



LUNDS
UNIVERSITET

AI and Human Rights

— An explorative analysis of upcoming challenges

Joel Bergenfalk

Avdelningen för mänskliga rättigheter

Historiska institutionen

Kurskod: MRSG31

Termin: HT 2018

Handledare: Olof Beckman



Abstract

AI has a great deal to offer human civilization, but AI is also accompanied with a great deal of risks. This paper aims to give an introduction to the field of AI from a human rights perspective and to account for future challenges AI could bring forth. I will argue that these challenges could come to put pressure on the contemporary notion of human rights as formulated in various declarations, covenants, conventions, and protocols by international community. And that it might prove difficult for the current human rights doctrine to survive the test of time as a functional axiom for rights and protection of humans in a conceivable AI future. A broad picture of risks and challenges is presented by analyzing four different topics related to AI and human rights: consciousness, rights and agency, bias and discrimination, and socio-economic rights. Moreover, contemporary guidelines are shown to lack a clear and feasible agenda for how to deal with future human rights concerns. I claim that two main factors could come to condition the trajectory of rights: efficiency and human imperfection.

KEYWORDS: *AI, Artificial Intelligence, Human Rights, Utilitarianism, Guidelines, Robots, Future*

Contents

1.	Introduction	3
1.1	Statement of Issue and Background	5
1.2	Delimitations	6
1.3	Method	8
1.4	AI Today & AI of Tomorrow	8
2.	Theory of Human Rights	10
2.1	Utilitarianism and Fairness	13
3.	Analysis	17
3.1	AI Related Science and Discourse	17
3.1.1	Studies of the Mind & Consciousness	17
3.1.2	Rights & Agency for Robotics and AI	22
3.1.3	Bias and Discrimination with AI	25
3.1.4	Socio-Economic Effects of AI	27
3.2	Regulations and Guidelines on AI and Human Rights	29
3.2.1	Guiding Principles on Business and Human Rights	30
3.2.2	The Asilomar Principles	31
3.2.3	The Toronto Declaration	32
4.	Discussion	34
5.	Conclusion	35
	List of References and Literature	38

1. Introduction

The field of *Artificial Intelligence* (AI) has undergone a meaningful shift towards inclusion into mainstream science in the last decade, although often wrongly perceived by media and elsewhere as mere fiction, i.e. science fiction. This rather deceiving depiction results in movies such as *The Terminator* and *Space Odyssey* being the first that comes to mind when thinking about AI.¹ This flawed way of thinking stands in the way of a much needed discussion on AI, especially since groundbreaking advances in technology will most likely have a significant effect on all parts of society: jobs, politics, and everyday life. The need to think ahead of the problems and challenges of tomorrow before we reach a point of no return is pivotal. One problem with AI is the fact that there is a huge discord in the public knowledge about the risks and benefits of AI. First of all, as with the reference to the 80s movie *The Terminator*, AI should not primarily be thought of as human-like (killer) robots. This fallacy of anthropomorphism limits our thinking about what AI is and what it can achieve. Instead AI should be signified by the notion of an algorithm that needs no physical body to perform actions. Secondly, there are far more pressing matters to attend to, both with currently existing technologies and some soon to exist before humanity has to worry about AI acquiring their own sci-fi inspired agenda and ultimately seeking vengeance. Instead, attention and effort should be put towards understanding the impact and consequences of the fact that AI is already becoming an integrated and indispensable part of society, which is evidenced by the current usage of AI already in place in fields such as transportation, police enforcement, judicial systems, and health institutions.²

This paper aims to examine possible challenges to contemporary human rights in the not too distant AI future. I will map out current and coming challenges with AI and human rights, therefore a delimitation of what human rights entails is necessary. By human rights in general I

¹ The Terminator, tells of a humanoid killer-robot who travels back in time to end humanity. Space Odyssey, features the supercomputer Hal 9000 who controls the spaceship Discovery One, Hal 9000 starts disobeying the crewmembers and then tries to kill them one by one.

² See for example, OECD, OECD Digital Economy Outlook 2017, *OECD Publishing*, 2017.

mean the codified rights adopted by the international community in various declarations, covenants, conventions, and protocols. This because codified rights are the most widespread, accepted, and adopted proclamations of human rights. Paragraph 5 of the 1993 Vienna Declaration and Programme of Action declares that:

All human rights are universal, indivisible and interdependent and interrelated. The international community must treat human rights globally in a fair and equal manner, on the same footing, and with the same emphasis. While the significance of national and regional particularities and various historical, cultural and religious backgrounds must be borne in mind, it is the duty of States, regardless of their political, economic and cultural systems, to promote and protect all human rights and fundamental freedoms.³

This is central throughout the paper, because human rights as a framework for AI ethics might not have any real bearing in its current form; even adaptation and additional protocols could be insufficient. My concern, which I will explore in closer detail, regards the formulation of human rights by the international community in various declarations, covenants, conventions, and protocols. These will not survive the test of time as a functional axiom for rights and protection of humans in a conceivable AI future. Perhaps the right to privacy will not be compatible with health services depending on the mass collection of data to be able to make correct diagnoses. Or there might be a risk that the justice system would implement automation systems that could greatly increase effectiveness even if there is a risk of discrimination. The demand for effectivity could come to trump individual rights, if the benefits outweigh the risks, the risks could be overlooked. Committees and policy makers can try to formulate regulations on existing weak human rights legislations whilst knowingly ignoring the inherent issues that comes with it. If human rights doctrine is not compatible with AI development, then adding additional protocol and conventions that relates back to the incompatible doctrines will not solve the problem. Human rights doctrine will face many challenges in the near future, which it will need to respond to in order to remain salient. In a way so rights do not become excessive for their lacking of feasibility, meaning, and serious authority.

³ UN General Assembly, Vienna Declaration and Programme of Action, 1993, para 5.

1.1 Statement of Issue and Background

My purpose with this paper is looking at AI development and its impacts on different parts of society and everyday life, and to do so through a human rights perspective.

There is no shortcoming in documents provided by governments and organizations that discuss different challenges accompanying the progress of AI; there is a broad consensus that AI needs to work for the benefit of human society and be aligned with what is often vaguely called *our goals*. A common feature for some of these documents is to focus on business and market opportunities, while rights and safety regards comes in second. For example, the Swedish government's first effort to investigate AI was consequently from a market perspective. In the 178 page long report there is a lot said about economic gains and the weighty issue of staying relevant and competitive on an international market.⁴ But rights was only mentioned in passing, stating such things as:

New products and services based on AI will offer great advantages for society but only if the users can safely be assured that the systems are reliable and will not threaten our security and individual rights and democratic freedoms.⁵

I will elucidate further upon the phenomena of speaking about rights and freedoms in general and in loose terms and how that consequently leaves much room for interpretation. It is well known in most widespread human rights discourse that there is a long standing tradition of a given hierarchy of rights, favoring individuals over groups, ICCPR over ICESCR, and negative over positive rights. Thus, when governments, organizations, and corporations express their efforts to secure that *all* human rights are accounted for and proper risk mitigation will be in place

⁴ Vinnova, Artificiell intelligens i svenskt näringsliv och samhälle, N2017/07836/FÖF, 30 April, 2018.

⁵ Author's translation of original quote: "Nya produkter och tjänster baserade på AI kommer att erbjuda stora fördelar för samhället men bara om användarna kan vara trygga med att systemen är pålitliga och inte hotar vår säkerhet eller individens demokratiska fri- och rättigheter." - Vinnova, 2018, p. 57.

to avoid any discrepancies, it is unclear what this alleged commitment surmounts to. I am interested in what that effort entails, to what extent, and if there are any boundaries. As previously mentioned, by virtue of the 1993 Vienna Declaration and Programme of Action paragraph 5, all rights must be considered on equal footing.⁶ Hence there *should not* be a given hierarchy of rights for states and other actors to abide by and consult in dire straits. It is my belief that equal consideration of all rights is de facto far from realistic and that treating the rights problem of the foreseeable future in a *laissez faire* manner is not a valid option. When rights to protect privacy, right to work, leisure, and questions about personhood etc were formulated and codified, the context and prerequisites were vastly different from that of today. Therefore a revision of what rights entail and what their objectives are, is a reasonable requirement to counter new challenges that were inconceivable when the rights were proclaimed.

Maybe the solution surrounding AI and human rights will be a true egg of Columbus, nevertheless it will most likely require immense effort regardless. The conjuncture of AI and human rights could conceivably result in some sort of tradeoff, where a possible outcome might even be that human rights does not prevail at all. By mapping out the current state of AI from a human rights perspective I want to examine the challenges and benefits abundant with a continued technological progression. **The statement of issue therefore takes the following form:**

What challenges will AI pose to contemporary human rights doctrine?

1.2 Delimitations

For the purpose and scope of this paper to be as stringent and clear as possible a few disclaimers and clarifications are needed. The scope entails technologies already present today, and conceivable development for the near future. To clarify henceforth, what I mean by the abbreviation AI throughout this paper will first and foremost be that of weak AI, explained in

⁶ UN General Assembly, Vienna Declaration and Programme of Action, 1993, para 5.

more detail. Throughout this paper, the terms *future* and *near future* will carry the meaning of roughly around 20-50 years. There is of course an undeniable notion of uncertainty paired with speculations of the future, but hopefully my delimitations will make the topic less surreal and easier to grasp.

I will make claims and argue about AI within the scope of my field: human rights studies, thus leaving out the specific challenges of programming and code mathematics associated with AI. I argue that the humanities have a very important role and function in conjunction with AI today and in the near future. But it is still important to note that the endeavour for a prosperous AI future will have to be multidisciplinary. Realizing and acknowledging that even if the humanities side of academia could, hypothetically, present an ethically sound approach for AI in their interaction with sentient beings. That would still not necessarily make the required programming feasible in an adequate way. The famously cited *Three Laws of Robotics*⁷ from the Isaac Asimov's *I, Robot* is known as a common example of how the complexity of programming and code is simplified and thus generally understood as less intrinsically challenging. Thinking about the loopholes in this *law* and using the laws of robotics as framework for how to reason when trying to solve ethical dilemmas with AI-alignment is as helpful as Douglas Adams famous contribution the eternal question of the meaning of life, i.e., the number 42. It goes without saying that a language meant to define and limit conduct within a system of ethics, whether in the form of Asimov's laws, or by a binary system of 1's and 0's, is inadequate in providing a clear cut in every ethical problem.

Most sources and facts regarding technological breakthroughs, science, and philosophy in this paper will be centred around the United States and Europe. I am aware that this paper in large part leaves out other major tech and world powers, e.g., China. This is mainly due to the lack of reliable, available sources in English. This is unfortunate from a rights perspective because rights

⁷ The most famous exposition of the “law” of robots comes from Isaac Asimov's *I, Robot*, where he lays out the Three Laws of Robotics: (1) a robot may not injure a human being, or, through inaction, allow a human being to come to harm; (2) a robot must obey the orders given it by humans, except where such orders would conflict with the First Law; (3) a robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

ought to be supranational and universal. If rights are to be recognized and respected they ought to have hold throughout the international community. I will speak of rights, policies, and politics in the very broad sense, thus leaving out the complex attention that follows from attempts to incorporate different cultural, historic and religious idiosyncrasies.

1.3 Method

I will demonstrate that human rights doctrine could be challenged by AI by mapping out the current state of AI and what challenges and opportunities that comes with our current understanding of the field. I provide a brief background about AI, where I explain what AI is, and what it currently can and cannot do. My theory of human rights that I will present provides an alternative framework as an example of who human rights could respond to the proposed challenges. The analysis will first look at four AI and human rights-related topics important to grasp to scope of the impact AI could have on human rights. Second, three different contemporary regulations and guideline on AI and human rights will be examined. These guidelines attempts to combine AI with human rights in a harmonizing way, but as I will elucidate on further, they are not found sufficient. Thus aligning with my thesis that human rights, in its current form as formulated in several declarations, covenants, conventions, and protocols will not cut the mustard as long term background for the future with the emergence of future technologies.

1.4 AI today and AI of Tomorrow

The community of AI scholars differentiates between three types of AI: weak, strong, and super. This is important to limit any speculations of what the term AI entails when referred in the context of this paper, since the term is sometimes used very loosely, especially outside of academia. AI throughout this paper will first and foremost be that of weak AI, if not specified

otherwise. I will followingly make clear the on the distinctions between different types of AI and their use for my paper.

The first type of AI is called weak AI or narrow AI, it is the AI in technical products already widely available and in use by most of the population in affluent countries today. This type of AI is narrow in the sense that it is made to perform a limited task or limited set of task. Winning a game of chess, facial recognition, calculate consumers' shopping patterns or driving a car are all categorized as narrow skills within this terminology. For example, the algorithm developed by the company DeepMind was sufficient in beating the leading champions in the complex strategy game Go, which humans have collectively over 2000 years of experience playing.⁸ However impressive, AlphaGo cannot yet compete with the human mind in other cognitive challenges, in the sense that it can only perform well in the specific tasks it was designed for.

The second type of AI is general AI, AGI, or strong AI, which is the long term goal for AI-developers to be able to program. In contrast to weak AI present today, AGI is believed to be able to outperform humans in nearly any cognitive task.⁹ AGI is as of today not within reach, this however does not mean that it will not be a reality in the future. The truth is that it is unknown when such technology will be at our disposal and there is not consensus among the AI-experts. Although there are reasons not to dismiss speculations and planning for the possibility with human level intelligence in AI. For example, in a survey conducted at the 2015 Puerto Rico AI conference arranged by the Future of Life Institute, the average (median) answer to the question of when the participant thought AGI would be plausible was the year 2045, but some researchers guessed hundreds of years or more.¹⁰

Thirdly, Super AI, or Super intelligence is the expected last step in the AI evolution.

⁸ DeepMind, *The story of AlphaGO so far*.

⁹ Future of Life Institute, *Benefits & Risk of Artificial Intelligence*.

¹⁰ Future of Life Institute, *Benefits & Risk of Artificial Intelligence*.

An ultraintelligent machine could design even better machines; there would then unquestionably be an “intelligence explosion,” and the intelligence of man would be left far behind [...] Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.¹¹

Super AI is expected to trigger the future period, which Ray Kurzweil coined as *The Singularity*. Therein technological change will be so rapid, due to machines learning and creating even smarter machines, resulting in a seemingly endless upward spiral.¹² If or when this occurs, human life and our role in the universe as Earth’s omnipotent rulers will be forever changed as we are superseded by a new intelligent being. The possibility of Super AI should be considered with great precaution and circumspection. Realistically, however, Super AI is most probably outside my scope of up to fifty years, and will thus be excluded from the scope of my paper.

2. Theory of Human Rights

My paper concerns AI and human rights, it is interdisciplinary in character with focus on human rights. Based on the academic field of human rights studies, it analyzes a rights-future wherein AI is present. Thus I will propose an alternative theory of human rights that I believe to be more compatible with the challenges which I shall elucidate on in further detail. I will make use of a critical standpoint against the current rights system that I believe is flawed, this critical approach is based on James Griffin’s account of rights and some ethical aspects from John Broome’s *Weighing Lives*¹³. The theory proposed by Griffin goes against the notion of rights as indivisible and interdependent. Which is central for the approach I am suggesting as an alternative, in short rights should be viewed differently:

The best account of human rights will make them resistant to tradeoffs, but not too resistant.¹⁴

¹¹ Good, Irving, John, *Speculations Concerning the First Ultraintelligent Machine*, 1964, pp. 33-34.

¹² Kurzweil, Ray, *The Singularity Is Near*, Viking Penguin, 2005.

¹³ Broome, John, *Weighing Lives*, *Oxford University Press*, 2004.

¹⁴ Griffin, James, *On Human Rights*, *Oxford University Press*, 2008, p. 314.

This is, as he himself is quick to point out, a mere truism; however, it is part of a bigger picture that captures what I have in mind for the future of rights integration with AI. Given the current state where rights are supposedly sanctified by claims of our human dignity and all rights are to be emphasized on equal footing, in a fair manner, and without discrimination, a great share of complications follows. When the right to privacy, owning property, security of persons, and a right to health are all indispensable, interconnected, and ought to be understood as absolute rights, no legitimate derogations are possible. I argue that such a scenario is not feasible, and thus rights discrepancies will instead likely be dealt with arbitrarily or not considered at all. This is why rights instead need to be set in a more narrow sense. There are a lot of heavy-weight moral claims and obligations that will consequently have no connection to rights, but that could be seen as a good thing to make the rights more viable. Griffin points out that it is a mistake to think that just because we tend to see rights as especially important in morality, that we therefore also must make everything of especially moral importance into a right.¹⁵

There needs to be what Griffin calls *existence conditions* for human rights; Griffin gives two such conditions as necessary and sufficient for a basis of human rights, the primary being *personhood*:

Out of the notion of personhood we can generate most of the conventional list of human rights. We have a right to life (without it, personhood is impossible), to security of person (for the same reason), to a voice in political decision (a key exercise of autonomy), to free expression, to assembly, and to a free press (without them, exercise of autonomy would be hollow), to worship (a key exercise of what one takes to be the point of life). It also generates, I should say (though this is hotly disputed), a positive freedom, namely a right to minimum learning and material resources needed for a human existence, something more, that is, than mere physical survival. It also generates a right not to be tortured.¹⁶

Personhood here means that rights are grounded in an account of agency. There are negative rights, but with a fair mix of positive rights to make the enjoyment of autonomy meaningful. As I

¹⁵ Griffin, James, 2008, p. 318.

¹⁶ Griffin, James, 2008, p. 311.

will go over in more detail, the basis of personhood guarantees rights to what is essential to live a decent human life, although there are limits to what rights gives entitlement to. Griffin maintains that a clear boundary is set by the condition of personhood that implies that for example torture would not be acceptable no matter what other interests are at stake.

Griffin calls the second condition *practicalities*, since he asserts rights as not absolute, they must adhere to practical considerations compatible with the society the rights aims to function within. Human rights' existence must to some extent depend upon them being an effective, socially manageable, claim on others. Therefore practicalities constitute the necessary second ground.¹⁷

Human rights are to be understood in narrow sense as that which protects our status as normative agents, this agency can be divided into parts of autonomy, minimum provision, and liberty. These are not absolute and can be threatened by other rival valid claims that can infringe on the other agents' normative agency, which in turn can be threatened with varying degree. This is determined by the effect on the subjects personhood. When a collective welfare or rights goal stands in conflict with a subjects normative agency, a tradeoff is possible if the infringement does not require a severe loss of life-quality.¹⁸

Griffin's account of rights are for good reasons considered controversial in human rights studies. For one thing his focus on the capacity of normative agency greatly risks putting a threshold conception of human rights. Some of those who may be in most need of protection could very well be excluded from human rights protection by this account. Children, coma patients, the severely mentally handicapped, and even future generations are perhaps not given appropriate consideration. I choose to go with Griffin's account because of my proposed challenges with AI development and the current rights doctrine. Therefore a narrowing of rights could result in better overall enjoyment of human utility. If the option is that rights become redundant and overlooked altogether I believe that a practical and feasible system, even if it fail to encompass

¹⁷ Griffin, James, 2008, p. 315.

¹⁸ Griffin, James, 2008, pp. 30-31.

all challenges, is still preferable. And I believe that Griffin's theory provides one such practical and feasible alternative.

2.1 Utilitarianism and Fairness

I will base my theory on Griffin's rights approach with one clear difference, I will add a utilitarian reading to the theory because I believe that the lack of some sort of measurement between incommensurable rights will be an obstacle otherwise. Griffin is not a utilitarian, and has strong arguments against why his theory should be interpreted as such:

For instance, it seems to me that, if what is at stake is my being able to live by my reasonable conception of a good life, then no amount of mere upset and distress for members of my community could outweigh it; it would not matter if a hundred people were upset and distressed, or a thousand, or a million... [My doubt is whether we could perform the remendously large-scale, long-run, cost-benefit calculations that this approach requires, or even arrive at probabilities reliable enough for action].¹⁹

The utilitarian addon is motivated by the fact that a narrowing and clarification is needed for rights to stay relevant in policymaking of the future. What I will propose is a system that needs to be able to make utilitarian motivated sacrifices to a reasonable degree, as Griffin points out:

If my blood had some marvellous factor and a few drops painlessly extracted from my finger in a minute's time could save scores of lives, then, on the face of it, the personhood ground yields no right that needs to be outweighed. The prick of my finger would hardly destroy my personhood. But what happens if we up the stakes? Does my right to security of person not protect me against, say, the health authority that wants one of my kidneys? After all, the few weeks that it would take me to recover from a kidney extraction would not prevent me from living a recognizably human life either. Where is the line to be drawn?²⁰

¹⁹ Griffin, James, 2008, pp. 314; 322.

²⁰ Griffin, James, 2008, p. 315.

As previously stated, personhood guarantees that some violation of bodily integrity would not be justified even if a strict utilitarian theory would allow for it. Interest can be weighed against each other in a utilitarian way but only to a certain extent. I do not claim to know exactly where the line ought to be drawn, and giving a full account of a functional theory of rights is beyond the scope of this paper. But what I will work with throughout this paper is to propose a mindset allowing for a more narrow set of rights that are not necessarily absolute. My addition is that I propose that some sort of utilitarian measurement needs to be in place, inspiration for such a system could be something quite similar to the system of QALY (quality-adjusted life years). John Broome describes in *Weighing Lives* how QALY is used among other things in health economics and is a functional way of determining how to allocate limited resources by measuring, in this case the benefits of treatment in terms of ‘quality-adjusted life years’.²¹

A person's quality-adjusted life years are the number of years she lives, adjusted for their quality. ‘Quality’ refers to the quality of the person's health only. For example, a quality of life might be: in constant slight pain and unable to walk. Another might be: deaf. A year in good health counts as one qaly. A year in less good health counts as less than one qaly; its value is reduced by a ‘quality-adjustment factor’. If a particular quality has an adjustment factor of .7, a year of life at that quality would be valued at .7 of a qaly, equivalent to .7 of a year in good health. To calculate a person's qalys, we add up across the years of her life, counting each year at its quality-adjustment factor.²²

This will work as an inspiration for how a utilitarian model could be developed. For the proposed guidelines I believe there is a need to consider justice and fairness, otherwise the risk of unjustly maximizing utility for those who are already rather well-off in accordance with utilitarian efficiency is abundant.

The Rawlsian idea of prioritizing the worst off is a well known political theory of justice, however, when allocating resources this theory involves a risk: giving absolute priority to the worst off is accompanied by a dilemma known as *the black hole problem*. There may be people

²¹ Broome, John, 2004, p. 261.

²² Ibid.

who are very badly off, very expensive to treat or help and who can get very little benefit from it. If giving absolute priority to the worst off as a consequent rule, it seems like all available resources should go towards them even if they would get very limited benefit from it. This would not be cost-benefit efficient at all and might even be considered highly unfair. A fair middle ground would be aiming at giving priority to the worst of but not absolute priority.²³

Considering the *black hole problem*, I believe that fairness should be considered, but for the sake of bringing large benefits, fairness may be outweighed by other goods.²⁴ Broome gives the following example:

The WHO has a project to distribute AIDS medicine to very large numbers of people in the poorer countries of the world, although it will treat a great many people suffering from AIDS it recognizes that it doesn't have the resources for treating everybody who needs treatment or would benefit from treatment, so from the time that it started this project, the WHO recognizes that there was an issue of what was the right way to decide or distribute medicine amongst the people who need it, and I think there are two criteria; the criterion of goodness which roughly is the number of people that you save, and there is the criterion of fairness which is distributing these resources in the appropriate way across the people who need them. And those two aims are really in conflict, a good part of the cost of delivery is getting access to the people who need it, some people you can get to quite easily, these are people who live in the cities, but there are people who live in the remote rural places which it's very hard to get to, so if you spend a lot of money on getting to those people you end up with less to spend on the medicines, so you have to balance the number of people you treat against the fairness in distributing the medicines, and I don't think either of those criteria dominates, I think sometimes we should sacrifice some unfairness for the sake of doing a greater amount of good with the available resources.²⁵

I believe this is compatible with Griffin's notion of *practicalities*: that practical consideration for the societal context must be accounted for in order to achieve the best outcome. The theory for this paper is constructed to be practical, as opposed to contemporary human rights doctrine. My

²³ Wolff, Jonathan, Jonathan Wolff on Political Bioethics, *Philosophy Bites*, 2012.

²⁴ See for example, Broome, John, 2004, p. 38.

²⁵ Broome, John, John Broome on Weighing Lives, *Philosophy Bites*, 2008.

approach is therefore suited to function as a template for future rights challenges by providing a system wherein rights are strong, but not inalienable and absolute.

There is a lot of available criticism to be made against utilitarianism, some of the most common involve well constructed, but unrealistic, thought experiments.²⁶ Even with real life situations considered, the objective here is not to formulate a waterproof theory, but to have a theoretical framework in which flexible and pragmatic choices can be made. I do not think any theory will ever be able to stand against every hypothetical problem. Therefore it is more important that the future of human rights is feasible in practice than theoretically good on paper.

In summary:

1. Human rights needs to be narrowed, the validity of right claims is determined by the effect on personhood.
2. Rights are grounded in personhood and agency and should be resistant to claims infringing on these, but not too resistant.²⁷
3. Rights are not absolute, reasonable tradeoffs are acceptable, but with limitations to their extent. Some violation of one's personhood will be considered unacceptable, but as with giving up some bodily integrity in the earlier example of giving blood will not.
4. A utilitarian value system, QALY or any analogous theory as a plausible option, should be used to help with large scale policy making to maximize benefits for humans.
5. Fairness and justice needs consideration, but should not be viewed with intrinsic value. Effectiveness can reasonably outweigh these.

These guidelines are not perfect, nor are they meant to be so. They are on purpose rather basic, to stand a better chance of meaningful implementation but still includes an attempt to consider weighing fairness within utility allocation. With the current unsatisfactory state regarding rights

²⁶ The list includes various *Trolley Problems* and different scenarios with atrocities towards some are weighed against utility for the larger quantities of others. Problems like this are meant to be hypothetical to test the limits of various moral theories.

²⁷ Supra note 11.

and its discourse, a change that would bring more good, estimated increased happiness, pleasure, and utility could be seen as an improvement.

3. Analysis

3.1 AI related science and discourse

In this section I will cover research in AI discourse relevant to human rights. First *Studies of the Mind and Consciousness* regards some of the major questions surrounding AI, breakthroughs in this field could possibly change the way we relate to machines entirely. This leads to the topic of *Rights and Agency for Robotics and AI*, covers the possibility of including machines and AI in the rights arena, and liability issues with AI when there is no human in control. *Bias and Discrimination with AI* showcase problems regarding non-representative data sets and the need for transparency within AI systems. Lastely, *Socio-Economic Effects of AI*, briefly considers some work and theories regarding AI and the socio-economic impacts this might bring, which could have a substantial carryover to current human rights doctrine.

3.1.1 Studies of the Mind and Consciousness

In this section, the question of what constitutes consciousness, and how it relates to AI will be examined. What it means to think and if a machine can be said to have a mind is connected to how we as humans ought to relate to AI, as it becomes an indispensable part of our everyday life. As will be further explored in the section *Rights and Agency for Robotics and AI*, the question of consciousness is closely related to the notion of agency, which is an important part of human rights discourse.

In academia, our understanding of the human brain, and consciousness in particular, is one of the most puzzling obstacles in natural science and philosophy alike. This Gordian Knot is indeed

very complicated and far from a sound solution or any sort of overarching consensus.²⁸ Even though there have been steady progress in cognitive sciences and neuroscience that help us better understand human behavior and the processes behind it, the *very hard problem* is to explain why and how physical processes of particles in the brain give rise to subjective conscious experiences.²⁹ The spectra of consciousness wherein the biological mind made of neurons is at one end and that of a silicon based or a simulated one is at the other keeps gradually getting blurred. I will give a philosophical background to some prominent theories of computer consciousness, by presenting four important scholars connected to the field of AI. I will elucidate on consciousness and the question of whether machines can *actually* think or not. For the sake of fair representation, two philosophers defending the capacity for AI's consciousness will be presented with two corresponding proponents of opposing views.

Alan Turing is perhaps one of the most famous names in computer science and was well ahead of his time in thinking about a future with intelligent machines and computers. In his paper written in 1950 *Computing Machinery and Intelligence* he rejected the importance of whether machines can think or not. Instead he focused on if machines have the capability of passing a behavioral intelligence test, named after himself the famous *Turing Test*. The test consists of a program or a computer to have a conversation via typed text messages with a corresponding human for at least five minutes, the human is tasked with telling if the conversation is with a program or another human person. The machine is said to have passed the *Turing Test* if it can consequently fool the interrogator 30% of the time. In the 1950s it was Turing's belief that by the year 2000 computers would surely be powerful enough to be able to pass this test, as of today, no proven case exists of a program passing the criterias.³⁰

As I will come back to in more detail, some philosophers claim that regardless whether a machine is capable of passing the *Turing Test* or not, proves nothing more than it being capable of simulating a process that is perceived as thinking.

²⁸ Chalmers, J. David, The Meta-Problem of Consciousness, *Journal of Consciousness Studies*, 25, No. 9–10, 2018, p. 7.

²⁹ Chalmers, J. David, The Meta-Problem of Consciousness, p. 6.

³⁰ Russell, J. Stuart, & Norvig, Peter, *Artificial Intelligence, A Modern Approach*, 2010, p. 1021.

Not until a machine could write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but know that it had written it.³¹

The argument here is that a conversation or a piece of art that is the result of an algorithmic process has no connection to the real world. Turing refuted this by pointing out that it is impossible to have any direct evidence about the internal processes and mental states of other humans either. There is no way to be certain that people around you, that you encounter everyday have consciousness. This courtesy of assuming is just extended to people without much thought and Turing believed that the same could one day also be true for machines.³²

Turing furthermore argued that in everyday life we take comfort in the convention that assumes that everyone around us are in fact thinking despite no test or clear evidence supporting such claim. Thus it is for Turing clear that we would and should extend this convenient convention to machines if we are experiencing them as acting in an intelligent way. The following dialogue is an example of Turing's point, which has become a well known part of AI academics tradition.³³

HUMAN: In the first line of your sonnet which reads “shall I compare thee to a summer’s day,” would not a “spring day” do as well or better?

MACHINE: It wouldn’t scan.

HUMAN: How about “a winter’s day.” That would scan all right.

MACHINE: Yes, but nobody wants to be compared to a winter’s day.

HUMAN: Would you say Mr. Pickwick reminded you of Christmas?

MACHINE: In a way.

HUMAN: Yet Christmas is a winter’s day, and I do not think Mr. Pickwick would mind the comparison.

MACHINE: I don’t think you’re serious. By a winter’s day one means a typical winter’s day, rather than a special one like Christmas.³⁴

³¹ Jefferson, Geoffrey, cited in Russell, J. Stuart, & Norvig, Peter. *Artificial Intelligence, A Modern Approach*, 2010, p. 1026.

³² *ibid.*

³³ *ibid.*

³⁴ *ibid.*

For some philosophers, this is far from convincing. Machines that processes data and acts according to given instructions proves nothing more than that the machine is capable of simulating a process perceived as thinking. Philosopher John Searle strongly advocates that no matter how impressive and developed an AI is, its capacity to *actually* think, in a strict ontological sense, will never be a reality.

No one supposes that a computer simulation of a storm will leave us all wet ... Why on earth would anyone in his right mind suppose a computer simulation of mental processes actually had mental processes?³⁵

Searle is the inventor of the famous philosophical thought experiment *The Chinese Room*, which according to Searle proves that a program cannot actually *understand*. In the hypothetical experiment you ought to imagine a system fully capable of passing the previously mentioned *Turing Test*. The system consist of a human who can only speak English being in a concealed room, the human has at his disposal a rule book, or instructions that are written in English. Outside the room is a Chinese speaking human who through a hole in the wall sends in a slip of paper with Chinese characters on it. Inside the room the English speaker then finds matching symbols in the rule book, follows the instructions and writes down the correct symbols on a new slip of paper and passes it back to the person on the other side. So even if it appears for the person outside that the other is a fluent Chinese speaker, the person inside the room does not have knowledge or understanding of the symbols the output yields and therefore does not understand Chinese. The rule book and stacks of paper are just pieces of paper and thus have no understanding either, which leads to Searle's point being that running the right program does not generate *real* understanding.³⁶

This thought experiment have received much critique, philosopher Daniel Dennett has said that *The Chinese Room* is an oversimplification that at best plays on human intuition. Calling the

³⁵ Russell, J. Stuart, & Norvig, Peter. 2010, p. 1027.

³⁶ Russell, J. Stuart, & Norvig, Peter. 2010, p. 1031.

scenario an *intuition pump* favorable only because it is conceived in such a way that Searles conclusion wrongfully seems to be the only possibility; and at worst that it simply shows that John Searle does not understand computers.³⁷ Since it relies on intuition rather than proof the very same argument could be made about the human brain as well, thus failing to answer the *very hard problem* of why a hunk of brain can be a mind while a hunk of liver can not.³⁸

Hubert Dreyfus is a philosopher that has been very outspoken about his disbeliefs around AI and the fundamental limitations that separates a machine, however smart, from a human's cognitive ability. Dreyfus is famous for his critic of AI in his paper *What Computers Can't Do* written in 1972 and a follow-up 20 years later, *What Computers Still Can't Do*. His main critique can be explained, simply put, that AI is suffering greatly from something called the qualification problem, which is the inability to capture everything necessary and sufficient in a given set of logical rules. Dreyfus maintains that human behavior is too complex to be captured in rules, saying that human behavior is in some sense determined by rules, but these rules are a sort of holistic context or background. This could be described as humans have a *direct sense* of how things operate in different scenarios.³⁹ He gives the example of appropriate social behavior in giving and receiving gifts:

Normally one simply responds in the appropriate circumstances by giving an appropriate gift.⁴⁰

This *direct sense* of what counts as appropriate in a certain context cannot be properly explained, therefore Dreyfus concludes, that it could not be feasibly programmed either.⁴¹

There is a plethora of work on this topic and therefore hard to give an all encompassing picture, but the big picture I want to mediate is that there is no side of the argument that clearly has the upper hand. It is hard to say whether our relationship to machines will change as their abilities

³⁷ Dennet, Daniel, Daniel Dennett on the Chinese Room, *Philosophy Bites*, 2013.

³⁸ Russell, J. Stuart, & Norvig, Peter. 2010, p. 1033.

³⁹ Russell, J. Stuart, & Norvig, Peter. 2010, p. 1024.

⁴⁰ Ibid.

⁴¹ Ibid.

improves and they grow harder to distinguish from humans. It is difficult to say if moral consideration would come from machines acquiring a greater intellect alone. Given the fact that machines are already “smarter” than most humans, but no one seems to regard their calculator as worthy of moral consideration even though it greatly out performs most in mathematical tasks. Then again being sentient, defined as being able to have goals, experiencing pain and pleasure, and the ability to form social connections has not stopped humanity from viewing non-human animals as greatly inferior, treating most of them as mere means to an end.

To summarize, the question if machines *really* can think and feel, or rather if they will be able to do so in the future is far from agreed upon. Some thinkers firmly believe that it is utterly inconceivable, while others believe it is possible, but science is just not there yet. There is still too much that is unknown about the human brain and our own consciousness to say anything for certain. Nonetheless, if machines reaches the threshold of being classified as sentient beings, it could come to be a pivotal moment in human history.

In the next section I will explore how machines being smarter, able to make decisions and possibility of consciousness affects human rights, moral consideration, and legal liability.

3.1.2 Rights and Agency for Robotics and AI

As of today, there are no serious efforts or large influential organizations working with AI and robots to be a serious contender to receive rights status. But given to efforts demonstrated below, there is an apparent worry surrounding the question that should be taken seriously nonetheless. There are arguments in academia, stemming from the same conclusions as shown in the previous section, that moral consideration can be plausibly defended even if it is not regarded as a practical possibility by states and institutions of the world.⁴²

⁴² Schwitzgebel, Eric & Garza, Mara, A Defense of the Rights of Artificial Intelligences, *Midwest Studies In Philosophy*, September 15, 2015, p. 1-44. See also Phil McNally and Sohail Inayatullah, THE RIGHTS OF ROBOTS, Technology, culture and law in the 21st century, *Futures*, April, 1988.

There are possible artificially intelligent beings who do not differ in any morally relevant respect from human beings. Such possible beings would deserve moral consideration similar to that of human beings. Our duties to them would not be appreciably reduced by the fact that they are non-human, nor by the fact that they owe their existence to us. Indeed, if they owe their existence to us, we would likely have additional moral obligations to them that we don't ordinarily owe to human strangers – obligations similar to those of parent to child or god to creature.⁴³

This has very limited impact to the overall rights discourse. Moral arguments aside there is a largely shared view that rights should not, at least not in the near future be extended to include other non-biological beings, the belief is that it would greatly undermine the rights discourse. For example the IEEE (Institute of Electrical and Electronics Engineers) stated that:

For the foreseeable future, A/IS should not be granted rights and privileges equal to human rights:
A/IS should always be subordinate to human judgment and control.⁴⁴

Also an open letter initiative targeted towards the European Union has been signed and endorsed by AI and robotics experts, industry leaders, law, medical, and ethics experts confirming that they do not believe that legal or rights status is appropriate for robots or AI.⁴⁵ Worth noting is that within the European rights doctrine, human rights are already extended to all legal entities and not limited to human beings exclusively, thus making rights for AI sound more plausible within this context. However the open letter initiative maintains that the legal status for a robot cannot derive from the Legal Entity model currently in place, since it implies the existence of human persons behind the legal person to represent and direct it. This is not always the case for a robot, especially with further progress making AI more independent and autonomous.⁴⁶

On the 25th of October 2017 the social humanoid robot Sophia made by the company Hanson Robotic was granted citizenship in the Kingdom of Saudi Arabia, becoming the first robot in the

⁴³ Schwitzgebel, Eric & Garza, Mara, 2015, p. 3.

⁴⁴ The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, Version 2, IEEE, 2017, p. 23.

⁴⁵ <http://www.robotics-openletter.eu/> The open letter has 285 signatories at the moment of writing (27/11).

⁴⁶ Ibid

world with full citizenship.⁴⁷ Even though, as a Forbes journalist pointed out; this might primarily be a publicity stunt that bears little meaning other than to generate headlines and put Saudi Arabia on the map for future innovation investment deals.⁴⁸ Sophia was nonetheless later named UNDP's first-ever Innovation Champion, making her the first non-human with a UN title which might be looked back upon as an historic moment in the future.⁴⁹

Another issue is the pressing matter of the legal status of AI. Having legal status and thus liability could be seriously considered since legal status is already extended to non biological, metaphysical entities such as corporations. This idea is discussed with regards to self driving cars. David C. Vladeck wrote in *Washington Law Review* on this topic suggesting that it is reasonable to imagine that the law could evolve to bestow *personhood* on machines in the same way that it has done just that for corporations and certain trusts.⁵⁰

A machine that can define its own path, make its own decisions, and set its own priorities may become something other than an agent. Exactly what that may be, though, is not a question that the law is prepared to answer.⁵¹

For now, he argues that the best option would be to construct a system of strict liability, meaning that the car would not be seen as an agent but carry its own insurance that could amongst different solutions be built in with the price from the manufacturer.⁵² There are only a few known cases where self driving cars, or semi-self driving cars, have caused injuries or crashes with fatal outcome.⁵³ Excluding the more common accidents where the self driving car was involved, but not at fault, like being rear-ended by a human driver. The first reported fatal crash was a Tesla self driving car in 2016 where the autopilot sensors on the the vehicle failed to distinguish a

⁴⁷ Stone, Zara, "Everything You Need To Know About Sophia, The World's First Robot Citizen", *Forbes*, 2017.

⁴⁸ *Ibid*.

⁴⁹ Risse, Mathias, *Human Rights and Artificial Intelligence An Urgently Needed Agenda*, 2018, p. 4.

⁵⁰ Vladeck, C. David, *Machines Without Principals: Liability Rules and Artificial Intelligence*, *Washington Law Review*, 2014, p. 124.

⁵¹ Vladeck, 2014, p. 145.

⁵² Vladeck, 2014, p. 146.

⁵³ Chang, Lulu and Dormehl, Luke, "6 self-driving car crashes that tapped the brakes on the autonomous revolution", *Digital Trends*, 2018.

white tractor-trailer crossing the highway against a bright sky, resulting in a fatal crash.⁵⁴ In March 2018 a Uber self driving car had an accident that was the first case involving the death of a pedestrian, this could have set precedent for more cases to come but was instead settled with unknown terms, supposedly to avoid the negative attention.⁵⁵ But as more autonomous cars hit the road and human control diminishes, more cases will probably appear in the near future. One thing that is import to note is that the probability of an accident with a self driving cars is negligible in comparison with the overall risks involved with human drivers. For example in the US alone, car accidents accounts for roughly 40.000 deaths per year.⁵⁶ So even with faults and errors, and the legal implications that needs a solution, AI technology could have a huge positive impact in lowering the current death rates.

3.1.3 Bias and Discrimination with AI

Bias within AI can refer to namely two things: the first is bias in statistic, which refers to statistical errors that can result from the use of incomplete or otherwise skewed data to train the AI system; the second is the more normative understanding of bias as judgement based on prejudices. *Selection bias* is among other things commonly exemplified with facial-recognition software being over representative of one particular group.⁵⁷ Because of the narrow data the AI has been trained with, the AI is vastly better at recognising the faces of white people as opposed to others. Here the statistical bias is shown by the varying levels of accuracy built in from *bad* training data, but the outcome could be said to be the result of impartial evaluation of the given facts.⁵⁸ This is only true in part, the AI itself cannot really be said to have committed a biased or discriminatory decision.

⁵⁴ Yadron, Danny & Tynan, Dan, "Tesla driver dies in first fatal crash while using autopilot mode", *The Guardian*, 1 July, 2016.

⁵⁵ Woodall, Bernie, "Uber avoids legal battle with family of autonomous vehicle victim", *Reuters*, 2018.

⁵⁶ Bureau of Transportation Statistics, *Transportation statistics annual report*, 2017.

⁵⁷ Campolo, Alex, et al, *AI Now 2017 Report*, The AI Now Institute, 2017, p. 13.

⁵⁸ Breland, Ali, "How white engineers built racist code", *The Guardian*, 4 Dec, 2017.

When examining technical systems, there can be a temptation to, or vested interest in, limiting discussion of bias to the first more ‘neutral’ statistical sense of the term. However, in practice there is rarely a clear demarcation between the statistical and the normative definitions: biased models or learning algorithms, as defined statistically, can lead to unequal and unfair treatments and outcomes for different social or racial groups.⁵⁹

What the AI *knows* and learns is based on what data it is given, and the humans behind it are the ones responsible for the input provided. Therefore it is quite hard to distinguish between the first and second notion of bias, since both are rather similar and the output can have negative, discriminatory effects regardless.⁶⁰ These problems are well known, and need to be addressed as machine learning technologies becomes more frequently used by institutions and elsewhere. One contemporary example of discrimination in AI systems was the use of COMPAS⁶¹, a risk-assessment tool that was used to make sentencing decisions in courts across the United States. The software does calculations based on a set of scores derived from 137 questions that are either answered by defendants or pulled from criminal records. And proceeds to give a rating of 1-10 for the estimated risk that the defendant would commit future crimes.⁶²

Researchers at Propublica released a report wherein they analyzed more than 7.000 cases to see how many of the defendants reoffended within two years. They found that white defendants were mislabeled as *low risk* more often than blacks, noting that black defendants were wrongfully labeled at twice the rate as opposed to whites defendants. The pilot project for COMPAS was launched in 2001, and today it is one of several similar programs used in the United States. The specific calculations used for the system are not available because they are protected as proprietary information.⁶³

A dangerous aspect of applying AI in society is that if it is launched prematurely, and without proper testing a process of trial and error have real, negative effects on people. At their best, well

⁵⁹ Campolo, Alex, et al, *AI Now 2017 Report*, The AI Now Institute, 2017, p. 14.

⁶⁰ Ibid

⁶¹ Correctional Offender Management Profiling for Alternative Sanctions (COMPAS).

⁶² Angwin, Julia, Larson, Jeff, and Kirchner, Lauren “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks,” *ProPublica*, May 23, 2016.

⁶³ Ibid.

constructed AI systems can potentially be used to augment human judgement and reduce both our conscious and unconscious biases. The examples above however points to the importance of transparency within AI development. It will remain important to perform proper due diligence and secure reliable AI that can be subject to critical review and testing, and that the discriminator practices are not overlook for convenience.

3.1.4 Socio-Economic Effects of AI

With continued progress of AI intelligence there is a potential for helping and enhancing human civilization to flourish like never before. Hopefully, not only people with capital and wealth but also workers around the world whose hard and demanding labor would no longer be needed could greatly benefit from automation and integration of AI. Some of these benefits are of course conditioned on a alternative to the lost, no longer needed occupations, or rather a substitute for the loss of income. This pressing matter will need a sound solution, and as I will go over, AI could have an impact on most known occupations. Beyond the transportation industry and physical labor; lawyers, doctors, bankers, and other *white collar* workers could face a staggering decline in the demand for their skills as well.

One analysis conducted by the McKinsey Global Institute, wherein 750 jobs were observed, concluded that roughly 45% of paid activities could be automated using already available technology.⁶⁴ A more recent report from the same institute found that 30% of *work activities* could be automated by 2030 and up to 375 million workers worldwide could be affected by emerging technologies.⁶⁵ But even with these alarming numbers most economists and public policy analyst routinely dismiss the idea of technological unemployment and embrace the idea that new innovations will and have always created new jobs, these jobs might very well be unimaginable today in the same way no one could foresee most *white collar* occupations some

⁶⁴ Schiller, Ben “How Soon before Your Job Is Done by a Robot?” *Fast Company*, January 6, 2016.

⁶⁵ Manyika, James et al, “Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation,” *McKinsey Global Institute*, December, 2017.

100 years ago.⁶⁶ This is because the fear of new technology leading to mass unemployment has not yet been justified, but it is reasonable to expect that this time could be different. The labor force participation rate in industrialized countries climbed from World War Two until the 2008 recession, but has since been declining overall.⁶⁷ In part because most affluent countries outsource a lot of work to low income countries, but also because productivity has been shown to steady increase with more investment in machines rather than human labor. The increased productivity has vastly improved corporate profits but has not given incitement to re-invest the increased profits to create new jobs or better wages for the workers.⁶⁸

Rising use of robots and artificial intelligence in advanced countries is likely to create benefits large enough that they could be used to compensate those workers who are substantially negatively affected for their lost wages.⁶⁹

The benefits to compensate for lost wages are of course dependent on the optimistic view point that the world's economic system will be able to adapt to some new or revised version where the now millions of people without an occupation have some sort of income or other means to get by. One frequently suggested solution is some sort of implementation of universal basic income, James Hughes, the Executive Director of the Institute for Ethics and Emerging Technologies argues followingly:

Policies that are being proposed to protect or create employment will have only a temporary moderating effect on job loss. Over time these policies, which will impose raise costs, lower the quality of goods and services, and lower competitiveness, will become fiscally impossible and lose political support. In order to enjoy the benefits of technological innovation and longer, healthier lives we will need to combine policies that control the pace of replacing paid human labor with a universal basic income guarantee (BIG) provided through taxation and the public ownership of wealth.⁷⁰

⁶⁶ Hughes, J. James, A Strategic Opening for a Basic Income Guarantee in the Global Crisis Being Created by AI, Robots, Desktop Manufacturing and BioMedicine, *Journal of Evolution and Technology*, February, 2014, p. 46.

⁶⁷ Hughes, J. James, 2014, p. 47.

⁶⁸ Ibid.

⁶⁹ Stevenson, Betsey, AI, Income, Employment, and Meaning, *Economics of Artificial Intelligence*, 2018, p. 1.

⁷⁰ Hughes, J. James, 2014, p. 45.

A paradigm shift in the world's economic structure would inevitably affect all social, economic and cultural rights. For example, it is difficult to grasp what will happen to the right to work and right to have leisure if there is no need for work and a surplus of leisure being the new norm. As I have sketched out, there are a myriad of challenges with AI and socio-economic rights, not the least with the concern for employment. But also, what is a good threshold for education, and moreover what is to be taught to people who are not fit for the limited selection of jobs? This might be more of a political challenge and a difficult task for policy makers world wide. I think that it is fair to expect that AI will have a noteworthy effect on the socio-economic side of human rights. However it is a very complicated task to speculate *how* socio-economic rights will respond to this. Socio-economic human rights, rather than political and civil human rights (PC), are formulated in terms of sufficient thresholds, while PC-rights are more absolute with written exceptions and limitations. Therefore these rights might not be incompatible *per se*, but rather outdated. The threshold could therefore be set higher or simply change focus to better fit a new paradigm.

3.2 Regulations and Guidelines on AI and Human Rights

In the follow section I will elucidate on some written regulations and guideline concerning AI in conjunction with human rights. Since there is a plethora of documents to choose from I have done a selective picking of three documents that I believe fit the scope of this paper. I have chosen these three on the basis that they have clear connections and association with human rights doctrine, with the exception of the Asilomar AI Principles that has the distinct quality of being a good representation of the forefront of AI-expertise. Which I believe provides valuable insight to the agenda of some of the most influential people associated with the development of AI-technologies. The other two are the UN's Guiding Principles on Business and Human Rights

⁷¹, and Amnesty International & Access Now's Toronto Declaration⁷² are elaborated upon further below.

3.2.1 Guiding Principles on Business and Human Rights

The document is referenced in *The Report of the Special Rapporteur of the promotion and protection of the right to freedom of opinion and expression*,⁷³ Stating that it serves as good guidelines for future developments and integration of AI.

Companies also have responsibilities under human rights law that should guide their construction, adoption and mobilization of artificial intelligence technologies (A/HRC/38/35, para. 10). The Guiding Principles on Business and Human Rights: Implementing the United Nations “Protect, Respect and Remedy” Framework provide a “global standard of expected conduct for all businesses wherever they operate” (principle 11), including social media and search companies. To adapt the conclusions from the Guiding Principles to the domain of artificial intelligence (ibid., para. 11), the Guiding Principles require that companies, at a minimum, make high-level policy commitments to respect the human rights of their users in all artificial intelligence applications (principle 16); avoid causing or contributing to adverse human rights impacts through their use of artificial intelligence technology and prevent and mitigate any adverse effects linked to their operations (principle 13); conduct due diligence on artificial intelligence systems to identify and address actual and potential human rights impacts (principles 17–19); engage in prevention and mitigation strategies (principle 24); conduct ongoing review of artificial intelligence-related activities, including through stakeholder and public consultation (principles 20–21) and provide accessible remedies to remediate adverse human rights impacts from artificial intelligence systems (principles 22, 29 and 31).⁷⁴

⁷¹ United Nations Human Rights, Office of the High Commissioner, Guiding Principles on Business and Human Rights, HR/PUB/11/04, 2011.

⁷² Amnesty International & Access Now, The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems, RightsCon, 2018.

⁷³ UN General Assembly, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/73/348, 2018.

⁷⁴ UN General Assembly, A/73/348, 2018, p. 10.

Suffice to say, with these principles present today, there is still a myriad of problems and ethical concerns regarding human rights and corporations. Which means that the hopes for them to make a promising foundation for more advanced technology and far more complicated technological development is not near satisfactory. My concern is that attempt like these will only be wishful thinking at best. It is good to encourage that companies ought to conduct due diligence, and engage in prevention and mitigations strategies. But I believe that the scope of commitments by corporations to *avoid causing or contributing to adverse human rights impacts*⁷⁵ will be severely limited by corporate competitiveness. To conduct proper due diligence is costly and time consuming, which means that if a competitor skips a few steps and launches its product or service faster. Then the company who tried commit to the *Guiding Principles* is likely to lose market shares and revenue. This is clearly already a possibility, but I think it is reasonable to expect an escalation of scenarios like this coming with new opportunities that AI will bring in the near future.

3.2.2 The Asilomar AI Principles

The Asilomar Principles are the result of a conference with some of the world's leading experts within the field of AI.⁷⁶ Being the first of its kind and the result of extensive work and international cooperation the 23 principles are by their very existence impressive in their own right. Tackling important questions of transparency and research funding which is a good step in the right direction. A selection of these principles relevant to my paper, are:

Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.⁷⁷

Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.⁷⁸

⁷⁵ UN, HR/PUB/11/04, 2011, para. 13 (a).

⁷⁶ Future of Life Institute, Asilomar AI Principles, 2017.

⁷⁷ Future of Life Institute, Asilomar AI Principles, 2017, para 7.

⁷⁸ Future of Life Institute, Asilomar AI Principles, 2017, para 8.

Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.⁷⁹

AI Arms Race: An arms race in lethal autonomous weapons should be avoided.⁸⁰

The private sector might in fact play a more prominent role than state actors in the future role of AI, which is why these principles are to be seen as a good example of aspirations of transparency and international cooperation. But they lack any real enforcing or coercive power to make good on the trajectory it aims for. The vague wording leaves room for many different interpretations, for example that an arms race in lethal autonomous weapons *should* be avoided, but not the weapons in particular.⁸¹ The word *should* implies that there could be just cause for an arms race given the right circumstances. The strive for AI to be compatible with ideals of human dignity, rights, and freedoms seems prima facie good. The fact that rights and freedoms are mentioned at all is of course better than a sole focus on economic and technological aspects. Nonetheless this principle unfortunately entails nothing much more than the simplistic conclusion that it is better for AI to do good rather than bad. It does not provide the framework for which rights, values, dignities and freedoms it is addressing nor their definitions.

3.2.3 The Toronto Declaration

The Declaration was written by Amnesty International and Access Now and presented at RightsCon Toronto in May 2018, it has since been endorsed by other influential organizations such as Human Rights Watch and Wikimedia Foundation. They point out that human rights law provide a solid base for establishing a ethical framework for machine learning and other aspects of future technologies.⁸² The goal is to lay the groundwork that ensures that systems of machine learning are created and used in ways that respect rights and non discrimination, particularly with

⁷⁹ Future of Life Institute, Asilomar AI Principles, 2017, para 11.

⁸⁰ Future of Life Institute, Asilomar AI Principles, 2017, para 18.

⁸¹ Supra note 80.

⁸² Amnesty International & Access Now, The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems, RightsCon, 2018, pp. 1-3.

regards to vulnerable groups in society.⁸³ The declaration demands that actors in the private sector should perform and follow a due diligence framework with special regard to human rights, to avoid discrimination and violation of human rights. When the risk is deemed too high or impossible to mitigate, private sector actors should refrain from using the assessed technology in that context.⁸⁴ The conclusion is that discrimination in machine learning systems should be avoided where possible, ending on the note that:

Technological advances must not undermine our human rights. We are at a crossroads where those with the power must act now to protect human rights, and help safeguard the rights that we are all entitled to now, and for future generations.⁸⁵

Once again the framework of contemporary human rights doctrine by which these NGOs operate leaves much room for interpretation. The fact that they affirm that Human Rights are *universal, indivisible and interdependent and interrelated*,⁸⁶ makes it less clear on what these recommendations actually mean other than the states must try their very best, which is already well established. If they stand by this fact, then the right to health, privacy, non-discrimination and leisure are all equal and absolute. If technologies in the full sense cannot or *must* not undermine human rights, including the right of future generations.⁸⁷ There is a prominent risk for either stagnating future development or what I believe is more plausible, for a lot of human rights to be overlooked all together in favor of development and economic gains.

Regulations on AI have the potential to accomplish what regulations on nuclear technology did not. To ensure that possible benefits are well distributed and utilized, and that risks and dangers of AI are well managed. But it will require immense efforts and transparent cooperation to do so. This can prove to be difficult since there is a clear discord between and within different states with various contradicting interpretations and reservations regarding scope and meaning of human rights. Making human rights far from one clear thing that states undertake to respect,

⁸³ Ibid p. 6.

⁸⁴ Ibid, pp. 12-13.

⁸⁵ Ibid, p. 16.

⁸⁶ Ibid, para 2.

⁸⁷ Ibid.

protect, and fulfill. Consequently, the multitude of rights in various conventions present today, cannot all have the property of *jus cogens*.

4. Discussion

This paper has aimed to give an entry to what AI is, how it relates to and affects human rights, and to show that current human rights doctrines, policies and guidelines regarding the future will face difficulties with the complexity of upcoming challenges.

The challenges contemporary human rights are facing does not begin with AI, but as I have analyzed they could become more apparent and consequently harder to ignore with continued technological progression. Therefore I proposed one possible alternative human rights theory aimed at being more compatible with coming paradigm shifts. A theory wherein rights are narrowed, non-absolute, and an addition of some form of utilitarian value system to give practical framework for contradicting right claims. Even with an added emphasis on fairness and justice, a full account on *how* to limit discrimination, consider fairness and justice, and many other challenges were left out of this paper. AI and Human rights urgently needs a clear agenda and I believe that means the field must be open to multi- and interdisciplinary collaboration. Much more research and efforts are needed for this topic, not the least regarding things that were mostly or completely left out of this paper; autonomous weapons, surveillance, AGI and more in depth studies of AI in specific areas are all welcomed for further research.

Better and more advanced AI will necessitate increased involvement from the humanities, the need to study and understand its impacts could prove to be a undeniably pivotal task for the near future.

5. Conclusion

Attempts to work out sufficient principles for human rights and AI have as of yet been unsatisfactory, this is due to several factors including uncertainty, lacking feasibility, and the discord regarding what constitutes human rights and their debated possible hierarchy. The notion of rights as being universal, indivisible, interdependent and interrelated is already questionable at best and could have dire consequences if it means that rights become overlooked and not seriously considered as a result. AI integration in the near future is likely put more pressure on this notion and could therefore make the infeasibility of contemporary rights doctrine more apparent.

My findings throughout this paper supports two major factors concerning AI that will prove challenging for human rights, namely: **Efficiency** and **Human imperfection**.

Efficiency: The extreme benefits AI could bring to labor, politics, health care, and everyday life will become too irresistible for society to pull the brakes on and therefore soon impossible to reverse. In the same way as the spread and integration of common internet usage is not withheld by the fact that the internet allows for proliferation on such things as trafficking and child pornography. Even though it is widely considered that these violations are among the most heinous crimes imaginable and ought to be stopped at all cost, there are still no serious efforts to dissolve and cut back on the world wide web. Similarly once society and policy makers grow custom to the huge improvements and comfort brought forth by AI, efforts to stagger its progress will be more difficult to implement. As AI helps societies to improve and flourish in various aspects, the rational and effective, but crude way to utilise might not be to altruistically care for the wellbeing of all mankind. If automation is effective enough, the common practice of outsourcing jobs to the poorest countries of the world becomes redundant, and a large amount of people consequently become economically superfluous. If the global economy shifts from a capitalist market wherein every human at least can have the minimum value of a consumer. A

huge portion of the world's population could consequently come to be of no substantial value to the affluent and powerful. Evidently the notion of the inherent human dignity is not sufficient to end current world poverty, discrimination, rights violation, and suffering in the world at present. Thus it could prove unlikely that the near future, without some major interference, will be any different. This is not the same as simply tolerating negative aspects of AI development and regarding them as inevitable, rather it means that rights trajectory could come to be conditioned by the strive for efficiency.

Human Imperfection: AI can outperform humans in almost all narrow tasks, and since there is a possibility that the cognitive limitations of machines will decrease, the line that currently separates human abilities from that of AI will gradually blur. This could put serious pressure on our current understanding of rationality as an axiom for morally valid claims and challenge human superiority. Moreover, Human bias whether it be motivated by race, sex, nationality or other factors is present in all levels and functions in society. Humans are imperfect, the notion that we have are capable of making rational, coherent and just decisions will be challenged by AI. There could be an exponential increase in the awareness of our flaws and prejudices as they will unavoidably surface at a vastly increased rate as AI mirrors human bias on account of being a man-made creation. Discrimination could therefore be harder to ignore if the discriminatory outcome is traceable in the AI's configuration, systems in place would need to be transparent and subject to critical review. This could either give opportunity to alter these systems to mitigate the negative consequences or alternatively the strive for efficiency will trump the risks of reproducing discriminatory practices and systems instead remains opaque. If increasing implementation of AI within the justice system results in cases being handled at a much lower cost and a fraction of the time, more cases could consequently be tried for people seeking justice. It is known that prejudice and bias towards different groups of people already affect the outcome of rulings made by human judges and juries in general. Therefore it is unclear if AI reproducing this tendency, just at a much more efficient rate will be compelling enough to be cause for urgent actions for change.

All speculations of the future are coupled with a unavoidable level of uncertainty, most events and outcomes are difficult or impossible to predict. Technological development and history does not follow a teleological trajectory and thus preliminary precautions are sometimes futile. The point being that the future is uncertain, therefore a static rights framework might need flexibility.

List of References and Literature

Amnesty International & Access Now, The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems, RightsCon 2018 Toronto, Canada, 16 May, 2018

Angwin, Julia, Larson, Jeff, and Kirchner, Lauren “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks,” *ProPublica*, May 23, 2016, found at:

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Betsey Stevenson, “AI, Income, Employment, and Meaning”, *Economics of Artificial Intelligence*, 2018

Breland, Ali, "How white engineers built racist code", The Guardian, 4 Dec, 2017, found at: <https://www.theguardian.com/technology/2017/dec/04/racist-facial-recognition-white-coders-black-people-police>

Broome, John, John Broome on Weighing Lives, Philosophy Bites, June 29, 2008, found at: <https://philosophybites.com/2008/06/john-broome-on.html>

Broome, John, *Weighing Lives*, Oxford University Press, 2004

Bureau of Transportation Statistics, *Transportation statistics annual report, 2017*. available at: <https://www.bts.gov/sites/bts.dot.gov/files/docs/browse-statistical-products-and-data/bts-publications/transportation-statistics-annual-reports/215386/2017-tsar-ch6.pdf>

Campolo, Alex, et al, *AI Now 2017 Report*, The AI Now Institute, 2017, available at: https://assets.ctfassets.net/8wprhvnpc0/1A9c3ZTCZa2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report_.pdf

Chalmers, J. David, The Meta-Problem of Consciousness, *Journal of Consciousness Studies*, 25, No. 9–10, 2018, pp. 6–61

Chang, Lulu & Dormehl, Luke, "6 self-driving car crashes that tapped the brakes on the autonomous revolution", *Digital Trends*, 22 June 2018

David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, Washington Law Review, Vol. 89:117, 2014

Dennett, Daniel, Daniel Dennett on the Chinese Room, Philosophy Bites, June 23, 2013, found at: <https://philosophybites.com/2013/06/daniel-dennett-on-the-chinese-room.html>

DeepMind, The story of AlphaGO so far, found at: <https://deepmind.com/research/alphago/>

Future of Life Institute, *Benefits & Risk of Artificial Intelligence*, found at: <https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/>

Good, Irving, John, *Speculations Concerning the First Ultraintelligent Machine*, Trinity College, Oxford, England and Atlas Computer Laboratory, Chilton, Berkshire, England, 1964

Griffin, James, *On Human Rights [Elektronisk resurs]*, Oxford University Press, 2008

Harari, Noah Yuval, *Homo Deus* (Audiobook), *Random House*, 8 September, 2016

Hughes, J. James, A Strategic Opening for a Basic Income Guarantee in the Global Crisis Being Created by AI, Robots, Desktop Manufacturing and BioMedicine, *Journal of Evolution and Technology*, Vol. 24, Issue 1, February, 2014, pp. 45-61

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, Version 2. IEEE, 2017, found at: http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

Manyika, James, Lund, Susan, Chui, Michael, Bughin, Jacques, Woetzel, Jonathan, Batra, Parul, Ko, Ryan & Sanghui, Saurabh, "Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation," McKinsey Global Institute, December, 2017

McNally, Phil & Inayatullah, Sohail, THE RIGHTS OF ROBOTS, Technology, culture and law in the 21st century, *FUTURES* April 1988

OECD, *OECD Digital Economy Outlook 2017*, OECD Publishing, 2017, available at: <https://espas.secure.europarl.europa.eu/orbis/sites/default/files/generated/document/en/9317011e.pdf>

Privacy International & Article 19, *Privacy and Freedom of Expression In the Age of Artificial Intelligence*, April, 2018

Risse, Mathias, Human Rights and Artificial Intelligence An Urgently Needed Agenda, Carr Center for Human Rights Policy, Harvard Kennedy School, 79 JFK Street, Cambridge, MA 02138, May, 2018

<http://www.robotics-openletter.eu/> (last visited 2019/01/09)

Russell, J. Stuart, & Norvig, Peter, *Artificial Intelligence, A Modern Approach*, Third Edition, Chapter 26, 2010, pp. 1020-1043

Schwitzgebel, Eric & Garza, Mara, Department of Philosophy, University of California at Riverside, Riverside, CA 92521-0201, September 15, 2015

Stone, Zara, "Everything You Need To Know About Sophia, The World's First Robot Citizen", *Forbes*, 11 July, 2017

United Nations Human Rights, Office of the High Commissioner, Guiding Principles on Business and Human Rights, HR/PUB/11/04, 2011

UN General Assembly, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/73/348, 29 August 2018

UN General Assembly, Vienna Declaration and Programme of Action, 12 July 1993, A/CONF.157/23, available at: <https://www.refworld.org/docid/3ae6b39ec.html>

Vinnova, Artificiell intelligens i svenskt näringsliv och samhälle, N2017/07836/FÖF, 30 April, 2018

Wolff, Jonathan, Jonathan Wolff on Political Bioethics, *Philosophy Bites*, June 11, 2012, found at: <https://philosophybites.com/2012/06/jonathan-wolff-on-political-bioethics-originally-on-bioethics-bites.html>

Woodall, Bernie, "Uber avoids legal battle with family of autonomous vehicle victim", *Reuters*, 29 March, 2018 found at: <https://www.reuters.com/article/us-autos-selfdriving-uber-settlement/uber-avoids-legal-battle-with-family-of-autonomous-vehicle-victim-idUSKBN1H5092>

Yadron, Danny & Tynan, Dan, "Tesla driver dies in first fatal crash while using autopilot mode", *The Guardian*, 1 July 2016, found at:
<https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk>