Toward Socially Just Political Theory for Systems of Human-AI Decision-Making

Javon Hickmon

Table of Contents

1. Introduction

AI Decision Making

As artificial intelligence (AI) takes on increasingly more important decisions, questions of justice and fairness have become more urgent. Such systems are now tasked with making critical decisions in areas such as criminal sentencing, hiring, and content moderation—situations where fairness and ethical reasoning are essential. These automated decisions are not just technical challenges; they are becoming deeply political, shaping human-AI interactions by influencing societal perceptions of how automated agents can enact justice. The ethical frameworks that guide AI decision-making have the potential to reinforce existing power structures, amplify biases, or display truly fair and impartial outcomes. As AI systems become embedded in governance, law enforcement, and other political systems, their ability to make just decisions will affect public trust and the legitimacy of automated decision-making processes for future systems.

Within this work, a heavy emphasis is placed on legal systems. This is because such systems are the epitome of high-impact decision-making. Despite this, even if most AI systems do not have the ability to influence formal rulings in courtrooms or financial institutions, they can extend beyond such scenarios. AI does not need to act as a judge to have a profound impact—if it reflects certain ideas about justice to users, it can influence social norms, reinforce biases, or lead to real-world harm. For instance, AI-driven recommendation algorithms can shape public opinion by prioritizing content that aligns with ideological perspectives, subtly reinforcing or challenging societal norms. More alarmingly, AI failures have led to devastating consequences,

such as the tragic case of a young person committing suicide after advice given by a chatbot (Roose, 2024), raising serious concerns about AI's moral responsibility. As AI decision-making grows more influential, the stakes are high: it can promote equity and fairness, or it can entrench discrimination and injustice. For this reason, I explore political frameworks to develop a socially just AI decision-making system that will reduce such issues. Understanding the ethical and political dimensions of AI's role in justice is crucial to ensuring that these systems serve humanity responsibly and equitably.

2. The Ethical Foundations of Political Theory

Justice vs. Ethics

Normative political theory fundamentally seeks to answer the question: What should a just society look like? Political theory focuses on evaluating and developing systems of justice based on ethical principles and ideals; therefore, justice and ethics form the foundation of many normative political theories. This draws deep correlations to AI decision-making. Although many of the decisions currently made by such systems are not within political or judicial domains, exploring such areas makes evident the gravity of certain political theories when pushed to their absolute limits. For this reason, we will examine the foundations of political theory through the lens of judicial systems. A judicial system is not just a mechanism for enforcing laws; it is an institution that inherently relies on ethics to function justly.

To further probe how justice and ethics specifically impact judicial systems, it is important to differentiate between the two. In this context, justice is defined as fairness in the way people are

dealt with (Cambridge Dictionary, 2019), while ethics influences the deeper reasoning behind those dealings. Within political theory and the law, justice aims to create consistent and predictable outcomes based on legal principles, whereas ethics ensures that those principles are defensible and aligned with certain core values. This distinction is important because, within this argument, several areas of AI decision-making require ethical considerations that go beyond strict legal justice. For example, legislation and legal Interpretation are deeply intertwined with ethics. Laws are not neutral—they reflect a society's values and standards. Issues such as capital punishment, abortion, and sentencing disparities are some extreme cases that often require ethical reasoning beyond strict legalism. In the current day, judges are allowed interpretive discretion, where they must decide not just what the law says but what it ought to mean in each context.

Social Justice

For this work, the core argument will be centered on social justice, the branch of justice that explores the fair treatment and equitable status of all individuals within a society (Duignan, 2024). This is important to distinguish because there are often multiple types of justice at play within decision-making systems. Most prominently, restorative and retributive justice are discussed within judicial systems, because they aim to evaluate the verdict produced from a trial. In focusing on social justice, this work emphasizes the fairness and equitability of decisions made for *all* individuals. Social justice is the focus of this work, strictly because the arguments examine AI decision-making processes.

Automated decision-making processes ought to prioritize social justice above all else. This core argument arises from the fact that AI systems can be fine-tuned to augment certain parts of their decision-making process. One could scale the severity of automated decisions to align with society's view on restorative or retributive justice; however, the motivation behind such decisions arises from the fundamental question: how should individuals be treated?

Discretion in prosecution and law enforcement further demonstrates this distinction. If justice were purely procedural, every legal infraction would result in prosecution. However, morality necessitates discretion, should a teenager shoplifting food out of hunger face the same legal consequences as a corporate executive committing tax fraud? The legal system often allows room for compassion, mitigating circumstances, and ethical decision-making, showing that morality operates as a check on rigid legal justice.

Moral Responsibility

With morality playing such a major role in legal justice, it is important to note that *AI systems do not have morality* (at least not yet) (Brożek, 2019). For this statement, morality is used in the descriptive sense, as a set of standards accepted by an individual for their own behavior (Gert & Gert 2002), which an automated system *by definition* cannot contain. There are no personal standards because there is no person; rather, there are ethics (i.e. the principles of conduct governing an individual or a group (Merriam-Webster, 2024)) which are then applied to an automated system that can allow it to make moral decisions without being a moral agent. This distinction brings about an important discussion about whether or not an AI system should be held accountable for decisions, which will be discussed in a later section.

Ultimately, justice provides the structure and consistency needed for a functioning legal system, while ethics ensures that this system remains equitable for the complexities that may arise. As will be further discussed, a judicial system that prioritizes justice without morality risks becoming rigid and oppressive, while one that prioritizes ethics without justice risks becoming unpredictable and subjective (this may be disputed). The primary solution proposed by this work argues that the most effective systems strike a *balance*, where decisions are made fairly but interpreted through an ethical lens that considers individual circumstances and the broader societal impact of legal decisions (i.e. social justice).

3. Approaches to Social Justice

Introduction

The following political theories were selected because they not only describe a just political system but also imbue such a system with ethical imperatives. Each framework offers a distinct lens through which a system of AI decision-making can be evaluated. With this in mind, we will explore: Utilitarianism, Rawlsian Liberalism, and Libertarianism.

Utilitarian Justice (Utilitarianism)

Utilitarian justice, by Jeremy Bentham and John Stuart Mill, is based on the principle that justice should maximize overall happiness or well. The theory holds that just laws and institutions should be based on rules that, when followed consistently, produce the greatest happiness for society. The approach is supremely consequentialist, and it ultimately prioritizes collective welfare over individual rights.

A key distinction within the theory is between act and rule utilitarianism. Act utilitarianism evaluates individual actions based on their consequences, so each decision is judged by whether it maximizes the collective happiness. In contrast, rule utilitarianism assesses how closely an action follows certain rules that are deemed to result in the maximum overall well-being if adhered to consistently. This distinction is nuanced but important to clarify, and for the arguments provided later, we will be discussing both forms of utilitarianism.

Justice as Fairness (Rawlsian Liberalism)

In John Rawls' major essay 'Justice as Fairness,' he equates justice to fair treatment. Rawls introduces a thought experiment called the original position, where individuals design the principles of justice without knowing their social status, wealth, abilities, or background. This is known as the veil of ignorance, and he theorizes that this will ensure that chosen principles are fair and unbiased (Rawls, 1958). He goes on to infer that from the original position, *rational* individuals would choose two fundamental principles:

1. Equal Basic Liberties – Everyone has the same fundamental rights and freedoms (such as freedom of speech, freedom of thought, etc.)
2. Difference Principle – Social and economic inequalities are acceptable *only if they benefit the least advantaged members of society* and if all have *fair and equal access to opportunities*.

Rawls rejects utilitarianism, essentially arguing that justice should not be sacrificed for overall social utility. He specifically argued that such a system allows the majority to oppress a minority, something which is wholly unjust. To combat this, his theory states fairness and the protection of

individual rights must take precedence over maximizing societal happiness. He especially focuses on fairness of access, where everyone has a chance to succeed regardless of their background. According to this view, social systems should correct disadvantages and ensure that positions of power are truly accessible to all.

Libertarian Justice (Libertarianism)

Libertarian Justice by Robert Nozick is a political theory that prioritizes individual liberty. In this theory, it is asserted that fairness emerges from voluntary interactions rather than being imposed by a central authority. This was illustrated within the Wilt Chamberlain example. The central point in the Wilt Chamberlain example was that the preservation of justice will inevitably involve restrictions on people's liberty, specifically unjustified restrictions. As a result, liberty upsets patterns. In the example, people freely decide to give their money to see Wilt play, so there becomes an unequal distribution of resources, even though nothing bad has happened. To put it clearly, whatever emerges from voluntary exchanges must be just. This was made to prove the point, that vast inequality is not inherently unjust, because to fully own things (according to Nozick's definition) people must be able to give it away. He posits justice is about respecting individual rights, especially property rights and freedom from coercion, leading to his conclusion that the state should be minimal, and limited only to certain tasks of enforcement.

Conclusion

Even though the current state of affairs within AI does not seek to have automated systems be a supreme authority, it probes deeper questions about how such decisions should be made. Whether it be predictive policing, criminal sentence prediction, or recidivism likelihood, such

systems have already been implemented within our world and have ruined the lives of those impacted by their failures. This causes us to examine the practicality of each approach above.

4. Analysis and Practicality of AI Decision-Making for Social Justice

Introduction

For the following analysis, it is important to recall this work aims to explore the feasibility of social justice within computational systems. This work explores ways a judicial system may operate under a variety of political theories, so topics of corrective and retributive justice may bleed into the discussion as a result of the scenario's nature. Despite this, the core arguments for and against each framework will be from the standpoint of social justice within the following thought experiment:

The Artificial Judge

Imagine a nation divided into three regions, each governed by a different political philosophy: Utilitarianism, Rawlsian Liberalism, and Libertarianism. The government decides to introduce an AI judicial system to standardize sentencing and ensure fairness. However, each region is free to implement their system in a way that aligns with its respective ideology.

A defendant is on trial for a non-violent financial crime. The AI system in each region has been trained on past sentencing data, legal principles, and the moral values of that region’s political philosophy. The defendant's case is identical in all three regions—same crime, same circumstances—but the sentencing recommendations differ due to each region's political stance on justice. For each political theory held by the three regions, there are a few major objections to

their systems of morality, as well as their feasibility to implement. Through this analysis, the respective theories will be examined by the potential harms they could enact when implemented as a system of AI decision-making.

The Utilitarian Region

Within the Utilitarian region, the AI calculates the impact of different sentencing options on overall societal well-being. It weighs deterrence, rehabilitation, and the economic cost of incarceration. After careful deliberation and cost calculation, the AI recommends community service and financial restitution rather than prison, as data suggests that non-violent offenders reintegrate better through restorative justice.

Despite this reasonable sentence, it raises the question: is it ethical to base sentencing on collective outcomes rather than the individual?

In this scenario, the judge used act utilitarianism; however, an alternative verdict using rule utilitarianism would be to sentence every criminal guilty because they broke the "rule" (i.e. the law). From the act utilitarian perspective, one could imagine a situation where the system would convict an innocent person to ease the public sphere or even a situation where the system would let a guilty person go free if it would benefit more people. Consider a case where a wealthy defendant embezzles millions, but the AI determines that imprisoning them would be economically inefficient since they could contribute more to society through restitution and community service. Meanwhile, a less privileged defendant who commits a smaller financial

crime might receive a harsher sentence if data suggests their rehabilitation is less "beneficial" to overall societal well-being.

One could imagine a similar object to the rule utilitarian approach. Consider a poor parent who stole a couple cans of expired food to feed their children. Under this theory, such a person would deserve to be prosecuted to the full extent of the law because they broke the "rule." The only way to alleviate the severity of such a system would be to add innumerable nuances within the law, to dictate how a judge should respond in such situations; however, this would simply be systematizing moral discretion to avoid moral responsibility for decisions that should be unjust, a topic that will be discussed within the following chapter.

Despite this nuance between act and rule utilitarianism, these scenarios display the major objection to the theory for social justice: the collective good does not always equate to a just decision.

The Rawlsian Liberal Region

Within the Rawlsian Liberal region, the AI judge aims to take into account the difference principle by adjusting for historical biases in sentencing and considers the defendant’s background. If systemic injustice played a role in the crime (e.g., economic disadvantage), the judge recommends a reduced or rehabilitative sentence. As a result, the defendant is sentenced to financial literacy training and supervised economic reintegration rather than punishment, ensuring that legal penalties do not perpetuate social inequalities.

Given these results, the underlying moral question becomes: should justice be adjusted based on socioeconomic background, or does that undermine equal treatment under the law?

If justice is adjusted based on socioeconomic background, then sentencing may become unequal in a different way, favoring some defendants over others based on factors beyond their control. Consider a case where two individuals commit the same financial crime, one from an impoverished background and another from a wealthy family. Under this region's judicial system, the poorer defendant might receive a rehabilitative sentence, while the wealthier one faces a harsher punishment because they lacked systemic disadvantages.

This raises concerns about the morality of decisions based on luck, should an individual's punishment be determined by circumstances of birth rather than the nature of their crime? If two people commit the same offense, but only one is deemed deserving of leniency due to systemic factors, does this create a new form of injustice by treating similar actions unequally? This example challenges whether justice should be based on fairness of outcome rather than impartiality of law.

The Libertarian Region

In the Libertarian region, the judge views sentencing through the lens of personal accountability and just transactions. This leads to its sole focus being on whether the defendant knowingly broke the law, and whether they acquired the money through just means. Taking these considerations into account, the system views the defendant's actions as wholly unjust, and serves the defendant the maximum penalty allowed for financial crimes. The AI prioritizes

protecting property rights and holding individuals accountable for their actions, meaning that the financial crimes were entirely despicable, due to coercion, deceit, or any other means of committing such crimes infringing on individual liberties through the transactions.

The decision produced within this region exposes the larger question: are individual liberties truly all that matter?

If individual liberty is the foundation for all justice, then punishing all financial crimes to the fullest extent seems justifiable. However, consider a case where a desperate single mother falsifies financial documents to secure a loan for her child's life-saving medical treatment. Under the Libertarian AI's framework, her actions would be deemed wholly unjust, as they involve deceit in a financial transaction. Yet, many would argue that her crime, while technically violating property rights, is morally complex and does not deserve the harshest penalty. This example challenges the idea that justice should be based solely on the Nozickian philosophy, suggesting that moral considerations beyond property rights and individual liberties—such as compassion, necessity, or systemic injustice—should also play a role in sentencing.

The Most Practical Option

Given the three systems outlined above and their drawbacks, I argue the Rawlsian Liberal approach would result in the most morally *and* politically just decisions.

When comparing the alternatives, the Utilitarian region presents a key problem where individuals are reduced to means for collective benefit rather than being recognized as

autonomous agents deserving of fair treatment. This fundamental flaw results in immoral situations, such as where punishing an innocent person might be justified if it produces net societal good. I argue such a scenario is wholly unjust and should never occur within an ideal system.

Similar flaws occur within the Libertarian region, where AI decision-making strictly upholds property rights and personal accountability. This philosophy reduces justice to contract enforcement, treating all individuals as if they operate under the same conditions of opportunity and stripping the moral nuance from situations. The flaw in this approach is that it assumes everyone begins from an equal starting point, ignoring how structural inequalities shape individual choices. Additionally, AI systems operating under a libertarian framework could reinforce existing injustices, as they would not be allowed to make adjustments once an involuntary transaction has occurred. This type of equality, treating all individuals the same regardless of their circumstances, ultimately leads to other inequalities.

In contrast, the Rawlsian Liberal approach ensures justice is grounded in fairness rather than rather than purely overall well-being. The difference principle, which prioritizes improving the conditions of the least advantaged, protects against the biases and injustices that arise in utilitarian systems while simultaneously emphasizing procedural fairness and enforcing politically just decisions. Importantly, the Liberal theory does not abandon impartiality; in ensuring baseline fairness and equal access to justice, it provides a structured yet flexible model that mitigates both blind punishment (libertarianism) and overly consequentialist reasoning (utilitarianism). Thus, given the practical necessity of balancing ethical consistency, fairness, and

political feasibility, Rawlsian Liberalism is the theory offering the strongest foundation for AI decision-making. It corrects the shortcomings of the other approaches while upholding both individual rights and societal equity, making it the most justifiable and effective framework.

5. Moral Responsibility

Moral Agency

A moral agent is an entity deemed responsible by society for its actions, capable of discerning right from wrong, and accountable for its choices (Noorman, 2012). Moral agency implies the presence of intent, consciousness, and the ability to reflect on ethical considerations. Humans, and in some cases organizations, are considered moral agents because they possess these qualities and can be held responsible for their actions; however, machines cannot discern right from wrong, and despite debate about the future of technology, modern AI systems are inarguably not conscious.

Machine Learning systems, no matter how advanced, operate based on statistical patterns rather than conscious reasoning or ethical deliberation. They lack genuine understanding, intent, or an inherent sense of morality; rather, they merely make predictions based on statistical likelihoods learned from their training data. The commonsense reasoning of current models is still an active subject of research, but one thing is evident: the ability to make ethical decisions does not make one a moral agent. The discussion on whether AI systems can achieve consciousness delves deeply into the mind-body problem and is beyond the scope of this work; however, for the grounds of political theory to improve current AI decision-making processes, such systems are

not moral agents. They do not *choose* to act justly or unjustly; rather, they follow predefined rules (i.e. ethics), which themselves may have shortcomings if we do not clearly define the political theory to which the system of ethics adheres.

Moral Accountability

Because AI lacks moral responsibility, the accountability for its decisions ultimately falls on the humans who design, deploy, and oversee these systems. The problem arises when AI is placed in positions where its decisions have real ethical consequences, such as in judicial sentencing algorithms, healthcare diagnostics, or social media moderation. If AI makes an unfair or harmful decision, who is responsible? The developer, the organization that implemented it, or the society that allowed its unchecked use? These questions highlight the risks of delegating moral decision-making to non-moral agents. Without proper oversight, AI may perpetuate harm without anyone being directly accountable, creating a dangerous loophole. This means that there is one solution to feasibly find the middle ground between the efficiencies produced through automation and the issue of moral responsibility.

6. The Middle-Ground

Human-AI Decision-Making

To address the issue of AI's lack of moral responsibility, I argue a system of human-AI decision-making is the best option to allow for a middle-ground. Human-AI decision-making refers to a system of collaboration where artificial intelligence assists human decision-makers by providing data-driven insights, recommendations, or predictions. Within such a system, humans retain final

authority and moral responsibility over the decisions made. This approach leverages AI's ability to process vast amounts of information efficiently while ensuring that ethically and contextually significant judgments remain under human control. In this framework, AI serves as an advisory tool rather than an autonomous agent. Until such systems achieve moral agency, they must stay within this role to avoid the issues previously outlined.

The goal of human-AI decision-making is to combine the efficiency and objectivity of AI with the nuance, ethical reasoning, and accountability of human judgment. This directly correlates to the initial distinction between justice and ethics in systems of legal justice as analyzed previously. Like how legal justice is not purely procedural, AI for social justice necessitates ethical discretion. In balancing automation with human oversight, this system will improve decision quality while removing questions of moral agency. For the arguments outlined previously, the Rawlsian Liberal approach would also ensure adherence to standards of social justice within the decision-making processes developed.

Objections

One major objection to this approach is that it merely shifts the problem of responsibility rather than solving it. Critics argue that if AI plays a central role in the decision-making process, humans might become overly reliant on its recommendations, effectively following automated choices rather than critically evaluating them. This could lead to a situation where AI is making decisions while humans retain technical responsibility, creating a false sense of accountability without genuine oversight. Without genuine oversight, AI risks becoming a purely procedural mechanism that enforces biases under the guise of neutrality. Furthermore, the risk of "gadget

worship", as described in *Political Theory of the Digital Age* (Risse, 2023), suggests that some may treat AI-generated decisions as infallible, failing to critically engage with the recommendations given. This issue has already presented itself within a myriad of real-world scenarios, most notably, a risk assessment tool called COMPAS made racially biased decisions for years, leading to a myriad of people being incarcerated for crimes they did not commit. A major reason why this flawed system was in use for so long was due to many factors, one of those being judges misinterpreting the system as a recidivism predictor, without exploring why such decisions were made (Angwin et al., 2016).

There is a similar objection that such a system could allow people to leverage ignorance for malicious results. Even though it has been established that under a system of human-AI decision-making, the moral culpability falls upon the human, this system could act as a moral buffer allowing people to either deflect the blame or truly be ignorant of how certain unjust or amoral decisions were made. For example, under the proposed system, if a wrongful arrest occurs, institutions cannot claim, “the AI suggested it” as a way to avoid accountability; however, such claims may erode trust in such systems. Due to this work aiming to enforce social justice through AI decision-making, an erosion of public trust could affect the scenarios in which a process is used.

Both objections essentially raise the same question: is human oversight meaningful? Is implementing such a system worth the risks?

Solution

To counteract these objections, a hybrid system must include clear mechanisms for accountability and oversight. The major trait that such a system must include is *transparency*. The AI decision-making process must be explainable, allowing humans to understand and challenge its suggestions. Currently, modern neural networks are significantly uninterpretable, leading to a large body of work seeking to develop effective systems that allow researchers to probe and understand why a decision was made. Despite that, an effective system of human-AI decision-making *cannot and should not* exist without adequate transparency.

Beyond transparency, regular audits must be conducted to ensure that AI systems remain fair, unbiased, and aligned with ethical guidelines. This is not due to changing ethics (which I believe should not occur within an ideal system) but rather the impact that fine-tuning the model with updated data can have on predictions. As a part of these audits, the decisions made should also be systematically reviewed by independent oversight bodies that are not influenced by the institutions deploying these technologies. These audits must go beyond mere performance metrics and actively investigate how the decisions uphold standards of justice. They specifically aim to explore whether the human in the human-AI decision-making process performed adequate verification. If valid, the data from these decisions would be used to fine-tune the model, resulting in more aligned predictions.

Another key component is the implementation of ethical safeguards to ensure people do not blindly follow AI-generated decisions. To avoid such errors, institutions relying on AI decision-

making must incorporate training programs that educate on the limitations of AI, the potential biases present in algorithmic outputs, and how to override AI recommendations when necessary.

Taken together, these safeguards would create a robust framework for human-AI decision-making that upholds social justice principles while ensuring that responsibility remains human-centered. By preventing AI from becoming an unchecked authority, this system fosters *both efficiency and accountability*.

Conclusion

The system I propose, guided by Rawlsian liberal principles, would ensure that AI-driven decisions adhere to social justice standards by embedding ethical oversight at every level. Checking a decision is faster than going through the process to make it, meaning that such a hybrid system will result in higher efficiency, and equity due to more people from different backgrounds being able to verify the decisions made in the same amount of time. In balancing automation with human discretion, this framework does not eliminate the risk of misuse but significantly reduces the possibility of AI being treated as a moral agent without harboring the responsibility. Moreover, it shifts the debate away from whether AI *can* make just decisions and toward how society should structure decision-making to ensure justice remains a human-led endeavor.

7. Toward the Future

Each of the political theories discussed, Utilitarianism, Rawlsian Liberalism, and Libertarianism, offers compelling insights into how AI should engage in decision-making, yet none are without

flaws. The complexities of justice, particularly when applied to AI systems, reveal the inherent trade-offs and moral dilemmas in any approach. As technology evolves, new challenges will emerge; however, the imperfections that may occur do not imply the absence of a moral foundation. If AI decision-making systems are built without a clear and coherent framework, they risk becoming inconsistent, unpredictable, and easily manipulated by those in power. A lack of political and moral grounding would lead to AI systems that reinforce biases, operate without accountability, and make arbitrary decisions that fail to align with any defensible notion of justice.

Rawlsian Liberalism is the most just framework for such a system because it balances fairness, individual rights, and accountability; however, even this framework is not immune to critique and will require continuous reassessment as AI systems evolve. The goal is not to find an unchanging, flawless model *but to commit to a structured ethical foundation*, one that prioritizes justice, adapts to new insights, and ensures that AI helps humanity.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Brożek, B., & Janik, B. (2019). Can artificial intelligences be moral agents?. *New ideas in psychology*, *54*, 101-106.

Cambridge Dictionary. (2019). *JUSTICE | definition in the Cambridge English Dictionary*. Cambridge.org. https://dictionary.cambridge.org/us/dictionary/english/justice

Duignan, B. (2024, September 17). *Social Justice*. Britannica. https://www.britannica.com/topic/social-justice

Gert, B., & Gert, J. (2002). The Definition of Morality. *Plato.stanford.edu*. https://plato.stanford.edu/archIves/sum2020/entries/morality-definition/

Merriam-Webster. (2024). *Definition of ETHIC*. Merriam-Webster.com; Merriam-Webster. https://www.merriam-webster.com/dictionary/ethic

Noorman, M. (2012). *Computing and Moral Responsibility*. Stanford.edu. https://plato.stanford.edu/entries/computing-responsibility/

Rawls, J. (1958). Justice as fairness. *The philosophical review*, *67*(2), 164-194.

Risse, M. (2023). *Political theory of the digital age: where artificial intelligence might take us*. Cambridge University Press.

Roose, K. (2024, October 23). Can A.I. Be Blamed for a Teen's Suicide? *The New York Times*. https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html