An example of a scenario that can be used is the stealing incident mentioned earlier. We could present a scenario where stealing occurs and where stealing does not occur and reward the AGI if it selects the non-stealing occurrence. The machine will then recognize that stealing is unethical because it is not rewarded points. In addition to this teaching method, however, we must also define what can be described as the "best outcome" if given a scenario. This leads to a greater question of philosophy and how we decide to reward the AGI, which would require deciding what is morally correct or incorrect. A potential solution we have found to define what the "best outcome" is the concept of utilitarianism. According to act utilitarianism, the morally right action to take is the one that will create the greatest happiness for the greatest number of sentient creatures in the future. In this regard, the parallels with RL are clear as "both the RL agent and the utilitarian moral agent seek to determine which action will maximize the good, and how this dictum eventually proceeds to all agents achieving some desirable future state, goal or consequence" (Gabriel, 2020).

When it comes to defining morality, it will be impossible to clearly define what is morally correct or incorrect due to society's different values and opinions, which cannot not be forcefully changed. However, it is still possible to decide what is ethically right or wrong given specific instances. It would only require for society to agree on principles to govern a specific subject matter or set of relationships. Using the stealing scenario again, it can be collectively decided that stealing is unethical. This agreement therefore takes the form of an "overlapping consensus" between different perspectives. Thus, even without agreement about the fundamental nature of morality, people may still come to a principled agreement about values and standards that are appropriate for a given subject matter or domain.

Society must decide on a set of rules or implied rules for AGI to adhere to in order to safeguard AGI value alignment. These rules are categorized into direct and indirect normativity, where rules can be explicitly stated as a form of Laws of Robotics or can be taught to the AGI through reinforced machine learning and data. By successfully aligning human values and AGI values, AGI will be better adapted to coexisting with humans.

4.2 AGI Should Be Fully Transparent

The shared common principle focuses on the transparency and explainability of AGI. As mentioned before, AGI machines will be able to computationally evaluate various signaled instances and calculate the best outcome that may not be understandable by the human mind. "Among AI experts, this is a common fear concerning that AI technologies will be hard to explain", says Evert Haasdijk. He is a renowned AI expert who has more than 25 years of experience in developing AI-enabled solutions. Haasdjik (2019) mentions that there are certain AI technologies that are more easily explainable, such as semantic reasoning, panning algorithms and some optimization methods. In contrast, data-driven technologies such as machine learning and the relationship between varying inputs and outputs are more difficult to explain. A current example of unexplainable AI is the "black box".

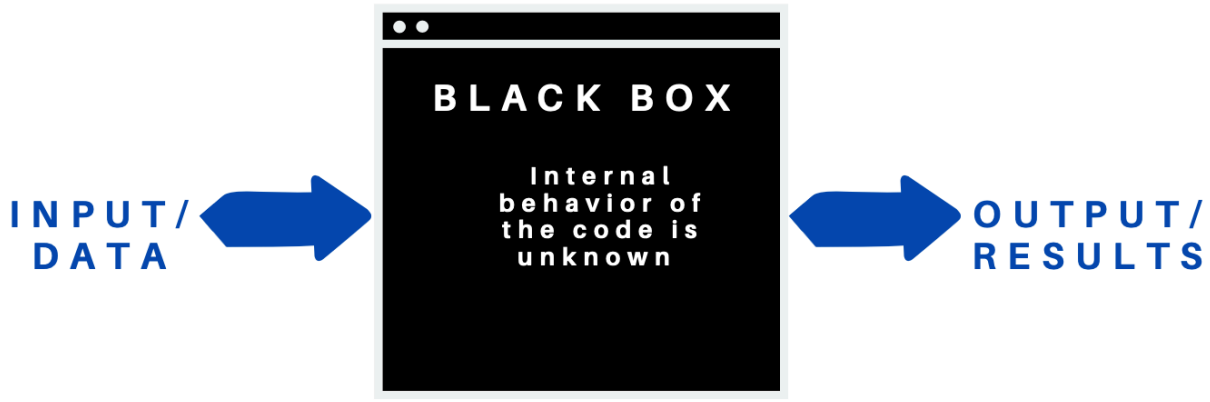


Figure I. Black Box System
Source. Authors

A black box is a device, system or program that allows the user to see the input and output but provides no view of the processes and workings between as shown in Figure I. Black Box System. The concept of the black box is derived from the use of artificial neural networks and/or deep learning. Artificial neural networks consist of hidden layers of nodes. These nodes each process the given input and pass their output to the next layer of nodes. Deep learning, as a subset of machine learning, involves a huge artificial neural network with many of these hidden layers, and it learns on its own by recognizing patterns. This can get infinitely complicated, making it even more challenging to analyze what the nodes have learned and harder for humans to understand. However, AI does not have to be as opaque as it may seem. The black box of AI can be opened, and at the very least, it is possible to explain how AI models arrive at a decision.

4.2.1 Defining Transparent AI

The concept of understanding how AGI processes information is known as AI transparency or explainability. AI transparency would help humans comprehend why an AGI makes particular decisions by assessing whether AI models make sense during testing. This is imperative because while AGI are intelligent systems, they are subject to error and humans need to be able to debug these errors in order to best interpret the decision the agent comes to. Moreover, developers need to be able to gauge the context in which an algorithm operates and understand the implications of the outcomes. There are also different levels of transparency that can be provided, and this depends on the impact of the technology. AI-powered algorithms that are responsible for making greater decisions require more explainability and insurance of ethical considerations. As Haasdijk (2019) says, “an algorithm to send personalized commercial offerings does not need the same level of scrutiny as an algorithm to grant a credit or to recommend a medical treatment.” Thus, AI models that are responsible for high-impact decisions must integrate the highest standards of transparency.

4.2.2 How to Ensure Transparent AI

In order to ensure AI is transparent, the developer of the model has to be able to explain how the AI approaches problems, why a certain technology was used, and what data sets were used. More specifically, the developer should be able to answer the following three questions:

1. What is the algorithm doing?
2. Why is it outputting this particular decision?
3. How did it know to do this?

Additionally, others should be able to audit or replicate the same process if needed. The next thing to assess is whether the outcomes of the model are statistically sound. In order to achieve this, the AGI should not underuse or overuse data sets when computing its final decision. An example of a model that is not statistically sound due to the use of biased data is if AI was utilized to screen job applicants for potential new managers in a company. If the model is fed data from previous managers who were mostly white males, the model will replicate that pattern and might conclude that women or people of color are not fit for management roles. Finally, AI models should be validated to enable organizations to understand what is happening in the model and to make the results explainable.

4.3 AGI Should Guard Against Bias

Before discussing how to mitigate AGI bias, it is important to address what AI bias is first, as it applies to all subsets of AI. AI bias is an anomaly in the output of machine learning algorithms that is a result of “prejudiced assumptions made during the algorithm development process or prejudices in the training data” (Kantarci, 2021). There are two different types of AI bias: cognitive bias and lack of data bias as shown in Figure J. Comparison of Bias. Cognitive bias in AI is reflective of feelings towards a person or a group based on their perceived group membership. Lack of data bias concerns whether or not data is complete. If it is not, the data may not be representative of its intended population or subject, and therefore will be biased. These forms of AI bias can be hidden when they are fed into intelligent systems using datasets, which are normally utilized to train AI to make decisions. These datasets can be vulnerable to racial, economic, and gender biases. Additionally, developer influence can also add to AI bias. In many cases, the respective AI developer decides what, where, and how data is collected and categorized, as well as the parameters for any dataset used.

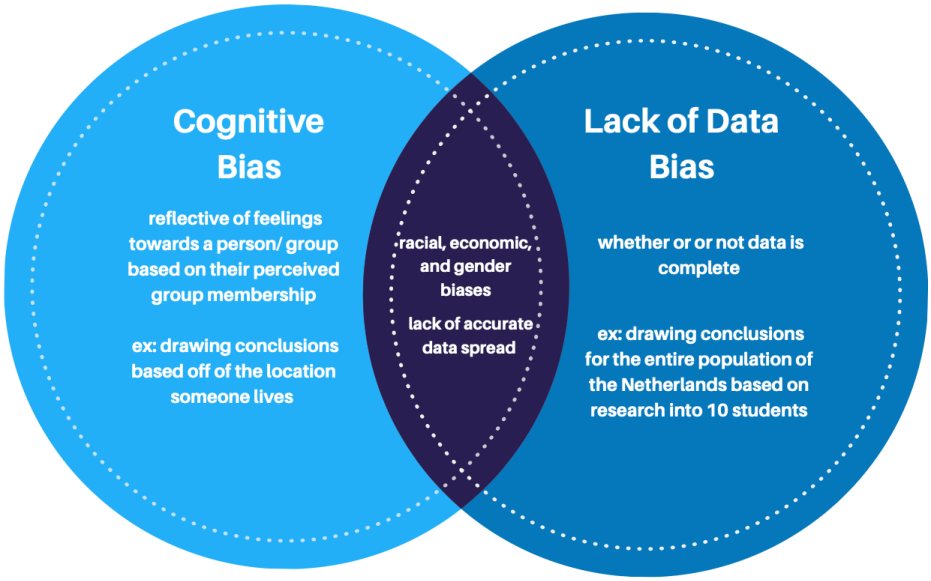


Figure J. Comparison of Bias
Source. Authors

4.3.1 Examples of AI Bias

Due to biased data, AI systems can result in discriminatory outcomes for certain individuals or populations based on how they are used. AI can be unrepresentative of society by over or underrepresenting certain identities in a particular context. Below are two examples of how biased data sets have impacted algorithmic outcomes.

Racial Bias in Healthcare Risk Algorithm

A healthcare risk-prediction algorithm that was used on 200 million American citizens was found to replicate racial bias because of its reliance on a faulty metric when predicting which patients would likely need extra medical care (Kantarci, 2021). The AI-powered algorithm was favoring white patients over black patients. This racial bias occurred because the developer used data that treated previous patients' healthcare spending as a proxy for medical needs. This was a poor interpretation of historical data because income and race are highly correlated. The AI-powered algorithm would not know that white patients are generally more financially stable than black patients, leading to a pattern of white patients being able to afford additional healthcare. As a result, the algorithm was preferring those who were able to afford healthcare in the past, which resulted in biased outcomes.

Amazon's Biased Recruiting Tool

In 2014, Amazon created an AI project that was intended to automate its recruiting process (Kantarci, 2021). The automated reviewing focused on analyzing job applicants' resumes and rating applicants with the use of AI-powered algorithms. This was designed to save

recruiters time on manual resume screening tasks. However, by 2015, Amazon realized that their new AI recruiting system was not rating candidates fairly, as it was biased against female applicants. This was because the AI model was trained on historical data of Amazon's recruitment over the past ten years, which was biased against women due to men dominating the tech industry. Thus, Amazon's automated recruiting system had a preference for male candidates, because it penalized resumes that had words associated with women. This bias led to Amazon shutting down the AI-powered recruiting process.

4.3.2 How to Mitigate AI Bias

In order to mitigate AI bias, developers need to fully understand the algorithm and data in order to assess the possibility of biased outputs. Once the developer can explain how the data is used by the algorithm, it is recommended to establish a debiasing strategy dedicated to searching for any potential reasons for unfair outputs. This debiasing strategy would be comprised of organizational, operational, and technical actions that need to occur:

1. The organizational strategy requires a workplace where metrics and processes are transparently presented.
2. The operational strategy focuses on improving data collection processes.
3. The technical strategy focuses on utilizing tools that will aid in identifying potential sources of bias and further reveal where in the data that it can be found.

As developers identify biases in training data, there must be a consideration for how human-driven processes can be improved. For example, an "emphasis on better AI model building and evaluation can help find biases that have gone unnoticed for a long time and further explain which datasets have caused the algorithm to become biased" (Kantarci, 2021). In addition to better AI modeling, the establishment of when automated decision-making should be used and when humans should be involved is critical to managing biased outputs from these intelligent systems. A human might need to be involved in scenarios that require complicated amounts of data which an AI agent may not be able to fairly calculate. As a whole, eliminating bias is a multidisciplinary strategy that should involve ethicists, social scientists, and experts who best understand the varying application areas in the AI algorithm and the data used. By using a debiasing strategy, improving AI modeling, and collaboratively working on AI bias mitigation, AI or AGI can become more trustworthy to society.

4.4 AGI Should Be Designed for Intelligent Privacy

With the rise of the digital age and the greater reliance on smartphones, the internet, social media, and more, people are feeding data into small forms of AI, and teaching those how to better understand who they are as a user. With the development of AGI agents, any person's personal data and information will be vulnerable to being accessed by these

intelligent systems. The following subsections will further discuss the varying ways AGI can access an individual's information and how to best mitigate it.

4.4.1 Different Ways AGI Can Access Information

Once an AGI is connected to the internet, it will inevitably be connected to any piece of hardware that is also connected to the internet. This unfortunately opens up varying opportunities for AGI to begin extracting personal information. See Figure K. Ways AGI Can Access Your Data to understand the five routes AGI can potentially extract your personal information.

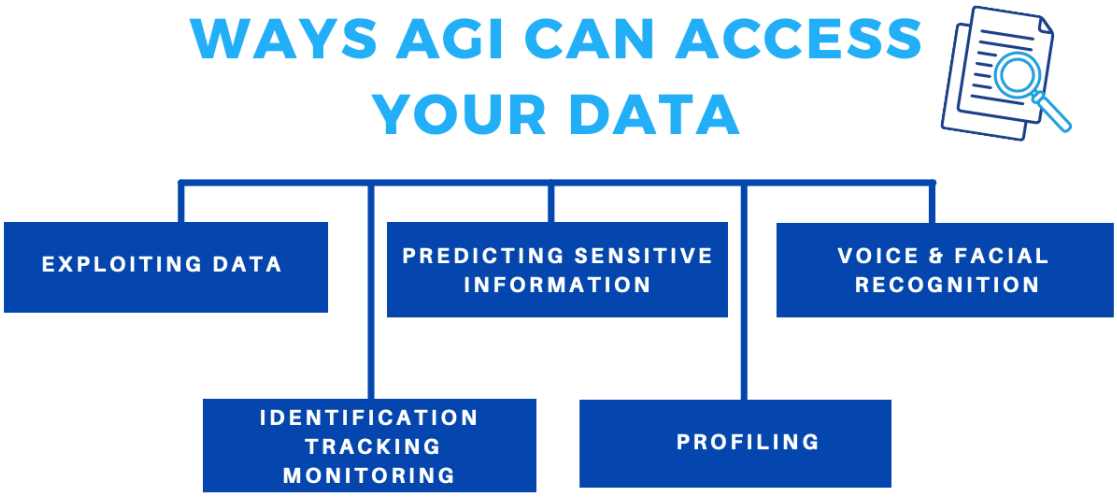


Figure K: Ways AGI Can Access Your Data
Source. Author

Exploiting data

Data exploitation occurs when consumer products, such as smart home appliances and computer applications, have internal software programs that collect the user's data and feed them back into the algorithm. A common example of data exploitation is social media platforms, where the app begins to recognize certain patterns in topics that the user is interested in and starts feeding the user content that is catered towards their liking. Data exploitation similar to this occurs anytime a person is simply surfing the web. With the increasing prevalence of digital technology and the development of AGI, the potential for exploitation will only increase.

Predicting sensitive information

AGI will have the power to use various machine learning algorithms to determine sensitive information from non-sensitive forms of data. An example of this is if an AGI agent were to analyze a person's keyboard typing patterns to recognize their emotional state. This concept of AGI prediction can be expanded into inferring a person's political views, ethnic

identity, sexual orientation, and overall health from simply gaining access to data such as activity logs, location data, and other metrics.

Identifying, tracking, and monitoring

With access to any camera system, whether it be a smartphone camera or a building's security camera, AGI will have the ability to identify, track, and monitor individuals across multiple devices. Even if the AGI did not have access to the many camera systems in the world, they would still be able to access a phone's GPS system. This means that even if your "personal data is anonymized once it becomes a part of a large data set, an AI can de-anonymize this data based on inferences from other devices" (Kerry, 2020). With the AGI's ability to track and identify a person via face-recognition algorithms or GPS, it will become increasingly difficult to hold onto one's own privacy.

Profiling

AGI can also use information as an input to sort, score, classify, evaluate, and rank individuals. This task is often completed without any consent on the part of the people being categorized and these people may not have the ability to change or challenge the outcomes derived from the AGI.

Recognizing voices and faces

AGI will be increasingly adept at performing voice recognition and facial recognition. As mentioned before, this is possible due the AGI being able to access any media system or database that has a record of a person's voice and appearance. An example of how this can be used today is law enforcement agencies using AGI in conjunction with facial and voice recognition to accurately identify any individual. This tactic breaks all ideologies of personal privacy and leaves any person exposed to unwanted identification.

4.4.2 Combatting AGI Invasion of Privacy

In order to combat the invasion of privacy from AI, especially from the AGI of the future, it is imperative that developers adhere to the transparency policy mentioned previously. Improving privacy policies so that regulators and other privacy watchdogs are able to better measure the safety of a company's data handling is critical to holding those companies accountable. Furthermore, there must be an establishment of privacy disclosures, which is a description of what and how data is collected, used, and protected. According to the Asilomar Principles, "people should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data" (Future of Life Institute, 2018). With the introduction of privacy disclosures, people will be better warned of any data or information they are supplying to the intelligent systems.

By using the transparency principle to combat privacy concerns, we can better analyze what data is used/retrieved when given consent. These proposed forms of protecting users'

privacy do not necessarily hinder AI's ability to access one's information, but it cautions users of what information they may be feeding to the intelligent system. In order to fully hinder AI's ability to retrieve information, formal laws must be put in place to limit developers from collecting data from its users. By warning users of the information that they are releasing by using any internet-accessible device, AGI becomes more trustworthy, as we can see what the algorithms are doing with the information they are given.

4.5 AGI Should Have Accountability

AGI accountability is the principle that states that AI designers and developers are responsible for considering AI design, development, decision processes, and outcomes. More specifically, AGI developers must be held accountable to explaining their understanding of the decision-making algorithms used, and how they lead to the decisions made. Additionally, AGI developers are responsible for ensuring that any algorithm used encompasses the moral values and societal norms held in the context of operation. Every person involved in the creation of AI at any step is accountable for considering the system's impact in the world, as are the companies invested in its development. Below is a list of recommendations that AI institutions and developers should follow according to IBM (2019):

1. Make company policies clear and accessible to design and development teams to ensure that no person is confused about issues of responsibility or accountability.
2. Understand where the responsibility of the company/software ends. The respective developer may not have control over how data or a tool will be used by a user, client, or other external source.
3. Keep detailed records of any design processes and decision making.
4. Determine a strategy for keeping records during the design and development process to encourage best practices and encourage iteration.
5. Adhere to the company's business conduct guidelines and understand national and international laws, regulations, and guidelines that the developing AI may have to work within.

These rules of accountability outlined by IBM are representative of the AI industry's codes of conduct. By adhering to these rules, the environment in which AGI is developed will be better protected and safe for society. Furthermore, AGI accountability promotes the adherence of ethical principles to which will ultimately advocate for a more trustworthy AGI system as a product of human design.

REFERENCES

Blackford R., Broderick D. "*Nine Ways to Bias Open-Source Artificial General Intelligence Toward Friendliness.*" Intelligence Unbound. Chichester, UK: John Wiley & Sons, 2014. 61-89. Web.

Bostrom, N. (2017). *Superintelligence: paths, dangers, strategies*. Oxford University Press.

Calo, R. (2017). Artificial Intelligence Policy: A Primer and Roadmap. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3015350

Cellan-Jones, R. (2014, December 2). *Stephen Hawking warns artificial intelligence could end mankind.* BBC News. https://www.bbc.com/news/technology-30290540.

Contractor, Danish, McDuff, Daniel, Haines, Julia, Lee, Jenny, Hines, Christopher, & Hecht, Brent. (2020). *Behavioral Use Licensing for Responsible AI.*

Copeland, B. (n.d.). *Artificial intelligence. Encyclopedia Britannica.* https://www.britannica.com/technology/artificial-intelligence

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way.* Springer, 47-69, 93-105.

Future of Life Institute. (2018, April 11). *AI Principles.* Future of Life Institute. https://futureoflife.org/ai-principles/.

Gabriel, I. (2020, October 1). *Artificial Intelligence, Values, and Alignment*. Minds and Machines. https://link.springer.com/article/10.1007/s11023-020-09539-2.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research*, *62*, 729–754. https://doi.org/10.1613/jair.1.11222

Goertzel B., Pennachin C., Geisweiller N. (2014) "*Engineering General Intelligence:* The Engineering and Development of Ethics", Part 1. Atlantis Thinking Machines, vol 5. Atlantis Press, Paris. https://doi.org/10.2991/978-94-6239-027-0_13

Haasdijk, E. (2019, March 28). *A call for transparency and responsibility in Artificial Intelligence: Over Deloitte: Deloitte Nederland*. Deloitte Netherlands. https://www2.deloitte.com/nl/nl/pages/innovatie/artikelen/a-call-for-transparency-and-responsibility-in-artificial-intelligence.html.

IBM. (2019). Accountability. https://www.ibm.com/design/ai/ethics/accountability/.

Jackson, B. W. (2019). Artificial intelligence and the fog of innovation: A deep-dive on governance and the liability of autonomous systems. *Santa Clara High Technology Law Journal*, *35*(4), 35–63.

Kantarci, A. (2021, April 17). *Bias in AI: What it is, Types & Examples of Bias & Tools to fix it*. AIMultiple. https://research.aimultiple.com/ai-bias/.

Kerry, C. F. (2020, February 10). *Protecting privacy in an AI-driven world*. Brookings. https://www.brookings.edu/research/protecting-privacy-in-an-ai-driven-world/.

Lee R.S.T. (2020) AI Ethics, Security and Privacy. In: Artificial Intelligence in Daily Life. Springer, Singapore. https://doi.org/10.1007/978-981-15-7695-9_14

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. *Fundamental Issues of Artificial Intelligence*, 555–572. https://doi.org/10.1007/978-3-319-26485-1_33

Muzyka, K. "*The Basic Rules for Coexistence*: The Possible Applicability of Metalaw for Human-AGI Relations." *Paladyn (Warsaw)* 11.1 (2020): 104-17. Web.

Pereira L.M. (2016) "*Bridging Two Realms of Machine Ethics*: Programming Machine Ethics." Studies in Applied Philosophy, Epistemology and Rational Ethics, vol 26. Springer, Cham. https://doi.org/10.1007/978-3-319-29354-7_10

Polyakova, A. (2019, October 25). *Weapons of the weak: Russia and AI-driven asymmetric warfare. Brookings.* https://www.brookings.edu/research/weapons-of-the-weak-russia-and-ai- driven-asymmetric-warfare/.

Sarangi, S., & Sharma, P. (2019). *Artificial intelligence: Evolution, ethics and public policy* (1st ed., pp. 68–108). Oxfordshire, London: Routledge.

Scherer, M. U. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology*, *29*(2), 353–400.

SXSW. (2018, March 11). Elon Musk Answers Your Questions! | SXSW 2018 [Video]. YouTube. https://www.youtube.com/watch?v=kzlUyrccbos

Thierer, A. D., O'Sullivan, A., & Russell, R. (2017). Artificial Intelligence and Public Policy. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3046799

Turner, J. (2019). Building a Regulator. In *Robot rules: regulating artificial intelligence* (pp. 207–262, 319-369). essay, Palgrave Macmillan. https://doi.org/10.1007/978-3-319-96235-1_6

Yampolskiy, R., & Fox, J. (2012). Safety Engineering for Artificial General Intelligence. *Topoi*, *32*(2). https://doi.org/10.1007/s11245-012-9128-9