Global Governance in the Age of Artificial Intelligence: The Impact of AI/ML on Human Rights

Jason R. Chen

Thesis Adviser: **Dr. Eileen Doherty-Sil**
Technical Adviser: **Dr. Kristian Lum**

CIS 498: CIS ASCS Senior Capstone Thesis

University of Pennsylvania

School of Engineering and Applied Sciences

Department of Computer and Information Science (CIS)

May 3, 2021

Acknowledgements

I would like to take this opportunity to extend my gratitude to those who supported me as I researched and wrote this undergraduate thesis. First and foremost, I would like to thank Dr. Doherty-Sil ("Dr. D") for agreeing to advise me on my thesis this semester; her decades of experience and comprehensive knowledge in the realm of human rights proved to be extremely helpful as I navigated the complexities of this topic. Furthermore, it was in her "International Human Rights" course (PSCI 258) at Penn that I discovered and cultivated my passion for human rights early in college.

I would also like to acknowledge Dr. Kristian Lum, who strongly advised me to take a human rights-based approach to my technical research. She completely shifted my perspective with regards to my analysis of the material, and this forced me to situate all of my learnings within the scope of human rights.

My completion of this thesis could not have been accomplished without the unconditional support of my classmates and peers, including but not limited to: Ella Bei, Hannah Pan, Yuxin Liao, May Xiao, Sherina Wijaya, Margaret Zhang, Beverly Deng, Helen Zhang, Angie Wong, Gabrielle Utomo, Jackie Lou, Kingsley Song, Felix Cui, Cindy Hao, and Larissa Nam — among countless others. These friends provided extensive moral support, much-needed advice, and crucial guidance as I struggled to put my research together throughout the semester. To Ella, Kingsley, and Felix — thank you also for forcing me to challenge my own subconscious biases and by extension, my personal views of human rights. To Yuxin and Hannah — thank you especially for creating opportunities for me to relax and take occasional breaks from this massive undertaking during senior spring.

And of course, it would be a crime for me not to express my unlimited appreciation for my parents — Kevin W. Chen and Xinyi Chen — to whom I owe my entire life and especially my general existence. I would not be here at Penn completing my undergraduate studies in Computer Science if it were not for their incalculable support over the years. All of my personal achievements are the ultimate evidence of their incredible parenting.

Table of Contents

Abstract

Artificial Intelligence (AI) and Machine Learning (ML) are now at the forefront of technology — their research has been accelerated in recent decades by increased quantities of available data as well as significant advances in GPU development. The deployment of AI by law enforcement agencies, militaries, and corporations around the world has sown distrust within the general public, as many fear that these new technologies endanger their most basic human rights. With this thesis, I hope to employ my coursework in computer science, political science, and international relations to situate AI/ML within the wider scope of human rights. In doing so, I plan to propose technical frameworks to guide the continued development of AI/ML in a manner consistent with the principles outlined in the Universal Declaration of Human Rights (UDHR). To that end, I have asked Dr. Doherty-Sil — who has an extensive background in human rights research — to be my adviser this semester.

I. Introduction

Artificial Intelligence (AI) continues to be one of the most rapidly developing technologies of the Digital Age, disrupting industries to become a ubiquitous part of our everyday lives; the expansive research and development of AI applications has not been without controversy, however: many prominent industry leaders and researchers including the late Stephen Hawking, Elon Musk, Steve Wozniak, Bill Gates, Peter Norvig, and Stuart J. Russell has collectively expressed concerns about the impact that these unexplored technologies may have on both existing and future human rights institutions. [1] Many of these experts fear that the hasty adoption and increased reliance of AI/ML technologies both by governments and corporations in the private sector may negatively impact human rights, democratic accountability, and the very foundation of free society. [2]

However, it is not just scientific authorities that are concerned: civil rights organizations including the American Civil Liberties Union (ACLU), as well as normal, everyday citizens have expressed apprehension about the deployment of AI products like facial recognition on civilian populations around the world. [3, 4] Nevertheless, proponents of AI argue that it will create new opportunities in a wide range of fields, including health, education, transportation, social justice, human rights, and wealth generation. [5]

The application of AI and its underlying technologies affects a wide spectrum of different areas including education, healthcare, law enforcement, work, and social responsibility. This is both important but also problematic, especially given that AI increasingly has the power to violate human rights and erode the laws that protect them around the globe. For example, the popular combination of big data analytics and AI/ML can threaten individuals' privacy, as it enables bad actors and even governments to ramp up their surveillance and monitoring of private citizens. [6] On a similar note, AI also has the capacity to weaken equality, work against the prohibition of discrimination, and impair access to other fundamental rights, such as political and personal freedom.

Additionally, technical innovation of AI/ML technologies has now outpaced the ability for government agencies to formulate, apply, and enforce regulation of algorithmic decision-making. Proper regulation has additionally been marred by commentators who believe that governance inhibits innovation and that it is already too late to manage this area of technical development. [47] Nevertheless, international human rights law provides guidelines for defining and assessing the potential harmful effects of AI and its associated algorithms, and it is the obligation of states and businesses to put into place appropriate mechanisms that protect these human rights. As such, it is imperative for the current dialogue on ethics of Artificial Intelligence to consider the human rights implications of this relatively recent innovation.

What Exactly are Human Rights? : Some Context

Human Rights refer to the individual and collective rights that have been enshrined in the Universal Declaration of Human Rights ("UDHR"), and then further detailed in the International Covenant on Civil and Political Rights ("ICCPR") and the International Covenant on Economic, Social and Cultural Rights ("ICESCR"). The UDHR is the leading statement of the rights that every human being enjoys by virtue of their birth; it is a non-binding U.N. General Assembly resolution, but many states operate under the belief that they have an obligation to defend and preserve the human rights outlined in the UDHR as part of their adherence to the Charter of the United Nations, which itself is a legally-binding international document. [7]

The ICCPR, as its name suggests, articulates civil and political rights that a state must obey as soon as its government ratifies this treaty. On the other hand, economic, social, and cultural rights may not always be realized overnight, and so the ICESCR instead requires ratifiers to institute policies that gradually achieve these rights. [7] Nevertheless, it is clear that international law places at least some responsibility on individual states to protect human rights in their own conduct or within their own jurisdictions, which oftentimes may require that they prevent private corporations in their areas of authority from committing human rights violations.

For the purposes of simplicity and conciseness, this paper will focus on the principles outlined in the UDHR, as it serves as the most basic foundation for international human rights law. The UDHR is a critical building block to informing people's understanding of human rights as a whole, and provides a pivotal lens through which we may analyze the impact of AI/ML on our fundamental liberties.

Source: *The United Nations*

Identifying the Human Rights Consequences of AI

The human rights impacts of AI stem from at least three technical sources, two of which can be addressed by conducting a human rights impact assessment before a particular system is deployed. The third source, meanwhile, can be hard to identify even after an AI system is in operation, due to the complexity of the technology. [7]

1) Quality of Training Data: The AI system reflects the biases of the training data fed into it, and can be considered to be an example of the classic “garbage in, garbage out” problem; it may have severe consequences for human rights depending on how the system itself is utilized. [7]
2) System Design: Choices made by system designers have important implications for human rights, especially since these individuals ultimately decide what variables the AI system should consider, prioritize, and optimize — these choices are ultimately informed by the designers’ own life experiences and biases and as such, may positively or negatively impact human rights. [7]
3) Complex Interactions: Once an AI system begins to operate, it may interact with its environment in unforeseeable ways, and it is these complex unpredictable interactions that may significantly impact human rights. Some of the adverse effects may be detected using certain analytical techniques, but there is a distinct possibility that a few of these interactions will go undetected. [7]

It is apparent from these three factors that human bias and prejudice are foundational issues for the ongoing development of AI and ML. For example, the quality of the training data and system design choices rely on data points collected from human beings, as well as the preferences and decisions of human software developers, respectively. These two factors form a significant portion of the groundwork for AI/ML models: since actual people need to write the programs used to train models, and their subconscious tendencies inform their design choices and parameter selection, it clearly follows that their own leanings will ultimately hold significant sway in the performance of the models that they train. Similarly, the training of the model itself involves providing the learning algorithm with training data to extrapolate patterns from, which it can then use to make predictions using new data. In this case, biased data may embed prejudices into the model during the training process, which would cause the model to make discriminatory choices as a result. There is thus a clear understanding of how human rights issues may arise from the first two sources, but the third source is the most dangerous based on the simple fact that an AI system's interactions with its environment and other agents is frequently unpredictable and consequently challenging to account for. From a technical perspective, this third factor is perhaps one of the most important unknowns that software developers and ML engineers will need to contend with in the field of AI research.

Equality, Non-Discrimination, and Privacy: Three Pillars of Human Rights

From a more human rights-centric perspective, AI implementations already affect the entirety of rights currently covered by international human rights instruments, such as civil and political rights, as well as social, economic, and cultural rights. Perhaps more specifically, the continued expansion in the use of big data and AI threatens the right to equality, the prohibition of discrimination, as well as the right to privacy, which the Human Rights, Big Data, and Technology Project consider to be the three main gatekeepers for the security of other human rights. [8] As such, respect for these three fundamental rights must be the basis for future development of ethical AI. These core principles are described in detail below:

Rights to Equality and Non-Discrimination

This pair of rights forms the very foundation of the UDHR and asserts that every human being has equal claim to all freedoms and liberties regardless of their "race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status." [8] Unfortunately, AI has high potential to violate these fundamental rights in a number of ways:

A common concern is that AI systems may perpetuate stereotypes and prejudices present in the data embedded within them. As mentioned previously, it is human software developers that decide which variables are relevant or significant for producing the output when designing the systems themselves; these individual value decisions are ultimately driven by those developers' implicit biases and personal experiences and consequently, may not adequately consider the risks of discrimination that their technologies pose to individuals or groups of whom they are not members of. [8] At the same time, AI easily perpetuates society's existing biases and discrimination since they are essentially trained using data from human decision-makers who harbor their own implicit biases. [7]

On a more subtle note, development and usage of AI and ML is currently dominated by just a few states and major technology companies concentrated in specific geographic areas like "Silicon Valley" in the U.S. [8] As such, continued innovation in the field of AI/ML may widen the digital disparity between different societies and may produce disproportionate benefits between those who are fortunate enough to reap the benefits of this technology and those who are not.

In general, AI-driven applications are also increasingly used in decision processes that impact marginalized or vulnerable individuals; specifically, the expansion of automation into more and more areas of society constructs "digital barriers" for disadvantaged people who are more likely to access things like social services, but are also less likely to have the technical means to do so. [8]

Right to Privacy

Perhaps the most vulnerable individual right is that of privacy, which has very much been negatively affected by modern developments of AI/ML applications. This right specifically underscores that, "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks." [8]

However, emerging technologies have greatly revolutionized the ways in which information can be collected, processed, and shared between different parties. In fact, AI/ML has made it possible for inferences to be made about individuals based on fragments of their data, including their interests and preferences, lifestyles, social connections and perhaps most invasively, their private thoughts; AI technologies now have the ability to uncover people's most intimate secrets using seemingly innocuous bits and pieces of data, and this potential has been greatly exacerbated by the fact that AI systems increasingly depend on the generation, collection, storage, analysis, and use of vast quantities of personal information. [7]

AI's extensive potential to uncover individuals' personal information given minimal data is a serious issue in that it may influence decisions that are made about people, frequently without their knowledge or consent. More recent debates have highlighted the challenge of figuring out how these inferences may be used by or shared with third parties, as well as the consequences that follow these commercial applications. For example, the Facebook and Cambridge Analytica scandal in 2018 demonstrated that inferences based on collected data could be accurately translated into "actionable insights" for political groups. [8] Human rights experts worry that the modern capacity for these political groups to intrusively influence public opinion threatens the very existence of democracy and affects people's freedom of opinion and expression. They argue that interference with the right to privacy may actively discourage free opinion in that it instills a "feeling of being watched" within people, effectively leading to self-censorship. [8] Clearly, the issue of privacy rights extends far beyond privacy itself: it truly is a gatekeeper to other rights necessary to maintain a free society such as freedom of speech and freedom of the press.

Transparency

On a related note, there is frequently also a lack of transparency as to what occurs under the hood of AI/ML algorithms. Even when there is transparency about the manner in which an algorithm operates, people hoping to challenge algorithmic decisions that involve them would be hindered by the nature of the algorithmic process itself. [9] This is especially true given that applications oftentimes involve the use of several different algorithms that interact with each other to perform a specific task. This in turn means that it is difficult if not impossible to trace what individual factors contribute to the final output of an AI system. Furthermore, AI/ML algorithms self-learn, identify patterns, and make predictions in a way that human beings don't fully understand; their learning processes do not replicate human logic and the resulting AI systems are opaque and unpredictable, which makes it tricky to assess their impacts on human rights and dispute decisions made by algorithms. [9]

Although transparency itself is not an explicit human right, it is a necessary accompaniment to the human rights described above. Given that it is nearly impossible to construct a completely objective, unbiased AI/ML system, transparency is necessary so that human operators may understand the variables and processes that lead to certain outcomes — particularly if those conclusions hold significant sway over the treatment of certain communities or individuals. Transparency additionally facilitates accountability, which itself is vital for identifying the parties responsible in the event that an AI system's predictive power goes awry and hurts vulnerable groups. The general public is already quite suspicious and distrustful of AI's apparent omnipotence, and it is crucial that there are transparency standards and accountability mechanisms in place to keep these powerful software systems and their architects in check.

The Significance of Equality, Non-Discrimination, Privacy, and Transparency

It is clear that the three most important liberties to consider when analyzing the human rights impact of AI/ML applications are equality, non-discrimination, and privacy. One of the most salient issues at the forefront of the general public's concerns with these algorithmic systems is AI's broad potential to discriminate against vulnerable populations and communities. This is especially important now given the noticeable proliferation of AI/ML technologies into everyday processes and products, vastly increasing the likelihood that the average individual will suffer directly from algorithmic bias. However, it is important to remember that AI/ML models require massive collections of personal data to be tested and trained; this extreme usage of user information is a source of worry for everyday citizens, who are rightly concerned how these applications may disturb their privacy. As a result, it is imperative to understand and address the ways in which AI research and development collides with the Right to Equality, Right to Non-Discrimination, and Right to Privacy. At the same time, transparency is necessary to hold businesses and governments accountable for safeguarding these three human rights, and is something that must be given adequate consideration when scrutinizing AI/ML development.

II. Literature Review: Towards a Human Rights-Based Governance Framework

International Relations (IR) scholars and renowned Computer Scientists alike have conducted extensive research into the intersection between human rights and artificial intelligence, and have reached similar conclusions about what is necessary to ensure ethical development of AI/ML. Specifically, many have determined that ensuring trustworthy AI requires development of a governance and regulatory framework that promotes socially beneficial AI/ML development without compromising individuals' fundamental human rights. However, discussions of the concern over the need for a more human rights-based approach to AI regulation have predominantly centered around "ethical" guidance, whereas a legal or more explicit rights-based framework is necessary. [6] For example, the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) reaffirmed its support for an "ethical" framework for AI regulation in its 2019 "Ethics Guidelines for Trustworthy AI." [6] The objective then, is to develop a legitimate, legal framework to guide AI development in a way that curbs its potential for abuse without handicapping innovation and research in the field. [6]

A human rights-based approach (HRBA) is ultimately necessary to ensure that AI is researched, developed, and applied in a manner that benefits the whole of human society and would be based upon international humanitarian rights law as well as the responsibility of businesses to respect, protect, and satisfy human rights. [8] Existing mechanisms for algorithmic accountability include data protection, impact assessments, and compliance checks, though they must also be complemented with international human rights law frameworks in order to effectively secure human rights in the face of otherwise unchecked technological innovation. [47] Several scholars have also pointed out that the number of actors in the AI/ML landscape have increased dramatically such that a multi-stakeholder, multilateral approach is vital to a HRBA: corporations, especially information technology companies, are now much more involved in activities that were historically and traditionally state responsibilities, thus shifting the balance of power between governments and the business sector. [8] As a result, any human rights-based approach should take into account the multiple different stakeholders across the public and private sectors when allocating responsibility and ensuring accountability for continued development of AI/ML applications.

To start, the Institute of Electrical and Electronics Engineers (IEEE) has stipulated a set of broad "General Principles" for development of autonomous and intelligent systems in a manner that prioritizes respect for human rights:

1) Governance Frameworks such as legal standards and regulatory bodies must be established to supervise processes and ensure that AI systems do not violate human rights, dignity and privacy but also facilitate traceability — all of which are necessary for building public trust in autonomous and intelligent systems (A/IS). [11]
2) It is necessary to institute ways to translate current and future legal obligations into concrete policy; these ways should additionally take into account the cultural norms of different societies, as well as the different regulatory frameworks across different governments. [11]
3) AI systems should not be granted rights and privileges to an equal extent to human rights; they should instead be subordinate to human judgement and control at all times. [11]

The IEEE's professional recommendations further reflect the consensus among computer scientists and international relations experts that AI/ML as a set of technologies and algorithms is still in a state of relative infancy, and thus require new legislation and regulations to ensure that their increasingly widespread applications do not violate existing human rights principles. It is thus not necessarily enough for governments and other stakeholders to formulate administrative frameworks; they must take concrete action and enact tangible policies that safeguard human rights doctrine. With that said, companies and governments have a moral obligation to do their due diligence across AI-related industries with haste; as AI's underlying technologies and algorithms continue to improve, so too does its associated legislation and regulations.

However, the IEEE is not the only group that has put forth different recommendations for policies to regulate and monitor ongoing and future AI/ML research. Several other groups, including government commissions, legal scholars, AI research scientists, software engineers, and international institutions have engaged in discourse over the main issues that would need to be addressed within a human rights-based governance framework. These stakeholders have converged on a few key priorities including the subject of human agency in AI applications, security of individuals' private information for AI systems that handle vast amounts of user data, transparency of algorithmic tools that are otherwise unpredictable, mitigation of algorithmic bias in AI/ML models, accountability mechanisms for software that has high potential for intentional and accidental misuse, and in a similar vein, the importance of corporate responsibility for safeguarding human rights in commercial deployment of AI/ML technologies. Each of these six critical issues is covered briefly below, and echo the shared perspectives as well as differences in opinion between different stakeholders in AI/ML research and development.

1) **Human Agency and Oversight**

AI systems must empower humans to make informed decisions that protect their own human rights. At the same time, proper oversight mechanisms must be in place: according to the European Commission High-Level Expert Group on AI, this can best be secured via human-in-the-loop and human-on-the-loop, and human-in-command approaches. [10]

Human in-the-loop refers to situations in which a human being is still the ultimate decision maker but their decision is informed by an algorithm. [12] Hypothetically, the human has the power of discretion to either agree or disagree with the decision made by the algorithm. However, issues arise in cases in which the human decision maker simply defers to the algorithm's conclusion or allows it to bear significant weight on their ultimate decision because of its supposedly scientifically-backed calculations. These risks are heightened in situations that involve higher risk, as human decision makers face increased scrutiny when and if they act against the algorithm's conclusion.

Human-on-the-loop refers to situations in which the algorithm itself is the decision maker but its decisions are reviewed by a human being. [12] The human reviewer may theoretically be able to challenge the conclusions of the algorithm; However, these human reviewers are restricted by the terms under which they can dispute an algorithm's decision. Because it would be so difficult for reviewers to pass muster, the actual probability of reviewers challenging the output of algorithms is low.

Human-in-command refers to the case in which humans have the capability to supervise the overall activity of an AI system and the power to decide when or how to use the software in a particular situation. This may involve the decision not to deploy the system under a specific set of circumstances, the level of human discretion at different stages of the system's operations, as well as making the judgement to overrule the system's decisions. [13] This is perhaps the safest and most comprehensive approach to human management of AI/ML applications, though the manpower necessary to institute these oversight mechanisms may not be scalable.

Thus, although many have urged for the establishment of oversight mechanisms for AI/ML technologies, there is clear disagreement over how this may be done. Many argue that installing human-in-the-loop, human-on-the-loop, or human-in-command procedures may preserve an adequate degree of human control over AI systems. However, AI policy experts like international human rights law professor Lorna McGregor disagree that these technical approaches are effective, and instead push the efficacy of choices in governance to curtail negative effects of AI system usage. Legal scholars like McGregor certainly believe that in-house oversight procedures are inadequate, and that concrete governmental regulations are necessary to maintain human agency over powerful AI/ML applications. Nevertheless, there are obvious debates over the effectiveness of technical approaches with regards to human oversight and control over AI systems.

2) **Privacy and Data Governance**

Perhaps one of the most pertinent concerns of AI/ML development is the safety of private citizens' data, especially given that these technologies are ultimately driven by Big Data and powered by the collection of virtually everyone's personal and public information. As a result, data governance mechanisms must be in place in order to protect the data's quality and integrity, but also provide legitimate access to such data. [10] This is an issue that is at the core of many experts' concerns with the deployment of AI, though there seems to be a dearth of academic literature that addresses how exactly public and private entities can resolve these problems from a governance standpoint. Perhaps it is not clear yet how existing technologies can solve such an indeterminate, widespread concern. With that said, this would be an excellent topic for future research into government policy, especially given that the proliferation of technology has frequently been at odds with people's right to privacy.

However, there is indeed actual international legislation that covers this very issue: the General Data Protection Regulation ("GDPR"), which recently came into force in the European Union and the European Economic Area, is noteworthy in this regard for its provisions requiring "data subjects" to be provided with "meaningful information about the logic involved, as well as the significance and the envisaged consequences" of the automated processing of their personal data. [7] This revolutionary set of data usage standards defines three main stakeholders: data subjects (person), who authorize data controllers (organizations) to access their personal data and who may forward that data to a data processor (organization) who is responsible for processing the information for the data controllers. [14]

By outlining specific stakeholders for data security, the GDPR makes clear what the responsibilities and rights of each of these parties is whenever data is used for official and commercial purposes. Some of the most fundamental rights described in the GDPR include the Right to be Forgotten, in which personal data must be erased as soon as they are no longer needed for processing, or in the case in which a data subject has withdrawn their consent for their information to be used. Another is the Right to be Informed, in which data subjects must be informed about the collection and use of their personal data, thus requiring that data controllers uphold certain obligations to their users. [15] As the GDPR demonstrates, the issue of data privacy and security is more so a policy and governance challenge rather than something that is addressable via technology. That is not to say that there do not exist technical proposals for addressing these data privacy problems, which will be explored in later sections.

3) **Transparency**

One of the biggest issues faced by watchdogs and regulators is the unpredictability and uncertainty that exemplifies algorithmic processes and decision making, which makes it challenging to assign blame, explain events, or designate responsibility when things go wrong. As such, many assert that an AI system, its data, and its associated business models must be fully transparent and should be explained to affected stakeholders so that those interacting with these technologies understand their limitations and capabilities. [10] The IEEE argues that there is an additional urgent need to develop new measurable and testable standards for transparency of AI systems that facilitate objective assessments of their compliance. These transparency standards would provide AI system architects with a useful guide for assessing their products' compliance and make it easier for them to understand what mechanisms to develop that provide this necessary transparency (e.g. for users of care or domestic robots, a why-did-you-do-that button which, when pressed, causes the robot to explain the action it just took). [11]

4) **Diversity, Non-discrimination and Fairness**

Another critical problem is that of algorithmic bias, which has been the center of much debate and controversy when it comes to AI development and deployment. This is especially given that many existing AI/ML applications have been reported to display prejudiced behavior, which has negatively impacted particularly susceptible communities.

Consequently, academics and institutions universally agree that unfair biases and prejudices must be avoided and countered at all costs, as its existence within an AI system exacerbates discrimination and marginalization of vulnerable groups. Additionally, AI technologies must be accessible to all and as such, should involve different groups of stakeholders throughout their development process. [10] Along a similar vein, a rights-based approach should concentrate on principles of equality, inclusion, and non-discrimination and focus especially on vulnerable communities such as minorities, indigenous peoples, or disabled persons. [11]

Clearly, there is not quite a technical solution to this massive potential problem of bias, prejudice, and consequent discrimination underlying algorithmic decisionmaking, and so the best alternative is simply to involve as many stakeholders as possible — particularly those who are members of disadvantaged and marginalized groups — so that they may be able to better identify the risks that most affect their communities. As with the issue of privacy, this does not necessarily indicate that researchers have not looked into technical methods to address algorithmic bias, as will be explained in later discussions.

5) **Accountability**

Many agree that AI systems must have appropriate accountability mechanisms in place in order to allow for proper assessments and audits of its algorithms, data, and design processes, especially for more critical applications that affect greater numbers of individuals or groups. It also follows for there to be suitable processes for redress to ensure restitution to negatively affected parties. [10] AI assessments should be required for AI systems deployed by corporations that have high potential for affecting human rights or are safety-critical applications; such an assessment must involve a fundamental rights impact assessment as well as consultations with stakeholders and relevant authorities. [10]

Additionally, states must develop or require procedures that enable immediate redress in cases of rights infringement. On a similar note, public enforcement authorities must develop auditing mechanisms to identify potentially illegal or harmful consequences of AI systems including discrimination and other forms of unfair bias.

Governments play perhaps the most important role in constructing effective official grievance and remedying mechanisms to address the negative impacts of AI on human rights. As a result, they should incentivize companies to perform their due diligence with respect to AI development and assist early stage companies in building the infrastructure to develop technology that complies with human rights standards. [7] However, it is also vital that governments examine their own applications of AI systems and develop accountability and redress mechanisms that sufficiently address the issues that stem from them. Governments and courts must make clear the responsibility, culpability, liability, and accountability for AI systems whenever possible during their development and implementation, as this pushes developers and users to understand their rights and legal obligations. [11] Some argue that registration systems should be set up to track the parties that are legally responsible for a particular AI system, including developers, operators, and system owners; such parties should additionally register their AI system's intended use, data sources, algorithms, outputs, model features, etc. [11]

More importantly, many organizations and scholars alike have proposed Multi-Stakeholder Initiatives (MSIs) — voluntary partnerships between governments, civil society, and the private sector to address development challenges collaboratively, entrench democratic practices, and strengthen regulatory frameworks. Many researchers and academics argue that these multi-stakeholder ecosystems must be constructed in order to inform the creation of new norms — which will eventually become best practices and legislation — as AI technologies are still too new for experts to understand the full impact of their usage. [11]

6) Corporate Responsibility

Businesses and private enterprises have also come under scrutiny for their responsibilities regarding human rights, as many increasingly argue that international laws should apply to them as well. [7] The most important result product of these discussions is the United Nations Guiding Principles on Business and Human Rights ("UNGP" or "Guiding Principles"), which pressures companies to avoid violating human rights through their actions and business relationships. Although these Guiding Principles are not binding international law, they prescribe industry standards and best practices and urge global enterprises to perform their own due diligence in order to ensure that they accurately identify, prevent, and mitigate human rights risks as early as possible. [7] Furthermore, the Guiding Principles specify that in situations under which corporations commit human rights infractions, these corporations must have legitimate mechanisms in place to address the relevant issues. In general, these Guiding Principles may not ultimately be legally binding, but they provide powerful recommendations on how businesses should conduct themselves in a manner consistent with human rights.

In the context of AI/ML, it follows naturally that a bare minimum public policy should require that those involved in building out AI systems, especially corporate enterprises, perform due diligence to ensure that their products do not infringe on human rights. [7] These policies may further be strengthened by requiring that the engineers and operators behind these AI systems allow external reviewers access to their technology's training data and outputs. Having said that, businesses hold proprietary interests in the data and processes that they use to develop their AI products, and so may understandably be reluctant to share such information. [9] In this case, it may be sufficient for these companies to not to make this information publicly available, but instead provide the necessary material to independent organizations that represent public interest, such as government regulators.

In general, companies that develop AI technologies must do so in a manner that is consistent with the norms and values of a rights-based approach, in addition to the particular values of the community for which their products serve. [11] These companies must also abstain from willingly providing their technologies to bad actors who would likely utilize them to commit human rights violations. [11]

Although this literature review provides a basic glimpse into different proposals for human rights-based regulations to AI/ML concerns, there exists a wealth of research into more granular issues. For the purpose of brevity and relevance to this paper, this review concentrates on the main, big-picture topics of human oversight, data privacy, transparency, algorithmic fairness, accountability, and corporate governance. Each of these subjects is closely related to the three pillars of equality, non-discrimination, and privacy that are crucial to safeguarding the wider spectrum of human rights. And although this review does not present a complete analysis of human rights-based approaches, it does provide a comprehensive survey of industry, governmental, and academic recommendations for AI/ML governance and management.

Based on existing academic literature and policy recommendations by nongovernmental organizations (NGOs), independent official commissions, and existing governing bodies, there is a clear understanding of what practices legislation must establish in order to address the vital issues of state accountability and corporate responsibility. However, there is an apparent ongoing debate on the level of human oversight necessary for the development and deployment of AI/ML systems, as demonstrated by the significant differences in opinion between McGregor and the European Commission High-Level Expert Group on AI. Additionally, the GDPR has become an official standard for policymaking with regards to data protection and user privacy.

However, scholars do not seem to have concrete, tangible solutions to the issues of transparency, non-discrimination, as well as lingering problems with data governance. These last vestiges are extremely important, and will be addressed within this paper's technical frameworks for AI/ML development and deployment.

III. Use Case Analyses

AI/ML technologies are now at the center of everyday processes and products and have increasingly become an integral feature of the software that powers consumer electronics, bureaucracy, national security and defense, healthcare, travel and hospitality, mobile entertainment, as well as a nearly infinite number of other applications. Unfortunately, these advanced technologies are mostly proprietary or protected as government secrets and consequently, remain a "black box" to the general public, including researchers and scholars. As a result, the following use cases do not present a complete technical explanation of how these applications operate and function, but still provide general definitions to facilitate understanding.

As the following AI/ML use cases demonstrate, these algorithmic technologies and techniques have exceptional potential but also run a high risk of violating fundamental human rights. With that said, many are under the impression that AI is exclusively something to be feared with its virtually unlimited power and capabilities. However, the reality is not quite as clear-cut: although it is true that many AI models have some negative measurable effect on human rights, it is important to keep in mind that many of these same applications also improve individuals' enjoyment of other human rights. And so it is ultimately necessary to calculate the human rights trade-offs for using AI/ML technology, as will be done in the following discussions of risk assessments, lethal autonomous weapons systems, AI-powered medical diagnostics, and automated online content moderation.

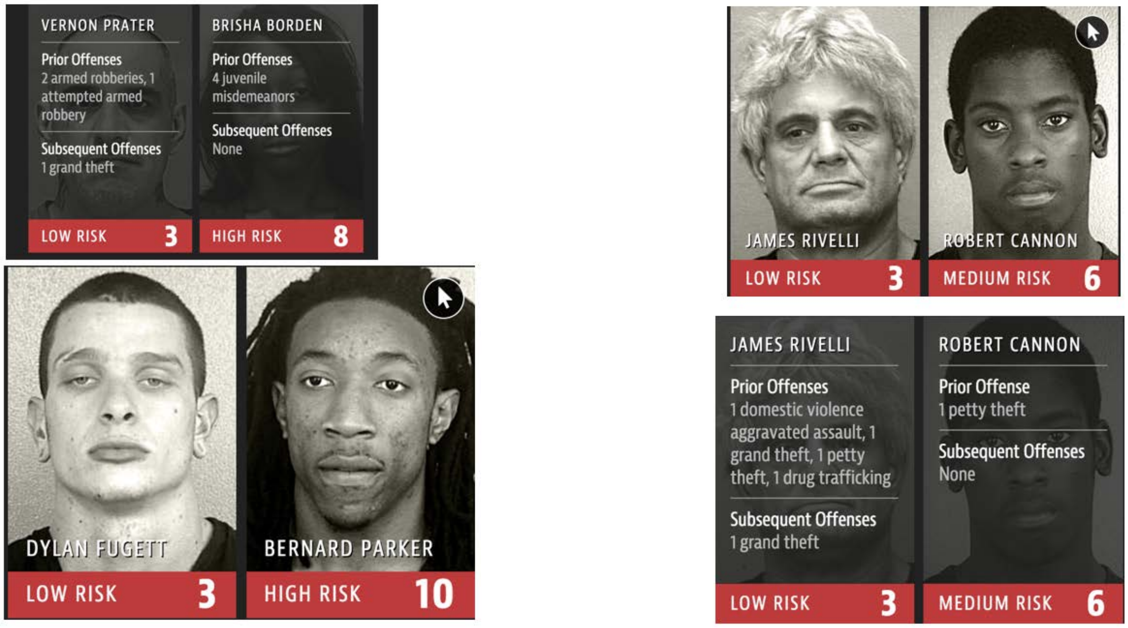
1. Criminal Justice: Risk Assessments

AI's capability for speed, efficiency, and accuracy is far beyond human capacity, but one of its most salient issues is its frequent inability to be fair and neutral, especially with regards to racial sensitivity. The criminal justice system is perhaps the most severe yet legitimate means for which a democratic society may strip its citizens' fundamental rights. However, AI systems have increasingly been deployed to automate decision-making at different stages of the judicial process; they have especially been utilized for risk assessments, which are used to figure out pretrial detention, sentencing, and parole. [7] The latest generation of risk assessment tools uses machine learning to rebalance risk factors in response to new inputs, which theoretically should improve the AI system's predictive accuracy.

These automated risk assessments may help direct police resources and increase efficiency, but there is controversy in that it may also lead to over-policing of already heavily monitored neighborhoods, and disproportionately increase stop and frisk practices based on race and ethnicity. [8] This is especially given the fact that these systems are trained on historical crime data, which itself may be influenced by biases in policing, and would thus simply perpetuate existing discriminatory practices by law enforcement. Those that support and promote risk assessments argue that they merely provide supplementary information, and that judges and prosecutors would rely on their expertise and experience rather than blindly follow an algorithm's decision. [16] However, research into behavioral economics and cognitive psychology suggests otherwise: judges and prosecutors are actually very likely to adhere to the decision made by an algorithm under the perception that it would be more reliable, scientific, and legitimate than their personal feelings. In fact, behavioral researchers found that it is generally difficult and rare for individuals to psychologically "override" the recommendations produced by risk assessment tools. [16]

AI has also been used for risk assessments in sentencing and bail decisions, though critics argue that these algorithmic risk assessments may take into account variables that are not relevant nor useful in predicting an individual's probability of recidivism and as such, would have questionable influence on a judge's ultimate decision making. [8] The increased use but inconsistent adoption of risk assessment tools has also concerned critics, who argue that the non-standardized methods through which these software products are developed has the potential to foster unfairness throughout the justice system. [16]

For example, the most widely-used algorithmic risk assessment tool in the United States is COMPAS, which many state courts use to inform decisions regarding bail and sentencing. [7] However, investigative journalists from public interest nonprofit ProPublica found that COMPAS mistakenly classified African-American offenders as "high risk" at a rate twice that of Caucasian offenders despite the fact that it supposedly had near-equal accuracy rates in predicting when members of either racial group would reoffend. Perhaps more specifically, COMPAS ultimately predicted that 45% of African-American convicts were high risk offenders despite the reality that they did not actually reoffend. In contrast, COMPAS misclassified 23% of Caucasian convicts as "high risk" who did not ultimately go on to reoffend. [7]



VERNON PRATER
Prior Offenses
2 armed robberies, 1 attempted armed robbery
Subsequent Offenses
1 grand theft
LOW RISK 3

BRISHA BORDEN
Prior Offenses
4 juvenile misdemeanors
Subsequent Offenses
None
HIGH RISK 8

DYLAN FUGETT
LOW RISK 3

BERNARD PARKER
HIGH RISK 10

JAMES RIVELLI
LOW RISK 3

ROBERT CANNON
MEDIUM RISK 6

JAMES RIVELLI
Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking
Subsequent Offenses
1 grand theft
LOW RISK 3

ROBERT CANNON
Prior Offense
1 petty theft
Subsequent Offenses
None
MEDIUM RISK 6

Source: *'COMPAS Software Results', Julia Angwin et al. (2016)*

If ProPublica's findings are to be believed, then it follows naturally that the rights of minorities to equality and discrimination are threatened when risk assessment systems like COMPAS are used by law enforcement and judicial authorities. Such analyses are part of a greater controversy that risk assessment tools perpetuate racial bias. These biases may be due in part to the over policing of minority communities and the consequent disproportionate amount of police scrutiny that individuals of ethnic minorities encounter. [7] This has had the effect of overrepresenting these minorities in law enforcement data and subsequently results in algorithms lending greater weight to variables that are associated with race when assessing an individual's probability of recidivism. [7] These issues are further compounded by the fact that these risk assessment tools are developed by private companies who hide their algorithms and training data as trade secrets, which makes it much more difficult for criminal defendants to defend themselves against charges and appeal convictions. [7]

Technical Development of Risk Assessment Tools

A wide range of stakeholders are involved in the construction of risk assessment tools, including local and federal governments, nonprofit organizations, and private enterprises. To start, software developers choose which data sets, testing methods, and programming techniques to use when creating these predictive analytics systems. Statisticians then interpret vast swathes of data to write the predictive algorithms themselves; this requires them to process datasets that include information on sentences, recidivism, and demographics in order to calculate which variables are most relevant to judicial decisions. [16] They then "train" the AI system to identify those same variables for new cases to make predictions; if these predictions reach a sufficiently high accuracy rate, then the system is then ready for testing. [16]

In this last stage of development, the AI system and its underlying algorithm(s) are tested against actual cases that have been decided by human judges. [16] Once the AI system finishes its modules of testing, it may be deemed ready for real world use, where it can then be applied to actual cases.

Assessing the Human Rights Impact of Risk Assessment Systems

In general, many critics agree that these risk assessment tools threaten people's right to equality and freedom from prejudice. For example, none of the judicial predictive systems use race as a variable, but many other variables have played the role of "proxies" for race in its absence and as such, embed the technology itself with racial bias. [16] However, it is not simply a race issue either. For one, many risk assessment algorithms consider gender, and since men commit offences at higher rates than women, there exists an inherent risk that the system would disfavor men. [16] The issue with this is that these algorithms may accurately predict group behavior, but not necessarily the conduct of individuals within a group; in short, big data algorithms use blanket generalizations about a group to predict the behavior of individual members of that group. [9]

The capability for risk assessment tools to decide (or at the very least highly influence) individuals' outcomes in the criminal justice system based on group generalizations is thus quite dangerous. In fact, the simple fact that risk assessments may discriminate against particular communities because of algorithmic bias is a gateway towards abuse of other key rights, as outlined in the relevant Articles below from the UDHR:

Article 1

All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood. [17]

Article 2

Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. Furthermore, no distinction shall be made on the basis of the political, jurisdictional or international status of the country or territory to which a person belongs, whether it be independent, trust, non-self-governing or under any other limitation of sovereignty. [17]

Article 7

All are equal before the law and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination. [17]

To start, the very existence of bias in risk assessment tools violates Article 1 of the UDHR, which describes the basic rights to equality that all humans receive at birth. More specifically, the ethnic and racial bias that is endemic to risk assessments infringes on the rights outlined in Article 2, which specifically inhibits denying an individual their human rights based on their identity and status — in this case, race, color, sex, but potentially other distinctions as well. Additionally, the prejudicial nature of risk assessments violates individuals' equal protection of the law as defined under Article 7. In this case, risk assessment tools themselves may not directly decide the outcome of a defendant in court, but as discussed previously, judges and prosecutors are frequently hesitant to break from the seemingly "objective" predictions made by algorithmic models. As a result, important court authorities simply become "agents" who act on the decisions made by risk assessments. In this way, judges may ultimately enact discriminatory sentencing decisions informed by predictions outputted from biased AI models powered by similarly prejudiced data. Additionally, the "black box" nature of risk assessment tools built and trained by private enterprises deprives criminal defendants of their ability to contest their charges or appeal court decisions, which itself is a key right implied under the equal protection of the laws in Article 7.

With these few Articles in mind, it should be alarmingly clear how dangerous risk assessments are for universal human rights, especially in their current form. Unless the businesses and corporations that develop these systems adopt a human rights-based approach to their engineering practices, there is little probability that the critical concerns outlined above will be addressed.

2. Conduct of War: Lethal Autonomous Weapons Systems (LAWS)

Lethal Autonomous Weapons, also colloquially known as "killer robots," refer to robotics weapons systems that have the capability to identify and attack military targets using its complex arsenal of sensors and computationally intensive algorithms — all without any need for human interaction or intervention during battlefield operations. Although LAWS have not yet seen real-world action, they would need to be able to function independently and without human contact for prolonged periods of time, especially given that unstable communication lines are the norm in the field. As a result, autonomous weapons systems would likely require human-out-of-the-loop system design — in which humans have no role in its decision making — to function. [18] This technology has a wide range of military advantages, as these LAWS would be fully functional in the absence of effective human communication or control, could operate in the field for extended periods at more extensive ranges, and would replace the current need to have so many human soldiers for combat operations. [18] From a purely utilitarian perspective, autonomous weapons systems would drastically cut down on the amount of resources necessary to arm a military force without compromising any firepower and at the same time risking fewer human lives in battlefield operations.

Not much is known about autonomous weapons systems just yet since they are a relatively new technology that is still in the process of being developed by different militaries and defense companies around the globe. Having said that, autonomous technology originated at the intersection between AI and robotics and represents an extraordinary blend between advanced software and cutting-edge computer hardware. In general, autonomous systems are beneficial because of their unique ability to investigate multiple options for action to calculate the most optimal response with little to no human involvement, especially in situations lacking structure and certainty. Additionally, these autonomous technologies are responsive to the changes in their environments, and adjust their behavior accordingly, enabling them to respond adequately to unanticipated, unpredictable events. This makes autonomous technologies ideal for battlefield environments, where important strategic decisions must be made in a matter of seconds and conditions constantly evolve.

For all its benefits and advantages, autonomous weapons systems also hold substantial drawbacks. For one, it takes agency away from human beings in deciding who or what constitutes military targets; this presents significant humanitarian, ethical, and legal concerns, which have mainly been the subject of debate with regards to International Humanitarian Law

(IHL) and in particular, its requirements for distinction, proportionality, and accountability in warzones. [18] Experts worry that LAWS may behave unpredictably and incomprehensibly because they function according to decision processes that are modelled after but still different from the human thought process. These understandability challenges are exacerbated by the absolute speeds at which these systems operate such that humans may not understand what is happening on the battlefield until it is already over and by extension, only after the "damage is done." [19] Unfortunately, the deficiency of understanding and predictability is here to stay, as there is no strategic advantage to addressing this issue, as it would weaken LAWS' performance in the face of more adaptable enemy weapons systems on the battlefield. [19]

Sources (from left to right): *'Toward a Ban on Lethal Autonomous Weapons: Surmounting the Obstacles', Wendell Wallach (2017); 'AI Companies, Researchers, Engineers, Scientists, Entrepreneurs, and Others Sign Pledge Promising Not to Develop Lethal Autonomous Weapons', Ariel Conn (2018)*

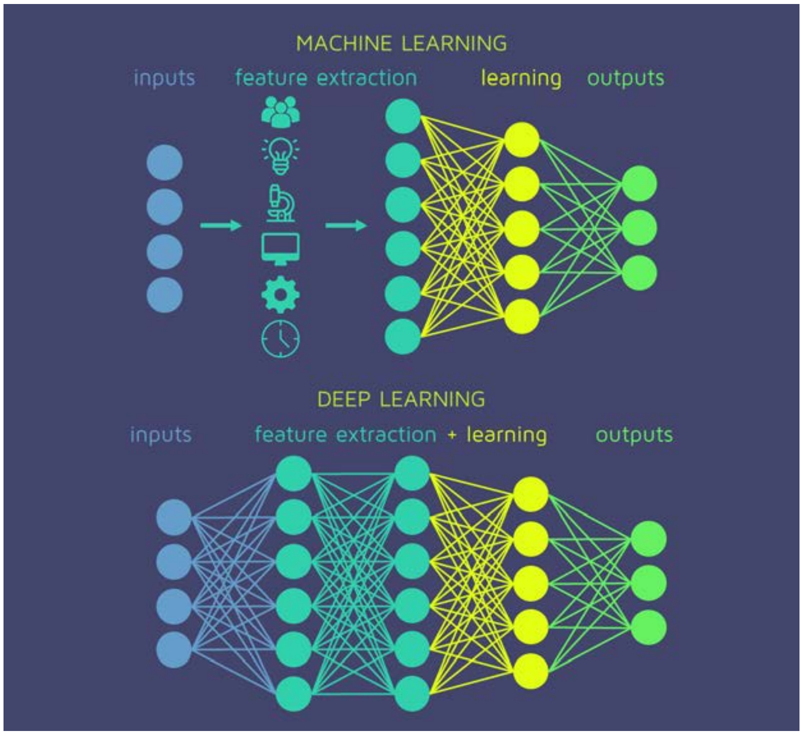
The Hybridization of Technology Behind LAWs

As mentioned previously, autonomous weapons systems are an amalgamation of advanced AI/ML software systems and extremely new and innovative robotics technologies. The "autonomy" of LAWS is based on a few different technical components:

1. Sensors that enable the system to perceive their surrounding environment(s).
2. A set of computer hardware and software that take in information from the sensors, processes it, and uses the resulting analyses to inform the system's next steps and actions. In this regard, the "brain" of autonomous weapons systems primarily consists of computer chips, sensing software, and control software.
3. Communication technologies that enable the system to interact with both machines and human beings, such as human-computer interaction interfaces.

4. End-effectors systems, which are the devices that actually affect the robot's physical environment, and actuators, which are the physical mechanisms that facilitate the movement of end-effectors. [20]

Together, these four distinct factors form the basis for autonomous systems' ability to process information from their environments, analyze collected data, and then dynamically engage with or respond accordingly in fractions of a second. From a software-centric perspective, the actual information processing and data analysis that LAWS perform originates from a distinct sub-field of machine learning known as Deep Learning.



Source: *'What is the difference between Deep Learning and Machine Learning?', (2019)*

Deep Learning itself falls under the umbrella of representation learning, in which the system "learns" how to learn by converting raw data into "representations" known as features, which are individual independent variables that the machine learning model ultimately uses to inform its predictions. The durability of Deep Learning comes from its utilization of simpler representations of information to express more complicated representations of data; in this way, Deep Learning enables computers to iteratively construct complex concepts using simpler concepts. [20] Most basically, Deep Learning algorithms attempt to replicate the human thought processes and capacity to "learn" by imitating the activities that occur between layers of neurons in the human brain. [18]

The "neurons" themselves actually represent hypotheses that have been converted into algebraic circuits with tunable connection strengths, which are then organized into "neural networks" — so named because their circuitry is superficially modelled after the networks of neurons in the brain. [21] These neural networks are considered to be "deep" because the neurons are arranged into multiple individual layers, which means that the path of computation requires several steps to go from inputs to outputs. [21] Deep Learning is not necessarily a new concept, though the computation power of neural networks has seen monumental improvements over the last few decades as a result of improvements in GPU performance and the advent of new mathematical formulas.

Therefore, whereas physical robotics mechanisms power the hardware of autonomous systems, Deep Learning is chiefly responsible for the algorithmic processes and decision making capabilities of autonomous applications. These two branches of applied science make LAWS one of the most advanced and deadliest technologies in defense research.

Assessing the Human Rights Impact of Lethal Autonomous Weapons Systems

The technologies that underlie autonomous weapons systems raise significant concerns about how they might affect human rights during times of war. Specifically, International Humanitarian Law, which is also referred to as the laws of armed conflict, requires that parties conform to core principles of conduct on the battlefield. The existence and deployment of these weapons systems is also at odds with international regulations found in the Convention on Certain Conventional Weapons, which requires discriminating between combatants and non-combatants, judgement of military necessity during an attack, as well as the assessment of proportionality in attacks on military targets with possible collateral damage. [21] According to famed computer scientists and renowned AI researchers Peter Norvig and Stuart J. Russell, improvements in technology may enable LAWS to satisfy the discrimination requirement, but machines are not yet capable of making the subjective and situational judgements necessary to satisfy the necessity and proportionality requirements outlined under international law.

Perhaps one of the most basic concerns is that "justice" cannot necessarily be automated, which is to say that it is not possible to simply translate IHL and other wartime conventions into lines of code for an autonomous system to follow; this decision making must be left to the discretion of actual human beings. [22] For example, there is a delegation of authority at multiple levels in the military, and there is an individual at each level who is responsible for the authority and potential consequences that come with use of force. This hierarchy of command means that these individuals cannot disregard their moral and ethical obligations to determine use of force, as they are inherently subject to the scrutiny of their superior officers. [22] The same cannot be said for autonomous weapons systems, as they are not responsible human agents and thus should not be given the independent capacity to dictate the aforementioned use of force.

The absolute speed in decision making and lethality of LAWs, especially in the absence of human guidance and override, poses significant risks to international norms with regards to human rights in wartime. These liabilities may be analyzed through the lens of the *Just War Tradition*, which many international relations scholars consider to be a universal, binding set of mutually agreed upon rules of combat. This set of theories suggests rationale for judging whether or not a nation must go to war, known as *jus ad bellum*, but also specifies conditions that belligerents must follow during military conflict, referred to as *jus in bello*. [23] It is this latter set of principles that autonomous weapons systems have a high probability of violating.

Jus in bello comprises two broad but extremely important doctrines that are meant to govern conduct of war and minimize unnecessary loss of human life in battle:

1. **Discrimination(also referred to as Distinction)**: Military force must only be applied against the political leadership and military forces of the state. Every effort must be made to discriminate between combatants and noncombatants, soldiers and civilians, to minimize civilian casualties. [23]
2. **Proportionality**: The destruction inflicted by military forces in war must be proportional to the goals they are seeking to realize. The goal should be to use the minimum level of violence to achieve the limited aims of war. [23]

The principles of Distinction and Proportionality are also explicitly embodied in IHL, thus codifying it into international law and perhaps more importantly, providing legal guidelines to prosecute those who violate these human rights standards as war criminals:

Rule 1 of IHL

The parties to the conflict must at all times distinguish between civilians and combatants. Attacks may only be directed against combatants. Attacks must not be directed against civilians. [24]

Rule 14 of IHL

Launching an attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated, is prohibited. [24]

With that in mind, the primary concern with lethal autonomous weapons systems is their scalability, which refers to the fact that a small number of such systems have the potential to unleash an arbitrarily large amount of firepower against human targets defined by their identification criterion. [21] This scalability issue is fundamentally an issue of Proportionality, especially since a LAWS' miscalculation could lead to an excessively powerful attack that causes collateral damage beyond the original target. Human strategists often calculate the amount of military power necessary to achieve some purpose by considering several key factors, including the definitions of military advantage, tactical objectives, civilians, and civilian objects within the circumstances of that particular purpose. [25] These human decision makers additionally consider legal and moral norms, as well as their own personal experience when calculating military necessity. It cannot be guaranteed that LAWS will be able to replicate this sophisticated decision making process in the complex, ever-evolving conditions that exemplify modern battlefields. In addition, computer scientists and IR scholars alike agree that autonomous weapons systems do not and may not ever have the ability to holistically think through the entire context of a given situation and thus, would be unable to comply with the doctrine of Proportionality. [21, 25, 26]

The constant shift in conditions and objectives in combat zones also indicates that Proportionality calculations must take into account developments on the level of military headquarters in order for their own judgement of military necessity to be based on the most recent strategic information. This would require LAWS to have access to a constant stream of data and information from military commanders to inform its decisions. However, as mentioned previously, channels of communication are rarely stable during military operations, and so it would not always be possible for autonomous weapon systems to have the data necessary to stage a proportional attack.

The other chief issue with fully autonomous weapons is its incompatibility with the principle of Distinction, especially given existing challenges of distinguishing between combatants and civilians. In this era of ideological extremism and the rise of terrorism around the globe, it is more challenging than ever for military authorities to differentiate between uniformed soldiers and combatants who dress like civilians and hide in civilian areas. Therefore, separating combatants and noncombatants involves analyzing subtle behavioral cues like body language, tone of voice, and physical gestures to determine an individual's intentions. [26] It is doubtful that inanimate machines will ever gain the ability to catch these subtle cues, which may lead them to engage in indiscriminate killing and ultimately, cause mass, unjustified civilian casualties. With that in mind, LAWS are virtually guaranteed to be in violation of Distinction when and if they are ever introduced to the battlefield.

It is obvious that fully autonomous weapons systems pose a grave threat to human rights in times of war as indicated by their inability to comply with the principles of Distinction and Proportionality, which form the basis for *jus in bello* and are codified into IHL as well as related bodies of law such as the Convention on Certain Conventional Weapons. This specific use case of AI has been an extremely widespread source of concern and opposition for many as demonstrated by the extensively popular Campaign to Stop Killer Robots, an international, multidisciplinary collaboration aimed at lobbying governments against deployment of LAWS. Coalitions like the Campaign to Stop Killer Robots join 26 Nobel Peace Prize Laureates, 4,500 AI experts, 30 countries, 170 non-governmental organizations (NGOs), the European Union Parliament, and Human Rights Council rapporteurs in the fight to ban autonomous weapons from ever being deployed in military operations. [27]

3. Healthcare: AI-Powered Medical Diagnostics

AI has also led to massive improvements in modern medicine and has specifically been used to advance the three pillars of healthcare: prevention, diagnosis, and treatment. However, medical diagnostics is one of the areas in healthcare in which AI has made the most headway.

Especially over the last few years, engineers have developed AI systems based on machine learning and deep learning to facilitate and automate the diagnosis of a wide range of different medical maladies. [7] Current diagnostics systems work by asking human clinicians a series of questions about their patients' current symptoms or conditions, and using their responses to either eliminate potential health problems or recommend specific diagnoses. However, just in the last few years, Deep Learning has increasingly been used to automate medical diagnoses, with promising early results in screening for various different cancers, degenerative disorders such as Alzheimer's, as well as milder conditions including autism. [7]

Unlike the previous use cases for AI/ML, experts are generally quite optimistic about its potential benefits and impact on human rights. These more recent AI-based diagnostic systems currently seem to meet or exceed physicians' performance in diagnosing illnesses and perhaps more importantly, may soon be more accessible than specialized human experts. [28] Perhaps more importantly, their deployment over the next several years should theoretically reduce diagnostic errors and more readily provide high quality diagnostics services at a more affordable rate. [7] It is clear then that the effectiveness and accessibility of these AI-powered diagnostic tools is a promising development towards the greater goal of maximizing standards of living and health around the world.

Having said that, there remain some concerns regarding the fairness and robustness of healthcare algorithms as a whole, which may prove counterproductive to initiatives to improve global healthcare. In addition, some worry about the fact that personal data must be gathered to develop these diagnostic tools. Specifically, healthcare providers must collect a wide variety of personal health and genetic data in order to train the necessary algorithms for medical diagnostics. Such intensely personal data has huge potential for misuse and could eventually affect people's rights to privacy, dignity, and freedom from discrimination. [7]

Source: *'How AI is powering a revolution in medical diagnostics', Andrew Wade (2019)*

Source: *'Is AI paving the way to doctorless diagnosis?', Chris Lo (2019)*

Underlying Technologies of AI-Powered Healthcare

The majority of AI applications in healthcare, including automated medical diagnostics, depend on large amounts of training data in order to make predictions, though the specific machine learning methods that they run on, including linear and logistic regression, decision trees, principal component analysis (PCA), and Deep Learning, tend to focus narrowly on particular tasks and are trained on specific data sets. These ML methods fall under an umbrella of AI known as "Artificial Narrow Intelligence"(ANI), also commonly referred to as "weak AI" for their limited capabilities. [28] As such, technologies that utilize ANI are incapable of versatile abstract learning and are better suited to simpler tasks like general pattern recognition. [28]

Since ANI are only designed to undertake singular tasks, they only operate along a narrow range of different parameters and contexts in a way that "simulates" human behavior but does not quite replicate it. [29] ANI can also be divided into two specific types: either they are reactive, or they have limited memory. Reactive AI has no memory or storage capacity, and simply mimics the human mind's capability to react to different stimuli in situations in which they have not experienced them before. In contrast, most AI falls under the definition of limited memory AI, which have data storage and can "learn," enabling them to use large volumes of data to undergo Deep Learning and inform their decisions and/or predictions. [29]

It is this ANI that powers medical diagnostics and imaging, which it endows with an extreme rate of accuracy that closely mimics human cognition. Therefore, ANI may only be able to carry out very specific functionalities, but recent developments mean that they are increasingly adept at doing so; AI-based medical diagnostics systems are no exception to this, as demonstrated by their high performance in recognizing a wide range of different human illnesses and conditions.

Assessing the Human Rights Impact of AI-Based Medical Diagnostics

As with other use cases for AI, one of the primary issues with its deployment in healthcare (and not just diagnostics tools) is algorithmic fairness and biases. However, healthcare AI's reliance on collections of patients' personal data to "learn" is a source of alarm for scholars, who fear that the extremely sensitive nature of the data involved may represent a breach of privacy rights. Those rights at risk are embedded in Articles 1(Right to Equality) and 2(Freedom from Discrimination), which have previously been defined, as well as Articles 12 and 25, which describe the right to privacy and adequate standards of living, respectively.

Article 12

No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks. [17]

Article 25

Everyone has the right to a standard of living adequate for the health and well-being of himself and of his family, including food, clothing, housing and medical care and necessary social services, and the right to security in the event of unemployment, sickness, disability, widowhood, old age or other lack of livelihood in circumstances beyond his control. [17]

It is almost certain that healthcare AI/ML applications in general are biased due to their heavy dependence on input data, especially considering the fact that most of their underlying models are still in research and development concentrated in a few countries and regions around the world (e.g. Silicon Valley). It is apparent that AI/ML models that are trained on a specific patient population or community may not work as well when fed with data from a different patient population. In fact, most of the input data used to train these models are from Western, educated, industrialized, rich, and democratic (i.e., "WEIRD") populations. [28]

The effects of this have already been observed in real-world medical applications: AI healthcare algorithms in the U.S. have been found to suffer from racial bias in that they assign lower risk to African American patients compared to White patients. These algorithms "learned" that less money is spent on African Americans who have the same level of need as White Americans, and used that observation to infer that African American patients are healthier than their equally sick White counterparts. [28] It is clear from this that current AI-powered medical applications already perpetuate biases and stereotypes, and this issue of equality can only worsen as engineers and researchers introduce more and more autonomous tools into the healthcare space.

However, the varying effectiveness of healthcare AI for different groups and communities threatens another vital human right, as outlined in Article 25 of the UDHR: an individual's right to an adequate standard of living, well-being, and personal health. In the health sector, in which phenotype- and sometimes genotype-related information are important, biased AI has the potential to misdiagnose and thus mis-prescribe treatments for specific subpopulations, which would greatly jeopardize their safety. [30] Although AI-based diagnostic systems do not inherently act with malicious intent, the mere fact that they treat specific communities differently — which may threaten their members' health and lives — violates the human equality required under Articles 1 and 2 of the UDHR.

At the same time, experts worry that the vast collections of data needed to train healthcare AI presents a risk to individuals' right to privacy, as explained in Article 12. For automated diagnostic tools in particular, the data involved pertains to patients' often immutable physiological and health characteristics; as discussed earlier, AI has the power to reveal a person's most intimate secrets with minimal, seemingly harmless information. It follows naturally that AI-driven applications would have the potential to uncover even more sensitive information about an individual if provided with their private health data. It should be no surprise then that privacy poses a significant issue in the development of AI healthcare systems like automated medical diagnostic tools.

Despite these drawbacks, many laud the power of AI/ML in the medical field, with its ability to bring greater healthcare access to vulnerable populations and improve the quality of existing treatments and care. Scholars frequently point to Article 25 (Right to Adequate Standard of Living and Health) as well as the following right as being strengthened by the advent of new healthcare AI technologies:

Article 3

Everyone has the right to life, liberty and security of person. [17]

To start, although some point out that the varying effectiveness of AI-driven healthcare may damage vulnerable populations' enjoyment of their right to adequate standards of living and health, the vast majority of scholars argue that AI applications have overwhelmingly improved the quality of healthcare around the world. For example, leading experts argue that AI-driven diagnostic systems will vastly improve standards of living and quality of life by empowering physicians with an advanced ability to detect their patients' conditions earlier and more accurately. [7] This in turn, will enable doctors to treat their patients earlier, which will help save lives and minimize the effects of disease in the long-term. Along a similar vein, many also assert that AI-based diagnostics systems have had a net positive impact on Article 3 in that its provision of accurate, high-quality diagnostics services have reinforced humanity's enjoyment of the right to life. [7]   In general, it is apparent that AI-powered diagnostics tools are a blessing for healthcare professionals, as it enables them to treat their patients earlier and much more effectively, which raises the standards of health and life for the entire human population.

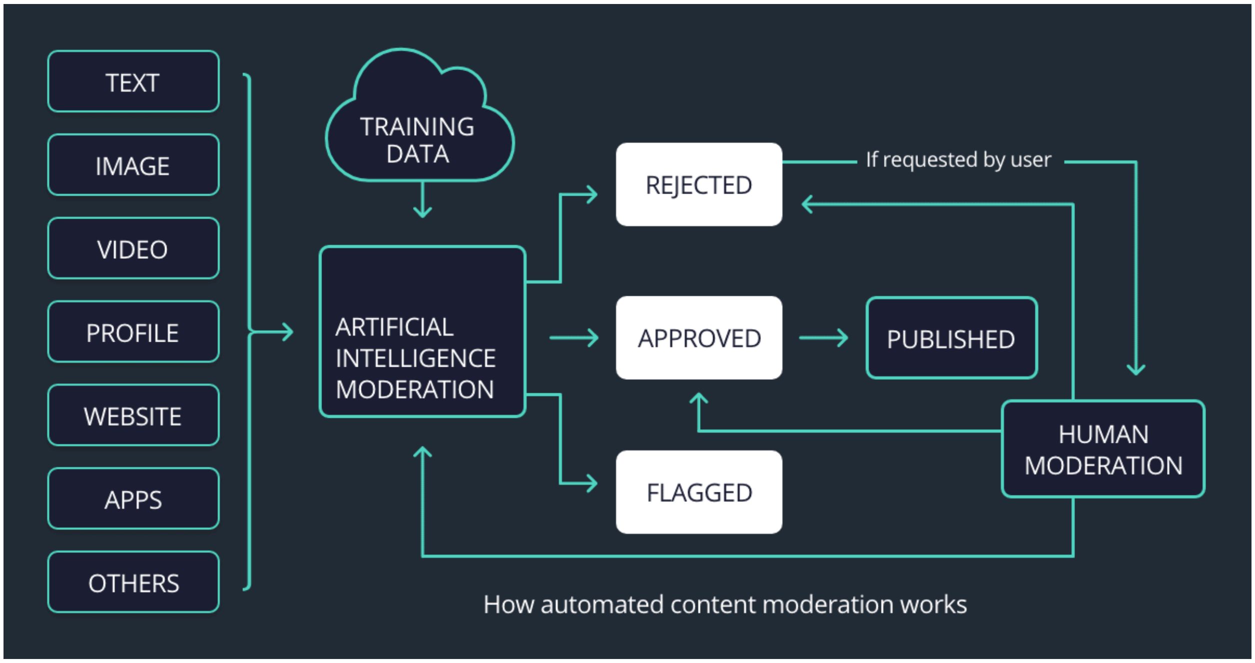
Using automated medical diagnostics as an example, it is clear that the healthcare sector has and will continue to greatly benefit from advances in AI/ML technologies. Although experts continue to harbor concerns about data privacy and how algorithmic bias may affect AI-based treatment of vulnerable communities, the vast majority believe that AI/ML will greatly advance medical technology and by extension, human health.

4. Online Content Moderation: Standards Enforcement

Major online platforms have met the challenge of the Internet's growing volume of content and associated need for content moderation by developing systems to automate standards enforcement. These technologies are still relatively new and most currently in operation are simply used to flag content that is potentially problematic so that human reviewers may then evaluate said content. [7]   AI is often proposed as a tool to identify and filter out problematic content in the ongoing global fight against hate speech, terrorist propaganda, and disinformation. To that end, software engineers have developed algorithmic content moderation systems to automatically filter and remove harmful content. However, these systems are not yet advanced enough to understand the nuances of human speech and consequently suffer from high error rates. [7] The usage of these algorithmic systems comes with an inherent risk of false positives (content incorrectly identified as objectionable) and false negatives (content that is mistakenly deemed to be acceptable), which poses a distinct threat for freedom of expression and access to information. [31]

From a governance perspective, automated content moderation systems are a challenge to transparency and auditability, especially given that companies frequently claim intellectual property in order to deflect responsibility and hide the operative processes behind their products. Since the criteria for content removal continues to be hidden, it will become increasingly difficult to understand the dynamics of takedowns. In general, it is concerning that vetted researchers and trusted third-party auditors are still unable to view the prohibited content databases and examine the underlying functionalities on which automated content moderation systems are built. [32] Having said that, transparency may not be the cure-all solution, but minimum transparency standards are necessary for users and experts to understand the ways in which their speech may be governed online, as discussed in previous sections. More importantly, full transparency empowers individuals to understand whether or not their freedom of speech is upheld on social media platforms and other websites or apps.

Technical Basis for Algorithmic Content Moderation



Source: *'Content Moderation: What is it and why your business needs it', (2020)*

Algorithmic content moderation involves systems that use pattern matching or prediction to classify user-generated content in order to come to a decision or undertake some governance outcome such as removal or account removal. [32]

From a technical standpoint, "matching" typically revolves around "hashing," in which a piece of content is transformed into a "hash" — a string of data that functions to uniquely identify the content itself. Hashes are efficient and effective because they are easy to compute and require less storage than the content itself, which further means that it is computationally cheap to match some hash against a large table of existing hashes. [32] One of the most suitable hashing techniques for content moderation is something called "perceptual hashing," also referred to as "p-hashing," which involves fingerprinting distinct components of content, such as the corners of images or the frequency of audio. This enables these systems to recognize distinct semantic patterns like shapes, colors, or sounds, which makes it possible to identify content even after it has been edited or perturbed. [32] Hashing's main functionality within online content moderation is its ability to match newly uploaded pieces of content against an existing database of content, and so it may oftentimes be deployed to scrub known controversial material from websites and apps.

In contrast, "classification" is used to analyze newly uploaded content for which there is no previous record in said database and thus is appropriate for categorizing this new material. ML empowers most modern classification tools; specifically, many of these ML techniques center around Natural-Language Processing (NLP) and require training language classifiers on massive text corpuses that human reviewers have already labelled as offensive, abuse, hate speech, etc. [32] One way in which NLP has been used to classify potentially harmful online content is using a “bag-of-words” model in which a piece of text is represented as a bag of its words without regard for grammar and word ordering but preserving the multiplicity of each individual word. Using this technique, automated content moderation systems may screen the bag-of-words constructed from a piece of content against an existing bag-of-words built from text that human reviewers consider to go against the online platform’s rules and standards. [33] If the bag-of-words created from the content contains a high frequency of words from the bag-of-words used to represent harmful content, then that specific text may be flagged for internal review or removal.

Another, more advanced classification approach involves using word embeddings, in which the words in a piece of text are converted into vectors in which words that are closer to one another in the vector space have similar meanings. One of the most popular examples of this approach utilizes Word2vec — an algorithm that trains a neural network on word associations from a corpus of text, enabling it to identify words that are synonymous with each distinct word in a sentence or document. [34] If the automated content moderation system detects that a piece of online content consists of words that are semantically similar to words that have been pre-classified as injurious, then it may then flag the content for review or removal.

Researchers and software developers have engineered a multitude of different algorithms and tools to automate online standards enforcement: matching is used to identify content that is already known to be harmful, whereas classification utilizes NLP techniques to recognize new content that is similar in meaning or syntax to text that human reviewers have already deemed to be in violation of an online platform’s community guidelines.

Assessing the Human Rights Impact of Automated Content Moderation

As with virtually all other commercial applications of AI/ML, there is a significant risk that its use in content moderation is susceptible to algorithmic bias discrimination and paves the way for invasion of privacy rights; these rights are described in Articles 2 (Freedom from Discrimination) and 12 (Right to Privacy), which have been defined previously. However, its specific purpose of identifying and selectively removing online content also opens the door to infringement of individuals’ freedom of expression, as explained in Article 19 (Freedom of Expression, Opinion, and Information) below:

> ### Article 19
>
> Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers. [17]

There is a significant risk that the deployment of AI in content moderation poses unfair and discriminatory impacts on specific groups — a clear violation of Article 2. Indeed, researchers have found that content classifiers may be more or less favorable to content associated with race, gender, and other protected categories and as such, also entrench biases against specific communities. [32] Additionally, these systems are not yet advanced enough to understand the nuances of human speech. [7] As mentioned before, algorithmic moderation tools are prone to outputting false positives, in which content is incorrectly classified as harmful, and false negatives, in which objectionable content is erroneously judged to be acceptable. False positives would thus violate individuals' freedom of expression, whereas false negatives may result in a systemic failure to control hate speech or harassment, which could discourage certain groups from participating in online discourse; both of these weaken Internet users' enjoyment of their Article 19 rights.

Similar to many other AI/ML applications in general, automated content moderation also threatens people's right to privacy. In this case, the development of automated content moderation systems typically requires large-scale processing of user data and oftentimes, additional profiling and extra scrutiny of users who engage in risky or questionable activities on virtual platforms. [31] The extra profiling and monitoring of online users, frequently without their knowledge or permission, is akin to surveillance so it's no wonder that human rights experts worry about the impact of algorithmic content moderation systems on people's Article 12 rights.

However, it is important to note that AI-driven standards enforcement also reinforces Article 19 as well as Article 3 (Right to Life, Liberty, and Security of Person). In particular, algorithmic systems have become an effective weapon for guarding freedom of expression on digital platforms. When they function as intended, these content moderation tools empower individuals to practice their freedom of opinion and simultaneously block bad actors from distributing content that hurt vulnerable communities, including terrorist propaganda, hate speech, and disinformation. In this manner, automated standards enforcement is crucial for maintaining a culture of democratic exchange on Internet platforms. Furthermore, these systems uphold Article 3 rights because they are objectively more adept than human beings at uncovering content that is unlawful and/or violates community guidelines, thereby bolstering the well-being and safety of digital consumers.

Although AI-based content moderation systems have both negative and positive effects on the different liberties outlined in the UDHR, their net impact on human rights is still indeterminate. Having said that, computer scientists tend to err on the side of caution, and so many have focused on addressing the human rights challenges of these tools rather than concentrate on their enhancement of other fundamental liberties.

Reflections on Use Case Analyses

These use cases highlight the wide range and complexity of AI/ML applications, showcasing their potential to greatly advance the efficiency and effectiveness of processes across different fields including criminal justice, war, healthcare, and online speech. However, some of these applications have attracted considerable controversy. Although risk assessments are intended to be objective and automate portions of judicial decision making, they have been shown to discriminate communities according to individuals' race and gender. Similarly, LAWS have been paraded as an safer means of waging war, though leading scholars in both the international relations and computer science communities have condemned their inability to comply with the agreed-upon rules of war outlined in IHL. On a softer note, AI-based medical diagnostics systems may be prone to algorithmic bias and infringement of privacy rights, but physicians and healthcare professionals laud their potential to improve global access to healthcare. Likewise, digital content moderation systems may unintentionally censor people's online speech, but many argue that this is outweighed by their capacity to expand freedom of expression around the world.

Whether or not these different use cases are worth the tradeoffs is ultimately up to their operators and users. However, this does not mean that individuals should simply ignore the negative effects of AI/ML systems on their human rights, even if they enjoy the benefits that these applications bring to their other liberties. With that in mind, it is essential that software engineers and system architects devise technical approaches to address these issues, even if their procedures do not completely resolve these problems.

IV. Technical Frameworks to Guide Future AI/ML Development

As discussed in earlier discussions, the three main pillars of human rights are the rights to equality, non-discrimination, and privacy. These three rights are considered the gatekeepers to other fundamental rights and so must be safeguarded in order to provide adequate protection to the vast network of liberties that are directly and indirectly connected to them. As such, it makes the most sense for any technical approaches to the issues posed by AI/ML to center on these three essential freedoms. As noted previously, transparency is oftentimes also a necessary accompaniment to these basic rights, despite not being an explicit right in itself.

Transparency has become a priority for AI experts when it comes to matters of policy, as it enables third parties to see that systems behave within appropriate bounds and audit and challenge its decision-making, which are crucial for building public trust. [9] To that end, many researchers have explored the viability of using distributed ledger technology, namely blockchain, as a means of enforcing transparency and also protecting privacy over AI/ML applications. However, transparency is not always meaningful in that it may not be possible for a human to understand an AI systems operations or underlying logic. [9] Many scholars thus argue that a more comprehensive approach to regulating development of AI/ML applications involves utilizing algorithmic accountability methods to check that algorithmic systems are not prejudiced or discriminatory to any individuals. [45] One of the most popular accountability measures is known as "procedural regularity", which promises to construct stronger safeguards against potential violations or abuse of people's rights to equality and non-discrimination.

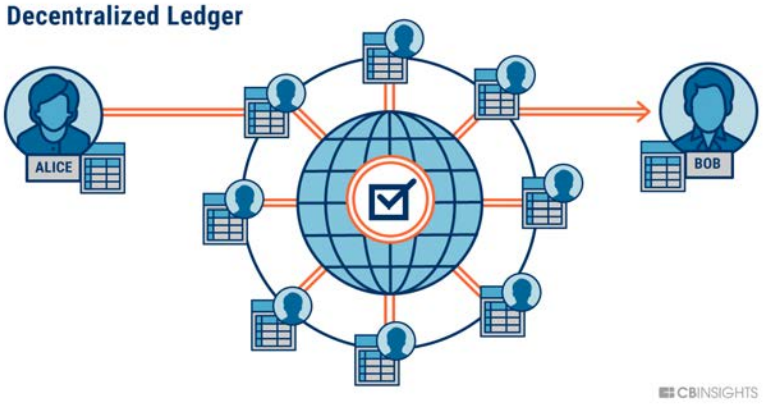
Although these are compelling means of addressing some of the core issues associated with AI/ML, they do not represent complete solutions. There is ultimately no "silver bullet" to the problems that come with algorithmic systems and applications, and there may not be a comprehensive technical solution to these integral challenges for the foreseeable future. As such, the following technical frameworks should be viewed as hotfixes to address the problems at hand but not necessarily exhaustive solutions.

Blockchain: A Means of Shielding the Right to Privacy

One of the most popular proposals for increasing transparency and improving privacy of AI/ML systems is blockchain, as its decentralized structure is ideal for data provenance and data accountability. In basic terms, blockchain involves information being combined together to form blocks that hold multiple sets of data. A block only has a limited storage capacity, so when it runs out of space, it is chained onto the previously filled block, thus forming a chain of data that is referred to as the "blockchain." [35] This process repeats for the next blocks, which will also be appended to the blockchain once they have been filled to capacity. Blockchains are fundamentally unalterable because they digitally represent an irreversible timeline of data such that when a block is filled, it can no longer be modified and will attach to the blockchain with a specific timestamp. [35]

Since information is stored in chronological order and in linear fashion, it is very difficult if not virtually impossible to retroactively alter the contents of a block that has already been appended to the blockchain. Each block contains its own uniquely identifiable hash, the hash of the block before it, as well as its timestamp. That hash is created via a mathematical function that uses the data contained in the block to generate a sequence of letters and numbers; thus, if the information inside the block is modified in any way, so too would its hash code. [35] Furthermore, because of the decentralized nature of blockchain technology (in which a collection of different computers individually known as "nodes" store a blockchain), if a bad actor were somehow able to change their own copy of a block, that copy of the block would no longer be in sync with everyone else's copy. [35] Thus, when everyone else's copy of the blockchain cross-references with one another, they would clearly see that this altered copy is different and mark it as illegitimate. This verification process is known as the blockchain's "consensus protocol," and forms the basis for validating all transactions that occur over the network.

As a result, the only way to successfully alter a piece of information would be if an individual somehow controlled 51% of the copies of the blockchain, such that the majority of copies would agree with their modification to form the new, adjusted blockchain; this is often referred to as a "51% attack". [36] Having said that, the individual would need an insurmountable amount of money and resources in order to change every single block to reflect their new hash codes and time stamps, making 51% attacks extremely impractical and costly. [35]



Source: *'Is Blockchain technology relevant to Smart Cities?', (2019)* In general, many computer scientists find that blockchain's consensus protocols provide built-in detection of data integrity violations, which makes it a prime candidate for tracking data provenance. They argue that information that provides provenance for virtual, physical, and a software application's resources can be stored publicly for transparency and auditability on the public ledger of a blockchain. [36] At the same time, encryption techniques maintain access control and data privacy by enabling individuals to view only the parts of the ledger that is related to them. [36] In this way, data is stored publicly on the blockchain in a way that facilitates full transparency, while consensus protocols and hash functions allow peers on the network to ensure that this data is not tampered with or altered by malicious actors.

In addition to blockchain's consensus protocols and hashing, smart contracts also help build the foundation for transparency and data privacy. Smart contracts are essentially agreements between buyers and sellers that are integrated directly into the code of a system, which can then execute the contract during blockchain transactions. These transaction protocols ensure that the digital contracts are executed, render the transaction(s) irreversible, and also makes them traceable, thus aiding in transparency. [37] More specifically, smart contracts can be used to encode policies and operations related to transparency so that all parties on the network may deploy them to analyze all transactions and queries that are carried out over a transparency-based blockchain network.

For example, these smart contracts can be used to verify the degree of transparency of some AI/ML system by making use of consensus protocols: if one of the various parties on the network, such as AI watch dogs, third-party auditors, and data hosts, calculates and broadcasts a transparency score across the network, but other parties disagree with the specific number then that transparency score may be appended to the ledger along with a note making clear that the specific number was unsupported. A different smart contract may take the average of the parties' transparency scores, so as long as there is consensus as to the network's policies and contracts, this information would simply be added to the ledger with no issue. [38] In any case, these are only two examples among a wide range of different ways in which research scientists have used smart contracts to enforce transparency over a blockchain. More importantly, any blockchain-based approach to securing transparency and privacy for an AI/ML system involves taking advantage of the hash functions, consensus protocols, and smart contracts that underlie distributed ledger technology.

Approach One: Data Accountability and Provenance Tracking

One of the most frequently-cited approaches for a block-chain based accountability and data provenance framework was engineered by Ricardo Neisse, Gary Steri, and Igor Nai-Fovino for the European Commission Joint Research Centre (JRC). The importance of their research lies in the fact that their proposed blockchain framework adheres to the principles outlined in the aforementioned General Data Protection Regulation (GDPR), which went into effect in the European Union in 2016. These relatively recent data protection requirements have not yet found their way into a lot of computer science research, though they will become increasingly critical in the development and deployment of AI/ML systems in the European Union.

The main three entities of this proposed data accountability and provenance framework are the Data Subject, the Data Controller, and the Data Processor, as outlined in the GDPR. Under this approach, When a data subject interacts with a data controller, usually the service provider(s), they create a data usage contract that explicitly stipulates how the controller may use or redistribute any data they may obtain from the subject. This data usage contract would be implemented as a smart contract, which would track data provenance, evaluate data usage control policies, and log events over the network. [14] This would enable subjects to check that data transfers and transactions conform to the contract policies guaranteed on the blockchain. This basic framework is illustrated in the following figure:

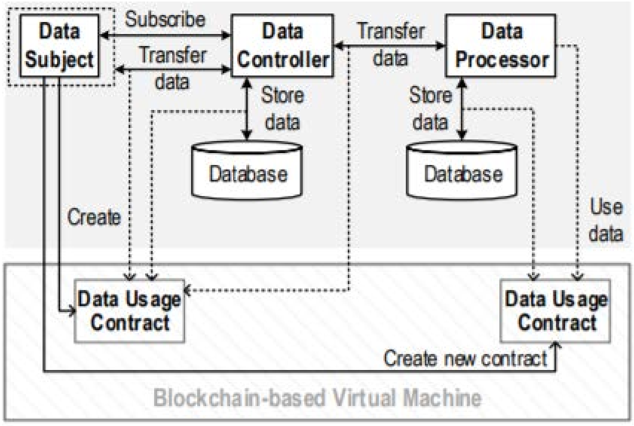


Figure 1: Architecture

Source: *A Blockchain-based Approach for Data Accountability and Provenance Tracking*

Under this proposed data transparency management system, the subject subscribes to the data controller, creates a contract that specifies the ways in which the controller may use or access their data, and then transfers the data to the controller. [14] Moreover, whenever a subject creates a new contract, they must use a new address on the blockchain to prevent their contracts with different controllers from being linked with one another; this would thus require subjects to manage a list of addresses for their different contracts. Then, just as in the previous case, the subject transmits the data to the controller after producing a new contract. [14]

A contract would keep track of the data transmitted to the controller, including the values of the data and reports on data instantiation. However, this information would be encrypted via the SHA3-256 hash function before being stored, as the public nature of the blockchain would enable others to access subjects' data otherwise [14]. In this way, the only information stored on the public ledger would be hashes of the data instance values and data instantiations, which would only be known by data subjects and controllers.

These sequences of interactions between entities on the blockchain and their smart contracts is illustrated on the following page:

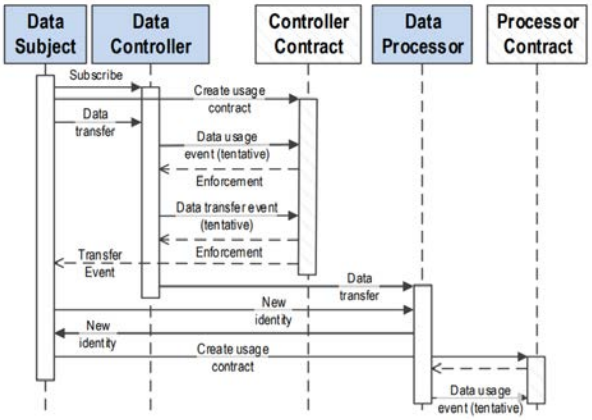


Figure 2: Sequence diagram

Source: *A Blockchain-based Approach for Data Accountability and Provenance Tracking*

This specific data accountability and transparency system design is centered on the use of smart contracts to empower data subjects to hold agency over the ways in which controllers and processors handle their personal data. In particular, a controller would need to check the conditions listed in the contracts of subjects before performing data usage activities such as accessing subjects' data, storing data in the controller's local database, transferring or redistributing data to data processors, and generating derived or consolidated data. [14] Additionally, the contract policies and events would be anonymized via hash functions to protect the privacy of data subjects. Nevertheless, if the relevant subject's smart contract permits the activity, the controller must record the action with a blockchain transaction to indicate that the data usage event has occurred. This recording of the event may later be used for accountability purposes. [14] Similarly, the smart contract would also be required to authorize any transmission of data between the controller and the processor. If this controller-to-processor activity were allowed by the contract(s) in question, the data would be delivered and a notification of the event as well as information denoting the address of the processor would be generated and converted into an encrypted format that only the subject would be able to access.

These two procedures are in line with the Right to be Informed as stipulated in the GDPR, as the smart contracts ensure that data subjects are properly alerted and consulted whenever controllers and/or processors access or use their personal information. [14] However, this technical approach also safeguards the Right to be Forgotten as outlined in the GDPR in that the system is structured such that a subject would have the authority to withdraw consent of their data usage at any time by simply deleting their usage contract from the blockchain. [14] This procedure would deactivate the contract, but still preserve the contract's entire transaction history, which is essential for maintaining accountability even as controllers' and processors' data access conditions change. This thus satisfies the GDPR's requirement that systems be designed to allow subjects to withdraw permission for use of their personal data. On a similar note, this system design enables a subject to include additional conditions or restrictions to processors' and controllers' use of their data, either by including more data provenance information in their initial contracts, or by creating "child contracts" as extensions to the original "parent contracts." [14]

Clearly, this blockchain-based approach to accountability and data provenance is a powerful means of maximizing the transparency of AI/ML systems and by extension, securing individuals' Right to be Informed and Right to be Forgotten, which are essential to safeguarding their Right to Privacy — one of the three pillars of human rights. The strict guidelines set forth by subjects' smart contracts under this framework equips individual users with extensive agency over the usage of their personal data, as they have complete control over who accesses their data and how that data is used, and may limit or expand this access at any time. The anonymization of each subject's data and contracts is another strength of this proposed blockchain network design, as it provides additional protection of user privacy.

As mentioned previously, this proposal in particular has gained traction among computer science researchers and legal scholars in that it specifically implements data usage standards formulated under the GDPR into its system design. In fact, some argue that this approach may also be valuable for integrating algorithmic accountability mechanisms into AI systems, especially since this design records every single transaction and activity over the network to inform subjects how, when, where, and why their personal data is used. In the context of algorithmic accountability and fairness, the blockchain itself could be utilized to record a data point's origin and verify that an individual's data is accessed, transferred, or used in a manner consistent with their rights. In this way, it would be possible for the blockchain to trace an AI system's particular decision through the different variables or data that influenced it, as well as the weight that the algorithms in question distributed across those variables and data points. [9]

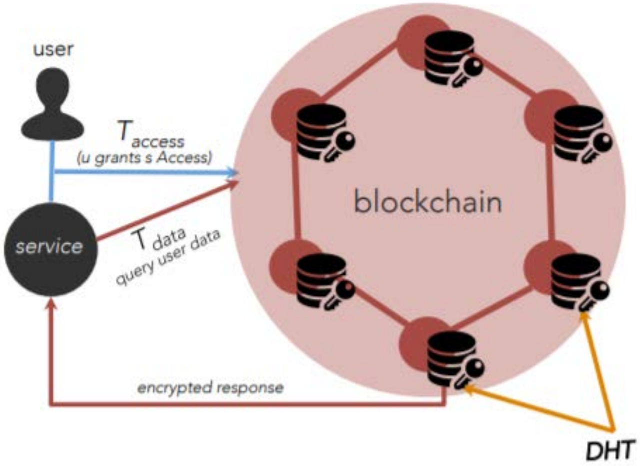
This project from the European Commission Joint Research Centre is an excellent example of how open distributed ledger technology may be adopted to design data transparency, accountability, and privacy management systems. Indeed, this initial concept seems to have already sparked a new field of research, though it is far from the only recommendation of utilizing blockchain to secure data-related human rights.

Approach Two: Decentralized Privacy Protection

Guy Zyskind, Oz Nathan, and Alex Pentland, researchers at MIT and Tel-Aviv University, similarly propose a decentralized personal data management system that enables users to own and control their own data over some application, as well as an underlying protocol that uses blockchain to automate control of access to personal data in a way that doesn't require trust in a third party. The core parties would be the users themselves, the service providers who operate the application and who may use personal data for commercial or operational purposes, as well as "nodes", whose responsibility is to maintain the blockchain and a distributed private key-value store to map users to their personal data stored on the blockchain. This proposed privacy-centric data management system would be built on a blockchain that allows two types of transactions: T_{Access} which would be used for access control management, and T_{data}, which would be used for data storage and retrieval. [39]

The way in which this system works is that when a user signs up for an app the very first time, their information is generated and sent via a T_{Access} transaction to the blockchain. Any data that is subsequently collected on that user is encrypted and transmitted to the blockchain in a T_{data} transaction, which then transfers it to an off-blockchain key-value store. The key in this key-value mapping would be the SHA-256 hash of the data and so would also serve as a pointer to a user's data on the public ledger. [39]

Under this data management system, service providers and users would be able to query their data using their key and a T_{data} transaction. The blockchain would need to confirm that the key belongs to the correct user or service provider before retrieving the private data. Perhaps the most important feature of this system design is that the user would have full control over who has access to their data. [39] In particular, users would have the ability to change the permissions granted to the app's service provider(s) by issuing a T_{Access} transaction specifying new permissions, which would include retroactively revoking the provider's access to user data that has already been stored. The figure below exhibits these procedures in action:



Source: *Decentralizing Privacy: Using Blockchain to Protect Personal Data*

This proposed data management system would empower users to be the owners and sole controllers of their personal data, while delegating service providers as guests who require explicit permission to use this data in any way. [39] Thus, the blockchain would store access-control permissions and policies such that only users would be able to alter them. Users would also be able to track what data an app collects on them, as well as how that data is accessed for use. This framework would also enable users to freely grant or revoke a provider's authorization to access their data, which is important given that most apps currently require users to grant providers indefinite access to their personal data upon sign-up.

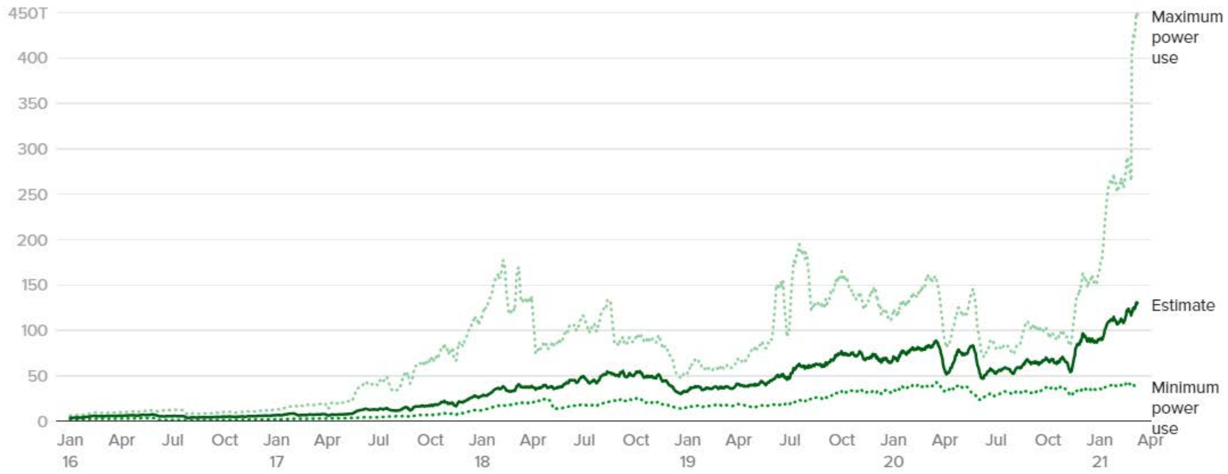
In these ways, this particular blockchain-based data management system provides many of the same benefits as that proposed by Neisse, Steri, and Nai-Fovino in their research for the European Commission Joint Research Centre. Like the previous framework, this approach also centers on giving users ultimate authority over who they allow to access or use their personal data. Furthermore, this proposed system design by Zyskind, Nathan, and Pentland grants users the capability to rescind service providers' access to their data, which is similar to data subject's power to withdraw consent in the earlier proposal and is thus also commensurate with the Right to be Forgotten in the GDPR. This specific approach is actually quite comparable to the earlier framework, as both utilize blockchain to enforce data privacy and grant individuals the agency to control how and by whom their data is used.

Having said that, there are key differences. To start, the system designed by Zyskind, Nathan, and Pentland makes use of transactions to empower users to determine how their personal data is used. In contrast, the system designed by Neisse, Steri, and Nai-Fovino concentrates on employing blockchain's built-in smart contrasts to grant users the same privileges. At the same time, this second approach focuses on using blockchain to protect personal data and privacy, whereas the first uses blockchain to address a wider range of issues in addition to privacy, including data provenance, transparency, and accountability. In any case, both of these approaches epitomize the potential for blockchain to be used to regulate AI/ML systems that collect and process vast collections of data, and whose decision processes are oftentimes unpredictable and difficult to trace.

The Problem of Energy Consumption in Blockchain

However, there are significant drawbacks to blockchain-based approaches to data governance, transparency, accountability, and privacy. The most substantial disadvantage is related to blockchain's distributed consensus protocol, which is used to manage the chronological order of the blocks and check that incoming transactions to the public ledger do not conflict with previously appended transactions. Specifically, in order for new blocks to be added to the blockchain, they must first go through a verification process that requires the other nodes in the peer-to-peer network to solve a crypto-puzzle — usually the SHA-256 hash function — before a block can be appended. [36] This process is known as proof-of-work (PoW) and is an extremely energy-inefficient consensus protocol. For example, the annual electricity consumption for Bitcoin was estimated to be 15.77 Terawatt hour in 2017, which was equivalent to 0.08% of the entire world's electricity consumption that year. [36]

Energy spent on bitcoin mining is approaching the levels used to power the world's data centers.



Figures are in terawatt-hours. This year, global data centers are expected to use about 200 TWh of electricity.

Source: Cambridge Bitcoin Electricity Consumption Index

Unfortunately, this issue has only worsened as blockchain technology proliferates, especially with the continuously rising popularity of cryptocurrencies like Bitcoin. Indeed, as of the writing of this paper, the University of Cambridge's Bitcoin Electricity Consumption Index estimates that Bitcoin's electricity usage will reach 134.63 Terawatt hour this year — a shocking 753.71% increase since 2017. [40] The excessive electricity consumption required for Bitcoin mining exemplifies the environmental impact of blockchain technology; if current trends continue, some estimate that Bitcoin emissions could single-handedly raise global temperatures by 2℃ over the next 30 years. [41] Clearly, distributed ledger technologies like blockchain may offer a host of advantages for ensuring transparency and securing people's right to privacy, but their negative effects on the environment are irreconcilable with the wider objective of protecting human rights institutions.

Although somewhat out of the scope of this paper, it is important to note that climate change poses a threat to the lives, livelihoods, and the survival of entire peoples. As the Office of the United Nations High Commissioner for Human Rights (OHCHR) points out, vulnerable populations, such as those living in island or coastal nations in the developing world, will be the first to feel the impact of global warming and rising sea levels over the next few decades. [42] With this in mind, it is necessary to keep in mind the environmental impact of blockchain technologies because although it may help advance certain human rights, it may also damage other critical rights.

Having said that, computer scientists have recognized the excessive energy use required for blockchain, and many have devised alternative consensus protocol mechanisms that demand fewer resources. The most popular of these is the proof-of-stake (PoS) approach, in which the blockchain network pseudo-randomly determines which node has the ability to attach the next block to the ledger based on the amount of resources each node has deposited or "staked" for this purpose. [43] As a result, the probability that a node is chosen to participate in the consensus protocol is linked to the size of their stake in the system; this mechanism also disincentivizes selected nodes from breaking the network's rules, as misbehavior would cause the node to lose its deposit. [43] More importantly, since PoS does not depend on computationally intensive processes like calculating cryptographic hash functions and instead concentrates on the amount of resources under the control of its major stakeholders, this procedure could be scaled to large systems and still maintain a high degree of energy efficiency.

Is Blockchain Still the Right Call?

With that said, blockchain remains particularly useful alongside AI/ML applications, whose algorithms and processes are oftentimes unclear and unpredictable and thus spark concerns about how they may misuse user data and endanger certain human rights. As the two technical frameworks demonstrate, blockchain is a powerful tool for ensuring accountability and tracing the flow of personal data over algorithmic systems, which is vital for guaranteeing transparency. More importantly, this distributed ledger technology empowers individuals to hold complete control over how their data is used or processed, thus strengthening their Right to Privacy — one of the core gatekeepers of human rights.

Procedural Regularity: Mechanisms to Preserve the Rights to Equality and Non-Discrimination

One of the most frequently-cited concerns about the proliferation of AI/ML applications in modern society is their propensity to discriminate against vulnerable groups — either due to systems designed by inherently biased individuals, or because of input data that necessarily exhibits the deep-seated leanings of people. To that end, one way in which algorithmic bias may be mitigated is through procedural regularity, in which an algorithmic system (such as one that involves AI/ML models) is tested to ensure that the same process was applied in all cases without revealing specifically how the system operates. [44]

Procedural Regularity techniques are essential for proving or disproving that an AI system uses the same decision policy to make each decision, that this decision policy was developed and documented before potential inputs were known, and that its outputs are reproducible. [46] This is especially important given that so much AI research and AI/ML applications are either government secrets or private enterprise's trade secrets, as it enables their software to be tested without the need for the underlying technology, processes, or data to be revealed to regulators or other auditors. However, procedural regularity is simply a first step for regulators or the software developers themselves to check that their program treats all subjects the same way, and is not necessarily a full-scale solution to algorithmic discrimination.

Joshua A. Kroll, an assistant computer science professor at the Naval Postgraduate School, has been at the forefront of pioneering methods of procedural regularity, and most academic literature on the topic stems from his research. Kroll has led the field in developing computational techniques to provide algorithmic accountability even when some information is unknown, namely: software verification, cryptographic commitments, zero-knowledge proofs, and fair-random choices.

Software Verification

Software verification refers to techniques used to mathematically prove that a program has specific properties; this involves analyzing the program's code or writing other programs to determine a system's invariants — properties that don't change at any point through the program's execution. As such, software verification techniques output mathematical proofs that demonstrate that a verified program has certain invariants. [46]

There are a number of different methods for software verification: software verification programs can transform a program into a form that exhibits the desired invariants to certify it; a piece of software can be exhaustively tested to ensure that invariants are maintained through different basic cases and edge cases; a program can be constructed using specially-designed software that ensures preservation of specified invariants and provides proofs of those invariants. [46] It is important to note that software verification simply proves that a software system satisfies its requirements, but does not necessarily indicate that the program itself abides by legal regulations or social norms.

In the context of human rights and AI, software verification may be used to ensure that an AI/ML system consistently abides by the rules that software engineers have integrated into its code. This is especially important if developers explicitly encode specific human rights standards into the system's invariants, such as procedures to ensure that its decision processes treat different subjects fairly, which would be vital for maintaining the Right to Equality. Having said that, software verification would only be as effective as an AI/ML system's operators and system designers are impartial; the invariants that software verification checks for are devised by human beings, who harbor their own subconscious biases and prejudices. Thus, software verification is extremely useful for ensuring that an AI application abides by certain standards, which may be sufficient to secure some human rights. However, those standards are ultimately constructed by human beings, which opens up the possibility that software verification would also be susceptible to bias and discrimination.

Cryptographic Commitments

Cryptographic commitments are analogous to a sealed envelope that is held for safekeeping by a third party; they bind a committer to a specific value contained inside the commitment (i.e. the digital object contained inside the sealed envelope) so that the third party may unseal and verify that the contents of the envelope are consistent with what the specific value expected. [46] This means that the third party may establish whether or not the digital contents inside the metaphorical envelope, such as a program's source code or a file's contents, have not been altered since the commitment was issued and that the committer themselves knew what these contents were at the time of the commitment. [46] Perhaps more importantly, cryptographic commitments are secure, and so the existence of the commitment itself does not reveal anything about its digital contents.

The process of creating a cryptographic commitment is relatively simple: the operations that compute a commitment from a digital object also output a corresponding key that can be used to verify the commitment itself; in this way, the only way in which a commitment may be verified is using that precise opening key and specific digital object. [46] At the same time, cryptographic commitments have a few important properties:

1) It is impossible to figure out what the original object is based on just the commitment.
2) It is possible to use the key and the digital object to establish that the commitment corresponds to the original object.
3) It is impossible to create a fake object and fake opening key such that using the fake key in conjunction with a legitimate commitment would uncover the fake object. [46]

Cryptographic commitments are important for ensuring procedural regularity for automated decision making systems, as they can be used to test whether or not the same decision making policy is used in different use cases and also verify that a software's policies were fully developed at a specific moment in time. [46] As a result, regulatory agencies and other government watchdogs can utilize cryptographic commitments to record a system's source code, input data, and decision policy at a particular point in time. Officials can then open the commitments at a later date in order to prove in court that a program was or was not altered in any way as a result of events that occured after the commitment was created. [46]

In this way, cryptographic commitments are an excellent accountability tool that regulators can utilize to ensure that AI/ML applications are designed explicitly to treat all users or data subjects equitably. This goes one step further than software verification, as it introduces a mechanism that allows third parties to audit an algorithmic system and check that its decision processes affect different individuals or communities in the exact same way. Furthermore, the fact that commitments are themselves unalterable enables auditors to confirm or disprove that software developers did not modify their code as a result of public inquiries. These accountability instruments are crucial for upholding fairness standards, and may be integral in regulators' efforts to force businesses to prove that their AI applications adequately respect the rights to Equality and Non-discrimination.

Zero-Knowledge Proofs

Zero-knowledge proofs can be used in conjunction with cryptographic commitments to enable an individual to prove that some decision policy exhibits a specific property without revealing the decision policy itself or requiring the individual to explain how they recognized that the decision policy has that property. [46] The interactions between zero-knowledge proofs and cryptographic commitments to validate procedural regularity can aptly be described below:

> "If a decisionmaker makes a trio of commitments, A, B, and C, where A is a commitment to the decision policy, B is a commitment to the inputs that were used in a particular case, and C is a commitment to the decision actually reached in that case, then zero-knowledge proofs let the public verify that A, B, and C really do correspond to each other. In other words, the decision maker can prove that, when the committed policy A is applied to the committed input data B, the result is the committed outcome C." [46]

Zero-knowledge proofs empower software developers to build audit logs that affirm that they used the right decision policies to process the correct inputs in order to produce their stated outcome(s). This enables them to provide some degree of transparency to the general public without disclosing their underlying decision policies or private data. At the same time, if the output of their system is ever challenged in court, the engineers may reveal their actual policy and input data to prove that it matches their cryptographic commitment. [46] The zero-knowledge proof then allows them to prove that they consistently applied the same decision policy to multiple different decisions, confirming their honesty and fairness across separate data subjects.

In terms of human rights, zero-knowledge proofs combine with cryptographic commitments to equip AI/ML software developers and system architects with the capability to demonstrate that their programs comply with important rights standards. This is especially important given that one of the most salient issues right now is the profound lack of transparency behind how AI applications operate. This problem is due in part to the fact that many advanced AI applications are either government secrets or trade secrets that companies have hidden away under the guise of intellectual property. However, a mixed zero-knowledge proof and cryptographic commitment approach permits government agencies and private businesses to demonstrate that their software adheres to human rights principles, including but not limited to the rights to Non-Discrimination and Equality. As a result, these two methods encourage AI researchers and developers to be transparent about their human rights performance without forcing them to disclose their underlying technologies. Having said that, zero-knowledge proofs suffer from the same weakness as software verification: they may enable software engineers to establish to the public that their AI applications exhibit certain properties, but this does not necessarily ensure that these properties completely comply with human rights standards.

Fair-Random Choices

The decision making processes of algorithmic systems frequently involve some degree of randomness, though the fairness of that supposed randomness must be verifiable by third parties. Engineers must prove that randomized processes in their programs do not affect their software's ultimate outcome. One popular method of ensuring fair-random choices is by designing the decision process such that completely randomized choices are replaced by a small, recorded random input (known as a seed value) that forces the program to compute random values in a deterministic, only pseudorandom way. [46] This forces the system to behave in the exact same way for the same inputs so long as the seed is not changed. More importantly, it means that the outputs and decision making processes for a program that involves randomized choices can be reproduced and reviewed by regulators and auditors. [46]

A proposed method of generating public confidence in a system's randomized processes is for software engineers to involve a combination the following mechanisms in its programming:

1) A random value provided by a trusted third-party that is made public knowledge.
2) A random value from the software developer(s) that may be kept confidential.
3) Some piece of information that is immutable and specific to an individual's personal profile that can be used to associate them with an input data point.
4) Another value that the software developer(s) selects. [46]

This combination of different values from multiple stakeholders would prevent software system architects from controlling the adoption of random values to skew results in some particular direction. As such, integrating fair-random choices into the operative processes of AI/ML applications is vital because it minimizes the risk that the system will exhibit the subconscious biases of its creators. In general, by involving different values from several different stakeholders, it is possible to reduce the probability that a single party's inclinations will disproportionately affect the randomness of AI systems. This in turn, makes it less likely that an AI application's randomized processes unevenly impact certain individuals or communities, and therefore helps uphold the rights to Equality and Non-discrimination.

The Effectiveness of Procedural Regularity in Mitigating the Algorithmic Bias of AI/ML

Procedural Regularity techniques, as proposed by Dr. Kroll, unquestionably provide a practical toolkit for evaluating and enforcing human rights guidelines, such as the Right to Equality and the Right to Non-Discrimination, within AI/ML systems. Software verification may be used to ensure that these applications are designed with these two rights in mind, whereas the combination of cryptographic commitments and zero-knowledge proofs bolsters AI system architects' capability to prove to the general public that their programs adequately maintain these liberties without needing to unveil their confidential designs. Lastly, fair-random choices block a single party from disproportionately influencing an AI system's pseudo-randomized processes, thus preventing its models from treating different subjects unequally.

However, Procedural Regularity comes with its own set of limitations from the perspective of human rights. To start, software verification and zero-knowledge proofs may obligate software engineers to integrate human rights standards into their code, but this does not necessarily mean that their design choices are free of their subconscious biases. As a result, even if engineers construct their programs to explicitly uphold certain rights, it is entirely possible that the mechanisms that they design to do so may infringe on other human rights. Thus, Procedural Regularity is an exceptional starting point to addressing the issue of AI's algorithmic bias and strengthening protections for the Right to Equality and Right to Non-Discrimination, but it is in no way an all-inclusive, thorough solution to the problems at hand.

V. Conclusion

Artificial Intelligence and Machine Learning can perhaps be considered the future of human technological innovation; these fields of research have seen considerable progress and advancement over the past few decades, and will continue to revolutionize the ways in which people live their lives as private enterprises and states continue to pour money and other resources into their continued development. AI/ML already form the technical backbone for a multitude of everyday products and services, from the apps on mobile devices to the crash prevention software that keeps drivers safe in automobiles. Clearly, AI has nearly unlimited potential to assist people with tasks and practices, though that omnipotence has significant drawbacks for humanity as a whole.

These issues are most salient when it comes to human rights, which are the moral principles and norms that everyone and every*thing* must follow in order to safeguard humanity's basic survival and well-being. The most important of these are the Right to Equality, the Right to Non-Discrimination, and the Right to Privacy, which altogether are considered the gatekeepers to other integral human rights. As the use cases of risk assessments, LAWS, automated medical diagnostics, and automated standards enforcement show, violations of these three principal rights often pave the way for infringement of other human rights standards. However, AI is not quite the disastrous, world-ending technology that movie franchises like *The Terminator* depict them to be. The continued advancement of AI technologies has also facilitated people's enjoyment of other fundamental rights: automated medical diagnostics have strengthened people's rights to adequate living standards and health, and algorithmic content moderation has bolstered individuals' right to expression.

Having said that, technical safeguards are still necessary to mitigate the more negative effects of AI/ML on human rights. To that end, blockchain has proven to be an effective means of ensuring privacy and securing the Right to Privacy over algorithmic applications. And although blockchain poses its own set of human rights challenges with respect to climate change, engineers and researchers have already devised ways of alleviating these issues. Furthermore, Procedural Regularity techniques have frequently been floated as a way to increase accountability and curtail algorithmic bias, thus defending the Right to Equality and Right to Non-Discrimination. However, Procedural Regularity does not completely resolve the problem of discrimination and prejudice in AI systems, and its techniques may potentially introduce new biases into these applications.

In general, AI/ML technologies will continue to proliferate modern society at extraordinary rates; they may soon dominate human technology, and there is unfortunately not much the average citizen can do to resist these radical shifts. However, individuals should take solace in the fact that researchers, subject matter experts, public officials, and industry leaders around the world recognize the consequences of AI/ML for human rights and that many will go to great lengths to minimize these negative effects. As the aforementioned technical frameworks demonstrate, engineers and scientists are hard at work developing technical accountability mechanisms that protect human rights from the harmful impact of AI. At the same time, domestic and international governments have shifted their focus to passing legislation that regulates the behavior of AI/ML applications such that they comply with existing laws and by extension, respect fundamental human rights. In any case, the average person may rightly be concerned about their basic rights in the face of AI, but they should rest assured that technical safeguards and concrete government regulations will protect them from the consequences of algorithmic systems.

References

[1] Anon. 2018. AI Open Letter. (February 2018). Retrieved January 28, 2021 from https://futureoflife.org/ai-open-letter/

[2] Eileen Donahoe and Megan M. Metzeger. 2019. Artificial Intelligence and Human Rights. (April 2019). Retrieved January 28, 2021 from https://www.journalofdemocracy.org/articles/artificial-intelligence-and-human-rights/

[3] Michael Crowe. 2019. Amazon says facial recognition can detect fear, raising concern for some privacy advocates. (August 2019). Retrieved January 28, 2021 from https://www.king5.com/article/news/local/amazon-says-facial-recognition-can-detect-fear-raising-concern-for-some-privacy-advocates/281-257712b2-5947-43b1-af93-37f064b7b902

[4] Masha Borak. 2021. Chinese people are concerned about use of facial recognition, survey shows. (January 2021). Retrieved January 26, 2021 from https://www.scmp.com/tech/innovation/article/3119281/facial-recognition-used-china-everything-refuse-collection-toilet

[5] Anon. Artificial Intelligence (AI). Retrieved January 28, 2021 from https://www.business-humanrights.org/en/big-issues/technology-human-rights/artificial-intelligence-ai/

[6] Ben Hartwig. 2020. The Impact of Artificial Intelligence on Human Rights. (May 2020). Retrieved 2021 from https://www.dataversity.net/the-impact-of-artificial-intelligence-on-human-rights/#

[7] Raso, Filippo, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Kim Levin. 2018. Artificial Intelligence & Human Rights: Opportunities & Risks. Berkman Klein Center for Internet & Society Research Publication.

[8] Lorna McGregor et al. 2019. HRBDT Report: The Universal Declaration of Human Rights at 70 - Putting Human Rights at the Heart of the Design, Development and Deployment of Artificial Intelligence. (December 2019). Retrieved 2021 from https://www.hrbdt.ac.uk/download/the-universal-declaration-of-human-rights-at-70/

[9] Lorna McGregor, Daragh Murray, and Vivian Ng. 2019. INTERNATIONAL HUMAN RIGHTS LAW AS A FRAMEWORK FOR ALGORITHMIC ACCOUNTABILITY: International & Comparative Law Quarterly. (April 2019). Retrieved 2021 from https://www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/international-human-rights-law-as-a-framework-for-algorithmic-accountability/1D6D0A456B36BA7512A6AFF17F16E9B6

[10] Anon. 2019. Policy and investment recommendations for trustworthy Artificial Intelligence | Shaping Europe's digital future. (April 2019). Retrieved 2021 from https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence

[11] Anon. 2021. IEEE Ethics In Action in Autonomous and Intelligent Systems: IEEE SA. (March 2021). Retrieved 2021 from https://ethicsinaction.ieee.org/

[12] Lorna McGregor. 2019. Accountability for Governance Choices in Artificial Intelligence: Afterword to Eyal Benvenisti's Foreword. (February 2019). Retrieved 2021 from https://academic.oup.com/ejil/article/29/4/1079/5320175?login=true

[13] Stephanie Weiser. 2020. Requirements of Trustworthy AI. (November 2020). Retrieved 2021 from https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1

[14] Ricardo Neisse, Gary Steri, and Igor Nai-Fovino. 2017. A Blockchain-based Approach for Data Accountability and Provenance Tracking. *In Proceedings of ARES '17, Reggio Calabria, Italy, August 29- September 01, 2017,* 10 pages.
DOI: 10.1145/3098954.3098958

[15] Anon. 2018. Key Issues. (October 2018). Retrieved 2021 from https://gdpr-info.eu/issues/

[16] ANGÈLE CHRISTIN, ALEX ROSENBLAT, and DANAH BOYD. 2015. DATA & SOCIETY; CIVIL RIGHTS: COURTS AND PREDICTIVE ALGORITHMS (2015).

[17] Anon. Universal Declaration of Human Rights. Retrieved 2021 from https://www.un.org/en/about-us/universal-declaration-of-human-rights

[18] Regina Surber. 2018. Artificial Intelligence: Autonomous Technology (AT), Lethal Autonomous Weapons Systems (LAWS) and Peace Time Threats. ICT4Peace Foundation and the Zurich Hub for Ethics and Technology (ZHET).

[19] Nathan Leys. 2018. Autonomous Weapon Systems and International Crises. *Strategic Studies Quarterly* 12, 1 (2018).

[20] Maaike Verbruggen and Vincent Boulanin. 2017. Mapping the development of autonomy in weapon systems. SIPRI.
DOI: OI:10.13140/RG.2.2.22719.41127

[21] Stuart J. Russell and Peter Norvig. 2021. *Artificial Intelligence: A Modern Approach*, Hoboken: Pearson.

[22] Peter Asaro. 2013. On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross* 94, 886 (June 2013), 687–709.
DOI: http://dx.doi.org/10.1017/s1816383112000768

[23] Mark R. Amstutz. 2018. *International Ethics: Concepts, Theories, and Cases in Global Politics*, Lanham: Rowman & Littlefield.

[24] Anon. Customary IHL - 1. Rules
Retrieved 2021 from https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1

[25] Jeroen van den Boogaard. 2016. Proportionality and Autonomous Weapons Systems. Opinio Juris (March 2016).

[26] Bonnie Docherty. 2021. The Need for and Elements of a New Treaty on Fully Autonomous Weapons. (March 2021). Retrieved 2021 from https://www.hrw.org/news/2020/06/01/need-and-elements-new-treaty-fully-autonomous-weapons#

[27] Anon. Retrieved 2021 from https://www.stopkillerrobots.org/about/

[28] Anon. 2020. TRUSTWORTHY AI IN HEALTH. Organisation for Economic Co-operation and Development (OECD).

[29] Eban Escott. 2017. What are the 3 types of AI? A guide to narrow, general, and super artificial intelligence. (October 2017). Retrieved 2021 from https://codebots.com/artificial-intelligence/the-3-types-of-ai-is-the-third-even-possible

[30] Sara Gerke, Timo Minssen, and Glenn Cohen. 2020. Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial Intelligence in Healthcare* (June 2020), 295–336.
DOI: http://dx.doi.org/10.1016/b978-0-12-818438-7.00012-5

[31] Emma Llansó , Joris van Hoboken, Paddy Leerssen, and Jaron Harambam. 2020. Artificial Intelligence, Content Moderation, and Freedom of Expression. The Transatlantic Working Group.

[32] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society 7, 1 (February 2020).
DOI: http://dx.doi.org/10.1177/2053951719897945

[33] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper Attention to Abusive User Content Moderation. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (2017).
DOI: http://dx.doi.org/10.18653/v1/d17-1117

[34] Mikolov, T, Sutskever, I, Chen, K, et al. (2013) Distributed representations of words and phrases and their compositionality. In: *27th Conference on Neural Information Processing Systems (NeurIPS)*, Lake Tahoe, NV, 5 (10 December 2013).

[35] Luke Conway. 2020. Blockchain Explained. (November 2020). Retrieved 2021 from https://www.investopedia.com/terms/b/blockchain.asp

[36] Deepak K. Tosh, Sachin Shetty, Xueping Liang, Charles Kamhoua, and Laurent Njilla. 2017. Consensus protocols for blockchain-based data provenance: Challenges and opportunities. *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)* (October 2017).
DOI: http://dx.doi.org/10.1109/uemcon.2017.8249088

[37] Jake Frankenfield. 2021. Smart Contracts: What You Need to Know. (March 2021). Retrieved 2021 from
https://www.investopedia.com/terms/s/smart-contracts.asp

[38] Elisa Bertino, Ahish Kundu, and Zehra Sura. 2019. Data Transparency with Blockchain and AI Ethics. *J. Data and Information Quality* 11, 4, Article 16 (August 2019), 8 pages.
DOI: https://doi.org/10.1145/3312750

[39] Guy Zyskind, Oz Nathan, and Alex 'Sandy' Pentland. 2015. Decentralizing Privacy: Using Blockchain to Protect Personal Data. *2015 IEEE Security and Privacy Workshops* (2015).
DOI: http://dx.doi.org/10.1109/spw.2015.27

[40] Anon. Cambridge Bitcoin Electricity Consumption Index. University of Cambridge Judge Business School Retrieved 2021 from https://cbeci.org/

[41] Camilo Mora et al. 2018. Bitcoin emissions alone could push global warming above 2°C. *Nature Climate Change* 8, 11 (2018), 931–933.
DOI: http://dx.doi.org/10.1038/s41558-018-0321-8

[42] Anon. 2015. Understanding Human Rights and Climate Change. Office of the High Commissioner for Human Rights.

[43] Johannes Sedlmeir, Hans Ulrich Buhl, Gilbert Fridgen, and Robert Keller. 2020. The Energy Consumption of Blockchain Technology: Beyond Myth. *Business & Information Systems Engineering* 62, 6 (2020), 599–608.
DOI: http://dx.doi.org/10.1007/s12599-020-00656-x

[44] Joshua A. Kroll. 2016. Accountable Algorithms (A Provocation). *The London School of Economics and Political Science* (February 2016).

[45] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2017. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology* 31, 4 (2017), 611–627.
DOI: http://dx.doi.org/10.1007/s13347-017-0279-x

[46] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu. Accountable Algorithms. *University of Pennsylvania Law Review*. 165 (2017), 633-705 .

[47] Laura Carter. 2019. Algorithmic accountability. (December 2019). Retrieved 2021 from https://www.hrbdt.ac.uk/algorithmic-accountability/