How Legal Systems Struggle to Maintain Accountability When AI Make Erroneous Decisions

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Arran Scaife

Spring 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Travis Elliot, Department of Engineering and Society

Introducing Legal Shortcomings for AI

Artificial Intelligence (AI) has become an increasingly dominant technology in today's age as we've seen developments over various sectors including transportation, criminal justice, and art to just name a few (Rigano, 2019). But, as these systems become more ubiquitous, it has also become evident that the existing legal frameworks may not be sufficient to determine who is accountable when AI make poor or unethical decisions that cause accidents or other harms to people. The US legal system currently determines this accountability through fault and liability, but through the STS framework of infrastructure by Susan Leigh Star, this paper aims to explore how these legal concepts erode as legal systems struggle to maintain accountability when AI make erroneous decisions.

The Ethnography of Infrastructure

The Ethnography of Infrastructure is an STS framework developed by Susan Leigh Star that explains the role of infrastructure in shaping social and technological systems. Her framework addresses three key questions: How can we characterize infrastructure, how should we study infrastructure, and how should we understand infrastructure.

To define infrastructure, Star explains there are nine characteristics that generally define infrastructure, as seen in Table 1 below.

Table 1

Infrastructure Characteristics

Term	**Description**
Embeddedness	Refers to how “infrastructure [integrates itself] into and inside of other structures, social arrangements, and technologies” (Star, 1999, p. 381).
Transparency	Refers to how infrastructure “does not have to be reinvented each time or assembled for each task, but invisibly supports those tasks” (Star, 1999, p. 381).
Reach or Scope	Refers to how infrastructure affects other aspects outside of its immediate applications. This can be “spatially or temporally [having] reach beyond a single event or one-site practice” (Star, 1999, p. 381).
Learned as Part of Membership	Refers to how infrastructure is uniquely learned by a group and subsequently taken for granted. The group's collective knowledge of the infrastructure is passed on to new members and becomes part of their everyday lives.
Links with Conventions of Practice	Refers to how infrastructure is designed based on past traditions of use, and the conventions of practice that have developed around it.
Embodiment of Standards	Refers to how infrastructure collaborates with other technology over interfaces and is built on a set of standards that have been agreed upon by

	industry or government bodies. These standards ensure that different components of infrastructure can work together seamlessly.
Built on an Installed Base	Refers to how infrastructure inherits the properties of legacy infrastructure it is built upon. New infrastructure is often built on top of existing infrastructure base and “inherits [the] strengths and limitations from that base” (Star, 1999, p. 382).
Becomes Visible Upon Breakdown	Refers to how infrastructure's invisibility erodes as the technology fails, and the infrastructure becomes visible. When infrastructure breaks down, people become aware of how much they rely on it and begin to notice its presence.
Fixed in Modular Increments	Refers to how infrastructure is not installed or reinstalled in an instant but rather over a long time. Infrastructure is built up in modular increments, with each module adding to the overall functionality of the system. This approach allows infrastructure to be updated and improved over time without the need for a complete overhaul.

These characteristics are important because Star explains that infrastructure is not absolute, but relational. The example she gives is a set of stairs. To us, stairs are a means of traversing levels but “for the person in a wheelchair, the stairs and doorjamb in front of a building are not seamless subtenders of use, but barriers” (Star, 1999, p. 380). Thus, we must use these nine characteristics of infrastructure to first help us classify the infrastructure of each case. Then, we can look towards studying it.

Star explains how to study infrastructure in three tactics. The first is identifying the master narratives and the non-master narratives. We do this by assessing the assumptions of the infrastructure and then analyzing what these assumptions may actually mean in practice. Second, we surface invisible work by looking at what labor becomes visible when the infrastructure breaks down. Finally, we can examine paradoxes of infrastructure, which can be shown by small changes causing significant technical challenges for users of the infrastructure.

Now that we understand how we can define and study the infrastructure in each case, we must aim to understand infrastructure across different levels of use. These levels are defined by Star under three categories: As a constructed material artifact that has pragmatic properties designed to be of a practical use to humans, as a trace of record of activities designed to maintain an imprint such that it is an information collecting device, or as a representation of the world designed to model the world in an abstract substitution.

Infrastructure in AI and Legal Systems

One way to understand the relationship between AI and legal systems to view them as two distinct yet interconnected infrastructures. AI can be seen as infrastructure that is embedded within other technologies while also demonstrating a reach or scope beyond single events, which we can see from multiple current examples of AI usage. GPT-3 is the most obvious example of such an embedded system with far reaching scope, as it has been a technological tool to develop other language models, such as Amazon chatbots and AI writing checkers, while also drastically affecting how schools handle academic integrity, as it has made it easier to detect plagiarism and academic dishonesty. AI’s transparency as an infrastructure is also notable, as it invisibly supports tasks such as language generation and translation, natural language processing, and even creative writing and art. (Floridi et al., 2020)

Similarly, legal systems can also be viewed as an infrastructure that are learned and taken for granted as part of membership in the legal profession. Many lawyers utilize technical terms, Latin phrases, and other legal jargon that may ostracize those outside the legal profession from understanding legal systems. (Eliot, 2021) Additionally, legal procedures can be a highly complex process of filing legal documents, presenting evidence, and making arguments that, in combination with the niche jargon, make it a common necessity to have lawyers defend a client in legal cases. We can see evidence of this by the instantiation of the Miranda right to be presented an attorney. (White, 1992) These and other standards and conventions are also embodied in legal practices and procedures and are linked with other social and political systems. The rule of law stating all persons are subject to the law, due process stating that the government must respect all rights owed to a person, and legal precedent, requiring the use of past court decisions as a basis for current cases, are all examples of how legal systems, by holding all persons to a similar standard and maintaining the protection of citizens, actively affects current social and political systems. This idea of precedent also illuminates the fact that legal systems are built on an installed base of laws, regulations, and precedents, ensuring it not be installed in one go but rather evolve over time. (Lewis, 2021)

From this, we can see that both AI and legal systems embody standards and interfaces that allow them to work with other infrastructures. Alongside this, they also share the characteristic of becoming visible upon breakdown, as failures and errors in AI decision-making can lead to legal challenges and calls for accountability. Since legal systems are designed to punish those who are responsible, many accidents involving the decision-making of AI and the resulting tort, which are civil cases to determine liability and compensation resulting from some harm caused by a responsible party, are clear examples of this apparent breakdown, as AI, although responsible due to its decision-making, cannot provide a compensation. (Doshi-Velez et al., 2017, p. 6) These infrastructures are also fixed in modular increments also contribute to their path dependency on past conventions and their potential resistance to change. Analyzing both as constructed material artifacts, we can see that AI, as a learning machine, must update its judgements based on new information gained and, thus, its current decisions must be improved in modular increments. Similarly, legal systems update their policy based on legal cases that may set new presidents or change existing ones such that this policy updating procedure is modular incremental and designed to fix the current legal system. (Lewis, 2021)

Throughout both legal systems and AI, we can see that they exemplify characteristics of infrastructure, but we've mainly focused on both as constructed material artifacts designed for pragmatics use. Alternatively, we can also analyze either as a trace of activities or as a representation of the world.

AI not only attempts to model the world, but also attempts to create such a model in a similar manner to human intelligence, which is clear by the many parallels made between the workings of deep neural networks and organic brains. This technology also integrates a trace of activity as the world is encoded into the AI's decision-making. Human biases often amplify themselves through AI decision making, allowing us to better understand and analyze our flaws these systems pick up. (Ntoutsi et al., 2020) But, as of today, this trace is largely unreachable as AI decisions continue to be a "black box", which is a system whose mechanisms cannot be internally understood from an outside observer. In this regard, the AI's decisions are so transparent that, when it makes a poor decision, the black-box nature of the system makes it difficult to understand how or why that decision was made, further complicating the issue of accountability when said decisions have significant impacts on people's lives. (von Eschenbach, 2021)

Similarly, legal systems can be seen as a trace of activities, as it collects information about legal cases, statutes, and regulations. Legal systems are also a representation of moral principles, as we generally consider the law as the moral minimum, as many precedents are created and updated based on moral or ethical considerations. (Lingwall) As such, when AI decisions impact individuals or society, questions arise as to who should be held accountable for these decisions. With AI systems being black boxes, it becomes challenging to identify who or what is responsible for such a poor decision. This creates a significant challenge for legal systems, as their legal decisions rely on the moral principle that those who cause harm should be held accountable and thus provide relief and/or be punished.

To study the infrastructures of AI and legal systems, we can identify the assumptions that these infrastructures make, such as the assumptions made by AI models about data and training sets, and the assumptions made by legal systems about what constitutes a fair trial. Secondly, we can surface the invisible work that pervades these infrastructures, such as the unnoticed labor that goes into training and updating AI models alongside explaining them, and the unseen work done by lawyers and judges in legal cases involving AI harms. Finally, we can find the paradoxes of these infrastructures through the technical challenges that arise when integrating new features into AI systems, or the paradoxical outcomes that sometimes result from legal decisions.

Cases of Complicated AI Accountability

As AI has become increasingly prevalent in various industries, many high-profile cases have highlighted difficulties in assigning liability, fault, or accountability, when AI make mistakes. To examine this phenomenon, we can analyze 5 cases and their legal consequences under Star's framework of infrastructure to define, study, and understand the interplay between the infrastructure of AI the infrastructure of legal systems.

Case 1:

For this study, we can analyze two cases. The first case happened in 2018, where a self-driving Uber test car in Tempe, Arizona struck and killed a pedestrian, Elaine Herzberg. Elaine was visible crossing the road for 6 seconds, but the driver was distracted watching a show. (Lee, 2019) The AI failed to detect her as a pedestrian and failed to apply the brakes only until 1.3 seconds before impact. Instead, the system classified her as an unidentified object, then a vehicle, then a bicycle, each of which generated different predicted paths for Elaine. (Issac et al., 2018) In this case, because the driver was not paying attention, the prosecutors ruled that Uber wasn't criminally responsible (Pavia, 2018), while the driver was charged with negligent homicide. (Lee, 2020) However, in civil trial, Uber and Elaine's family ended up reaching an undisclosed settlement, suggesting some civilly justified cause of action against Uber. (Neuman, 2018)

The second case happened in May 2016, where a Tesla Model S car on autopilot collided with a semi-truck, killing the driver, Joshua Brown (Figure 1). After investigation, it was found that the driver did not have his hands on the wheel, the driver had their laptop mounted to their dashboard at the time of the crash, which was playing a movie at the time of the crash, and that the autopilot did not engage the brakes at 74 mi/h when it passed under the trailer. Tesla claimed this was a result of the brightly lit sky making the white trailer difficult to detect by the driver and the AI system. In this case, the truck driver was found at fault for failing to yield the right of way to the Tesla. Despite this, the National Transportation Safety Board found no fault against Tesla since Joshua didn't have his hands on the wheel. (National Transportation Safety Board, 2020) Thus, no lawsuit was filed since "if the families of Brown … file suit against Tesla, they will face significant challenges… [as] the available legal theories for product liability and accident compensation claims are traditional… Tesla requires buyers to consent to contract terms which require drivers to keep their hands on the steering wheel at all times, including when the autopilot system is engaged. And technically, Tesla's current autopilot technology is not "self-driving." Thus, Tesla would not be legally liable for Joshua's misuse of their technology. (Baiamonte, 2017)

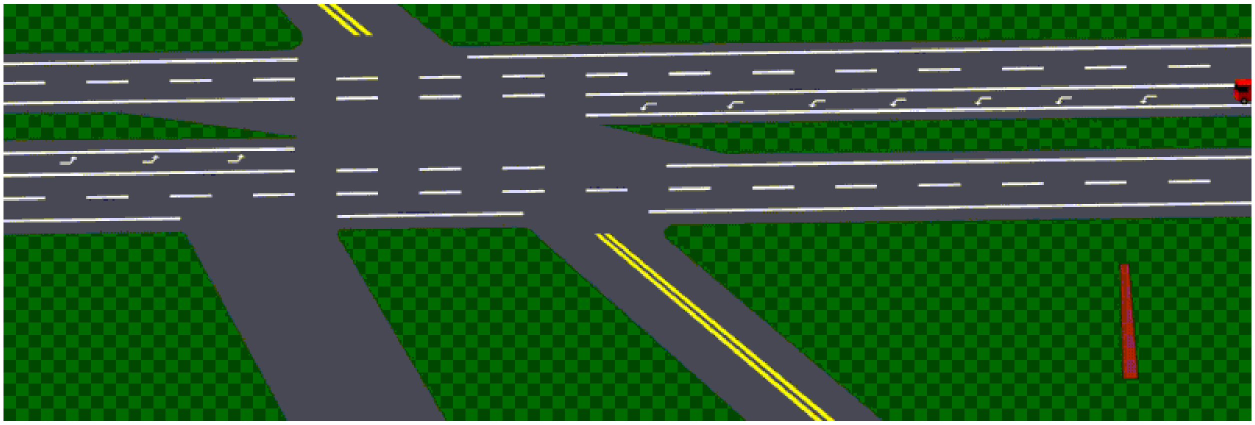

Figure 1. Depiction of the Tesla autopilot crash between Joshua Brown and a semi-truck (Tesla Autopilot, 2023)

We can see that AI is an infrastructure in this case by its embeddedness in the car, its transparency in invisibly supporting the driving of the car evident by the drivers not paying attention to the road, and its reach by its mistakes causing death and family grief. We can equally see that the legal system is infrastructure in its embeddedness within society, its visibility when broken, as it failed to find any person criminally responsible for the crash, and its scope, as rulings can have far reaching consequences on parties and future cases.

Now, towards studying these cases, we can identify the assumptions made by both systems. Both cases illuminate the AI's assumptions about how it recognizes and classifies objects in its environment, and its limitations in understanding the context of situations. In the pedestrian case, we see this in the AI system's false classifications of the person and the false path trajectories that caused the crash whereas in the truck case, we see this in the AI system's null classification of the trailer and its reluctance to apply the brakes even after the collision. Both cases also falsely assume that the driver is fully aware while continuously assuring the driver of its driving capabilities. This ignores the contradictory fact that by transparently supporting driving, the AIs are indirectly causing the driver to be comfortable and, thus, less aware.

The legal system's responses, on the other hand, reflect their assumptions about accountability and responsibility for mistakes involving the accident. Here, the legal system makes a false assumption that only humans can make deliberate decisions that cause harm, where non-human decisions must then be the result of human error or negligence. This assumption is backed by the fact that the ruling in the pedestrian case was that Uber was not criminally responsible for the death of Elaine and for the truck case, the its assumed that the ruling would be that Tesla was not at fault, as the driver is legally responsible for safely driving the car due to Tesla's statement that autopilot is not designed to replace human drivers, despite this being the exact pragmatic purpose of autopilot. By maintaining the driver be fully aware, the legal system assumes the driver should have intervened and prevented the accident, even though the AI system was designed to operate autonomously. This assumption, similar to the AI, ignores the fact that the driver may have been lulled into a false sense of security by the AI's capabilities and may not have been fully aware of the potential risks involved. Thus, a paradox of this infrastructure arises, as Uber and Tesla legally require the driver to remain attentive to avoid liability while simultaneously causing the driver to become inattentive through the AI's capabilities. For the driver, this creates a situation where they are held accountable for the actions of the AI system, even though the system was designed to operate autonomously, which directly contradicts the legal system's purpose to hold accountability to those responsible.

To understand the complex interplay between AI and the legal system in this case, we can view both infrastructures as constructed artifacts with pragmatic properties. The AI's pragmatic purpose is to drive for the driver while the legal system's pragmatic purpose is to prevent harm by punishing those who are responsible for harm. In these regards, both infrastructures failed their purpose by the Ais causing accidents and the legal system not finding any fault from any parties, making both system's limitations visible.

Thus, as pragmatic infrastructures, the AI's failure to detect the pedestrian and the truck highlight the importance of AI design and testing, as these AI demonstrate clear design flaws and limitations in detecting and responding to pedestrian and truck movements. The cars' inability to detect such an important classification shows how AI should prioritize edge case classifications, as they hold the most difficult classifications for the AI while also potentially having drastic consequences for failure.

These cases also highlight the need for clearer legal frameworks for determining liability and responsibility involving AI systems where responsibility may be shared by humans and machines. Here, the drivers’ inattention is used as a scapegoat for avoiding responsibility for the AI systems’ failures, as the legal system avoids addressing how the AIs indirectly cause the drivers to become inattentive by falsely assuming they were attentive enough to grab the wheel and avoid the crash. The legal system should aim to look past mere legal disclaimers protecting companies from liability and look more towards the unintended consequences, such as inattention, AI may have on its users.

Finally, we find a need for greater transparency and accountability in the development and deployment of AI systems. The mere fact that the AI system was not able to accurately detect and respond to a pedestrian question why the driver, nor the court, were informed of the AI’s decisions. As justifications behind the AI’s decisions reduce transparency to the driver and the court, there is a growing need for the development of explainable AI systems that reduce AI’s black-box nature, where the decision-making process of AI is made clear and understandable to both developers and end-users. Explainable AI not only increases transparency but also ensures that the AI’s decision-making aligns with ethical principles and legal frameworks, as well as mitigating potential biases and errors that may be present in the system. The development of explainable AI systems is particularly relevant in safety-critical applications, such as autonomous vehicles, where lives are at stake. Thus, as the use of AI systems in society continues to increase, it becomes imperative that we develop systems that are transparent, trustworthy, and accountable, to ensure that they benefit society as a whole while minimizing potential harm. (Deeks, 2019, p 1829)

Case 2:

Outside of autonomous driving, not only does AI make accountability hard to determine, but it can also mask the actions of those truly responsible. In 2018, Trivago, a hotel booking website, was sued by the Australian Competition and Consumer Commission (ACCC) for misleading advertising practices, claiming Trivago misled consumers by representing its hotel search results as impartial while fraudulently biasing its AI-driven search results towards hotels that paid Trivago higher fees for referrals.

Although Trivago claimed that the algorithm didn’t take into account the fees paid by hotels, expert witnesses analyzed similar decisions of different AIs, revealing otherwise. Thus, the court held Trivago was misleading consumers by not maintaining an impartial AI-driven hotel search whilst claiming an unbiased search. (Fraser et al., 2022, pp. 186-187)

Here, we can start by defining both the AI system and the legal system as interconnected infrastructures. The Trivago AI system is embedded in the Trivago website, which integrates it into the larger system of hotel bookings. It operates invisibly, supporting the hotel search function of the website. The reach or scope of the infrastructure extends beyond the immediate hotel search function to include the reputation and trustworthiness of the Trivago brand as a whole. The AI system is learned as part of membership, with the system being designed to learn and adjust its search results based on user behavior. It links with conventions of practice by utilizing past user data to inform future search results. The system embodies standards, both in terms of the search algorithm and the business model of generating revenue from hotel referrals. It is built on an installed base of past user data and search results, which shape the system's future decision-making. The system becomes visible upon breakdown, as evidenced by the court case. Finally, the system is fixed in modular increments, with adjustments being made incrementally over time based on user feedback and data analysis.

Similarly, the legal system is embedded in society and integrates with various technologies, such as AI algorithms used by companies like Trivago. It is also transparent in that its ruling indirectly discourages AI abuse by demonstrating detection and punishment for such fraud. Finally, the legal system is linked to conventions of practice, as it is based on past traditions and precedents. In this case, the legal system was used to hold Trivago accountable for misleading advertising practices, which can serve as an example of how legal infrastructure can shape and influence social and technological systems through reach and scope.

Now, understanding that both are infrastructure, we can continue to study Trivago’s AI and the Australian legal system. For AI, we can identify the master narrative and non-master narratives of the infrastructure. The master narrative is that the Trivago AI offers an unbiased hotel search that helps consumers find the best deals while the non-master narrative is that the AI biases search results towards hotels that pay higher referral fees. Secondly, we can surface invisible work by looking at what labor becomes visible when the infrastructure breaks down. Here, the designers of the Trivago AI became visible as their decisions and actions were put under scrutiny during the investigation and trial. Finally, we can examine paradoxes of infrastructure where the paradox in this case is that the AI system is designed to optimize revenue for Trivago through hotel referrals while also claiming to provide an impartial search function for consumers,

More interestingly though, for the legal system, a master narrative it maintains is to hold companies accountable for fraudulent practices involving misleading advertising with AI and that simulations of AI doing different activity is evidence of tampering, but it also demonstrates a non-master narrative that it becomes incredibly difficult to determine accountability when AI is involved and that the legal system may not have the necessary expertise to fully understand the complexities of AI systems, as finding differing AI simulations is only an indication of tampering while showing bias in the original Trivago AI is proof of tampering. For invisible work, the work of expert witnesses analyzing the AI's decision-making process became visible in the court case as they examined multiple AI's to objectively find Trivago had biased their system. Finally, for paradoxes of infrastructure, although the law is meant to ensure accountability and protect consumers, it can also be used by companies to shield themselves from accountability, as, even with the expert witness' AI simulations, Trivago may have argued that their AI-driven search results were not intentionally biased and that they were acting in good faith, which may have been a legitimate defense under the law. Thus, a paradox arises where the legal system is both a tool for accountability and a means of protecting companies from accountability. In the same vein, while the legal system is designed to be impartial and unbiased, it can also be influenced by factors such as the expertise of the lawyers involved or the biases of the judge or jury. In the Trivago case, the decision of the judge to hold the company accountable for misleading advertising may have been influenced by a range of factors, including their interpretation of the law, their understanding of AI and its limitations, and their assessment of the evidence presented. (Landsam et al. 1994) This creates a paradox where the legal system is both impartial and subject to bias and influence. Thus, these limitations between Trivago's AI and the legal system highlight many necessary normative claims.

Finally, we can understand both systems as a trace of human activities. Here, the Trivago AI's recommendations served as a history of Trivago's deception as expert witnesses compared their Ai to unbiased AI models. This discrepancy was enough evidence to conclude that Trivago must have biased the AI's training to get such a differing activity. This is unique to this case since expert witnesses can directly compare the current model against their built models, whereas this is not possible when the AI is destroyed or makes a decision in a fringe case, as seen with autonomous car accidents. Similarly, the legal system can also be viewed as a trace of human activity, as it sets a precedent that condemns past fraudulent AI biasing without necessarily removing bias from AI.

As seen by this case, AI can be easily used to mask fraudulent practices and make accountability difficult to discern. Thus, to ensure companies are held accountable for the actions, it's essential to have field experts who can understand the complexities of AI systems to identify bias or fraud, as this case demonstrated that this evidence was crucial to obtaining a fair ruling. This can be achieved by many potential solutions, such as training programs, workshops, and/or educational incentives that focus on AI usage in differing industries, thus helping the alleviation of jury, judge, or lawyer bias with AI. A similar solution to this problem involves educating all parties involved prior to the ruling with an understanding of AI basics to avoid bias in the court.

We should also invest in increasing transparency and accountability for AI systems. Although this case was able to discern a trace of Trivago's wrongdoings, the current inability to investigate the internal working of AI decisions makes this practice fallible, as experts must investigate other AI systems that could behave differently to even impartial AI due to inherent bias present in many AI systems, which we investigate in the next case.

Case 3:

In 2012, an AI algorithm called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), used a variety of factors to assess a defendant's likelihood of reoffending, including criminal history, age, and gender. COMPAS was naively used by several states, such as Wisconsin, in the US to help judges determine the appropriate sentence for a defendant. (Fraenkel, 2020) However, in 2016, an investigation by ProPublica found that even when controlling for factors such as prior offenses and demographics, black defendants were almost twice as likely to be labeled as high-risk compared to white defendants. (Angwin et al., 2016) Thus, in July 2016, the Wisconsin Supreme Court ruled that judges can use COMPAS risk scores in sentencing, but there must be warnings alongside the scores to highlight its limitations. (Kirkpatrick, 2017) From then, this tool has been argued to violate the 14th amendment by providing disparate treatment based on race. (Thomas et al., 2022)

Like before, we begin by analyzing the infrastructure in this case by examining the characteristics of COMPAS. Since the legal system is the same as before, we can assume from here that it is infrastructure. Now, COMPAS is embedded in the criminal justice system embodies standards by integrating with other technologies such as court databases and criminal records. It is also transparent in that it invisibly supports societal functions by providing risk assessment scores to judges, which they use without consideration of score bias. It has a wide reach and affects other aspects outside of its immediate application, such as the sentencing of defendants by doubling sentences based on race. It is learned as part of membership, as it is uniquely learned by criminal justice practitioners and taken for granted in their decision-making process. It links with conventions of practice in the criminal justice system, as it uses factors such as prior offenses and demographics. It is built on an installed base, as it inherits the properties of legacy infrastructure it is built upon, such as the criminal justice system itself. It becomes visible upon breakdown, as its invisibility eroded when its limitations and biases were exposed by ProPublica. Finally, it is fixed in modular increments, as it is not installed or reinstalled in an instant but rather over a long time, which is demonstrated by the court revision in 2016.

Thus, since both COMPAS and the legal system are infrastructure, we can continue to study them using Star's tactics. Firstly, we find that COMPAS presents a master narrative that is it an unbiased tool that helps judges make more informed decisions. Despite this, it also projects a non-master narrative that it is in fact biased and doubles sentencing for black defendants. This in conjunction with the master narrative that defendants should receive equal treatment in rulings regardless of race under the 14th amendment, judges are led to false assumptions that they're giving fair sentences, which is a paradox in the judge's determination and the implementation of COMPAS, as its original intended purpose was to provide unbiased sentences. Thus, after 2016, the need for judges to manually adjust the rulings provided by the tool were surfaced as judges aimed to reduce its bias. Finally, we can also find a paradox of COMPAS in that it tries to be impartial in its rulings such that it won't include race or gender in accordance with the 14th amendment, but also considers demographic and gender directly to determine its ruling, which can be used by the AI to estimate race or sex.

To understand COMPAS, it can be best understood as a representation of the world and how it affects human bias. As a representation of the world that models how bias is formed, we can analyze the assumptions that are built into the infrastructure. For example, the factors that are used to calculate the risk of reoffending, such as criminal history, age, and gender, can be considered as assumptions. Thus, we can study how these assumptions may lead to biased results by analyzing how they reflect societal biases, such as racial profiling, and how they may perpetuate these biases. Further, this model can illuminate that, even if we find a correlation between these factors and reoffending, this does not necessarily mean that they are the direct cause of reoffending, thus indicating the necessity to separate correlations from causations in human and AI judgements. (Kadiresan et al., 2022)

From this, we now understand the importance of AI infrastructure being designed with transparency and accountability in mind to prevent biases from perpetuating in the decision-making process. Here, the lack of transparency led to 4 years of racial discrimination in multiple states, which could only be proactively resolved by 2016 since many defendants may have already experienced longer sentences than necessary. Especially in criminal justice systems, the use of AI algorithms should be accompanied by ongoing monitoring and evaluation of its usage lifetime to prevent unintended consequences. In this case, although ProPublica’s study helped find the bias in COMPAS, the data had already been present and could have been used to prevent COMPAS’s usage much earlier by tracking its sentencing statistics over time. Finally, AI should be initially designed with the understanding that it is severely limited in distinguishing correlation from causation. Thus, in applications where the distinction is most severe, such as court rulings, they should avoid using such technology either until the distinction is separated for AI decisions or until they can eliminate all inputs to the system that could contribute to biased outcomes. Thus, we see that it is important to prioritize the principles of fairness, accountability, and transparency in the development of AI infrastructure and its integration into critical decision-making processes. This includes considering the potential social and ethical implications of AI algorithms and continuously monitoring their impact to ensure that they align with societal values and legal standards. By doing so, we can avoid the unintended consequences of biased decision-making and ensure that AI is used for the betterment of society as a whole.

Case 4:

Finally, AI has also raised questions on intellectual property and copyright about who takes accountability for AI-derived work. In 2018, a group of French artists and programmers created an AI system called "The Next Rembrandt" that was designed to create a new Rembrandt-style painting, which analyzed Rembrandt's previous works and used this information to generate a new painting that resembled his style. The resulting painting, titled "Portrait of Edmond de Belamy", sold for over $400,000 at a Christie's auction in New York. (Schlackman, 2020) However, the legal status of the painting's copyright was unclear, as the AI system was programmed by human creators, but the painting was generated autonomously by the system.

In this case, the copyright for the painting was held by a company called Obvious, which was formed by the French artists and programmers who created the AI system. Obvious co-founder, Hugo Caselles-Dupré argued that they should hold the copyright because they've contributed significant modifications to the art and efforts to the model that should classify it as new work. (Flynn, 2018). Despite this though, some legal experts argued that the copyright should belong to the AI system itself, as it was responsible for the actual creation of the painting.

(Jones, 2018) However, since copyright law continues to recognize only human authors and creators of AI are not direct authors of the AI’s generated works, neither argument has been widely accepted. (Fischer, 2014)

The system is embedded within other technologies and is built upon an installed base, as it was an established AI method created by a group of French artists and programmers who used Rembrandt's previous works as data for the AI to generate a new painting. The system is also transparent in that it invisibly supports the societal function of creating art. Its reach is not just limited to generating one painting, but the potential to generate many more in the future, indicating a shift from individual artists’ work to commissioned AI art for cheap. The AI system was learned as part of membership by the creators who designed it and was based on past traditions of use to find unique ways of creating art, linking it with conventions of practice. Finally, the system also embodies standards as it collaborates with other technologies over interfaces, such as the database of Rembrandt's previous works. Thus, we can confidently claim that the AI is an infrastructure.

Using this, we can study the infrastructure under Star’s tactics. We find that the legal system makes a bad assumption that humans are the sole creators and owners of copyrightable works, thus creating an equivalent master narrative. On the other hand, it neglects the fact that AI systems could be creators of copyrightable works such that a non-master narrative becomes that creators of AI that make copyrightable works should hold the copyrights to works generated by their systems. For inviable work, the creation of the AI system took significant resources and labor that wouldn’t generally be readily available to the public without this legal intervention. The artists and programmers had to analyze Rembrandt’s works, develop algorithms to imitate his style, and train the AI system using large amounts of data, which is often taken for granted despite being crucial to the creation of the AI painting. Finally, although the AI system was created and claimed to autonomously generate artwork without direct input or control from the human creators, the French artists are also claiming ownership of the AI and the data it was trained on alongside the generated painting, which creates a paradox where the AI is seen as a self-sufficient creative entity, yet it relies on human input and expertise to function, and its creations are ultimately claimed by human owners.

Finally, to understand the infrastructure, we can analyze the AI as a representation of a human with individual autonomy. In this case, the AI was designed to analyze and learn from Rembrandt's previous works, which can be seen as a representation of the world of art. Similarly, by using this data, the AI system was able to create a new painting that resembled Rembrandt's style, which can be seen as a representation of human creativity and expertise in art. However, the fact that the AI system was created and programmed by humans also highlights the limitations of human autonomy in relation to AI. While the AI system was able to generate a new painting autonomously, it still relied on human input and expertise to function properly, including the data it was trained on and the algorithms used to generate the painting. This creates a paradox of autonomy where granting copyrights to the French artists stifles autonomy and creativity of independent artists not using AI while providing copyrights to the AI stifles autonomy and creativity of artists using AI to generate art.

We can see from this that we desperately need to update copyright laws to recognize AI-generated works as having their own independent copyright status, and to establish guidelines for determining ownership and accountability in such cases, as current copyright law will not find resolving grounds due to the previously mentioned paradoxes. Additionally, groups and individuals who create AI systems capable of generating copyrightable works should be required to clearly specify their roles and responsibilities in the creation process. Otherwise, we are left with situations of unclear responsibility for derivative AI works.

Discussion

Firstly, legal frameworks must evolve to accommodate the complexities of AI and its decision-making capabilities where policies and regulations should be developed to hold AI systems accountable for their decisions and actions while also protecting the rights and interests of those affected by these decisions. We've seen from the autonomous car cases that it became difficult to determine who was at fault for the case without defaulting to the driver being required to maintain awareness despite the AI supporting the driver's unawareness. In the COMPAS case, the fairness of the AI's risk assessments caused significant unfair harms to those incarcerated such that, without regulations for such AI, such systems could perpetuate biases and inequalities free of consequence. Finally, the Next Rembrandt case demonstrated questions about copyright and ownership of the AI-generated artwork where, without clear regulations and laws in place to define legal responsibility and liability for the actions of AI systems, it is difficult to determine who should hold the copyright for such creations.

Additionally, the development of AI should prioritize transparency and inclusivity to minimize the potential for biased decision-making, especially in areas such as criminal justice, which should be subject to rigorous testing and auditing to ensure that they do not perpetuate or exacerbate existing biases in society. This also extends to other legal areas where biases can affect status, such as employment. We saw from the autonomous car cases that transparency and inclusivity were not a focus of either case, as both drivers were never alerted to the wrongful judgments of the AI, each having only seconds to react to the poor AI planning prior to the collisions. Not only this, but we also saw that the transparency in the Trivago case was situated by intentional misleading under the guise of black boxed AI. Here, even with the help of expert witnesses, the lack of transparency in AI decisions has led to not only unclear AI judgements, but also its abuse.

Finally, as AI continues to transform various aspects of society, there is a need to critically evaluate and rethink the current legal and economic frameworks surrounding intellectual property and ownership, which includes reevaluating who can be considered an author or creator of AI-generated works, who is responsible for the faults of AI's actions, as well as considering the implications of AI-generated art on traditional artistic practices and commercial value. The Next Rembrandt case demonstrated the potential for AI-generated art to challenge our understanding of creativity and authorship. While the artwork was generated by an AI algorithm, it required human input to design and create the final piece, raising questions about who can claim authorship and ownership of such works and whether the contributions of the AI system should be acknowledged and protected under current intellectual property laws. Not only does this apply to who claims works, but also who claims the faults. In both autonomous car cases, we saw that the responsibility for the accidents did not fall on the manufacturer due to a lack of evidence towards negligence and a lack of awareness on the driver's part, despite AI being inherently black box while also incentivizing the driver to become unaware of the surroundings without telling the driver its planning process.

Counterargument Responses

A main counterargument against this would be that AI is constantly evolving, and it may be challenging for policymakers to keep up with the rapid pace of technological advancements. Thus, it may be difficult to develop policies and regulations that hold AI systems accountable for their decisions and actions while also protecting the rights and interests of those affected by these decisions. To respond to this, although it is true that is constantly evolving, it still remains imperative to develop legislation that can keep up with these advancements so that AI systems and their developers are held accountable for their decisions and actions. Otherwise, cases such as Trivago would have no remedy to stop them from misleading customers. Thus, if we were to wait for AI to become stable or mature, which may not be guaranteed, before implementing legislation, we'll likely be led to significant harm being done in the interim where it may become more challenging to develop regulations as AI becomes more widespread.

Another counterargument could be that since AI systems often rely on complex algorithms that may be difficult to understand or explain, it may be challenging to ensure transparency and inclusivity, making prioritizing transparency and inclusivity in AI decision-making difficult to implement in practice. To this, I would respond that although it may be challenging to ensure transparency and inclusivity in AI decision-making, it is still crucial to prioritize these values to minimize the potential for biased decision-making. Additionally, there are already efforts underway to develop explainable AI systems that can provide insight into how decisions are made, allowing policymakers to already encourage the adoption of such systems to ensure greater transparency and accountability. (Dosilovic, 2018)

Finally, one might claim that current laws already provide sufficient protections for AI-generated works and those affected by AI decisions such that rethinking the current legal and economic frameworks surrounding intellectual property and ownership may be unnecessary. My response here would be that, although current laws may provide some protections for and against AI-generated works, we've already seen that they aren't sufficient to address the unique challenges posed by AI. As with the Next Rembrandt case, traditional copyright laws already do not account for the contributions of AI systems to creative works, leading to ambiguity over who can claim authorship or ownership. Additionally, the COMPAS case highlights how the current frameworks find difficulty in addressing AI systems that perpetuate biases or cause harm, where the ownership of such harmful works becomes more important.

Conclusion

The cases discussed demonstrate the need for legal revisions to guide the development and use of AI. These revisions should prioritize transparency, inclusivity, and accountability in AI systems, ensuring that AI systems are subject to rigorous testing, auditing, and explainable AI measures over time to minimize potential biases and prevent harm to individuals or society as a whole. The cases also highlight the need to reevaluate legal and economic frameworks surrounding intellectual property and ownership as AI-generated works challenge our understanding of creativity, autonomy, and authorship. As AI continues to transform various aspects of society, it is crucial to have ongoing discussions and revisions about the ethical and social implications of these technologies to ensure their responsible and beneficial integration into our lives.

References

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016). Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Baiamonte, C. (2017). Available at: https://lawreview.syr.edu/teslas-autopilot-cleared-by-government-investigation-but-questions-remain-about-liability-for-accidents-involving-self-driving-and-safety-technology/

Deeks, A. (2019). THE JUDICIAL DEMAND FOR EXPLAINABLE ARTIFICIAL INTELLIGENCE. *Columbia Law Review*, *119*(7), 1829–1850. https://www.jstor.org/stable/26810851

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S. J., O'Brien, D., … Wood, A. (2017). Accountability of AI under the law: The role of explanation. *SSRN Electronic Journal*. doi:10.2139/ssrn.3064761

Dosilovic, F. K., Brcic, M., & Hlupic, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. doi:10.23919/mipro.2018.8400040

Eliot, L. (2021). Considerations about legal jargon and the use thereof by ai. *SSRN Electronic Journal*. doi:10.2139/ssrn.3989332

Fischer, M. (2014). Available at: https://www.lexology.com/library/detail.aspx?g=62031706-af26-4463-82b4-4c2864becc2d

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*(4), 681–694. doi:10.1007/s11023-020-09548-1

Flynn, M. (2018). Available at: https://www.washingtonpost.com/nation/2018/10/26/year-old-developed-code-ai-portrait-that-sold-christies/

Fraenkel, A. (2020). Available at: https://afraenkel.github.io/fairness-book/content/04-compas.html

Fraser, H., Simcock, R., & Snoswell, A. J. (2022). Ai opacity and explainability in tort litigation. *2022 ACM Conference on Fairness, Accountability, and Transparency*. doi:10.1145/3531146.3533084

Isaac, M., Wakabayashi, D., & Conger, K. (2018). Available at: https://www.nytimes.com/2018/08/19/technology/uber-self-driving-cars.html

Jones, J. (2018). Available at: https://www.theguardian.com/artanddesign/shortcuts/2018/oct/26/call-that-art-can-a-computer-be-a-painter

Kadiresan, A., Baweja, Y., & Ogbanufe, O. (2022). Bias in AI-based decision-making. *Educational Communications and Technology: Issues and Innovations*, 275–285. doi:10.1007/978-3-030-84729-6_19

Kirkpatrick, K. (2017). It’s not the algorithm, it’s the data. *Communications of the ACM*, *60*(2), 21–23. doi:10.1145/3022181

Landsman, S., & Rakos, R. F. (1994). A preliminary inquiry into the effect of potentially biasing information on judges and jurors in civil litigation. *Behavioral Sciences & the Law*, *12*(2), 113–126. doi:10.1002/bsl.2370120203

Lee, D. (2019). Available at: https://www.bbc.com/news/technology-50484172

Lee, T. (2020). Available at: https://arstechnica.com/cars/2020/09/arizona-prosecutes-uber-safety-driver-but-not-uber-for-fatal-2018-crash/

Lewis, S. (2021). Precedent and the rule of law. *Oxford Journal of Legal Studies, 41*(4), 873–898. doi:10.1093/ojls/gqab007

Lingwall, J. (n.d.). Available at: https://boisestate.pressbooks.pub/businessethics/chapter/what-is-the-idea-of-a-moral-minimum/

National Transportation Safety Board. (2020). *Highway Accident Brief: Collision Between Car Operating with Partial Driving Automation and Truck-Tractor Semitrailer Delray Beach, Florida, March 1, 2019* (Publication HWY19FH008). Available at: https://www.ntsb.gov/investigations/AccidentReports/Reports/HAB2001.pdf

Neuman, S. (2018). Available at: https://www.npr.org/sections/thetwo-way/2018/03/29/597850303/uber-reaches-settlement-with-family-of-arizona-woman-killed-by-driverless-car

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., … Staab, S. (2020). Bias in data-driven Artificial Intelligence Systems—an introductory survey. *WIREs Data Mining and Knowledge Discovery, 10*(3). doi:10.1002/widm.1356

Pavia, W. (2018). Available at: https://www.thetimes.co.uk/article/driverless-uber-car-not-to-blame-for-woman-s-death-klkbt7vf0

Rigano, Christopher. (2019, January). Using artificial intelligence to address criminal justice needs. Available at: https://nij.ojp.gov/library/publications/using-artificial-intelligence-address-criminal-justice-needs

Schlackman, S. (2020). Available at: https://artrepreneur.com/journal/the-next-rembrandt-who-holds-the-copyright-in-computer-generated-art/

Star, Susan Leigh. “The Ethnography of Infrastructure.” *American Behavioral Scientist*, vol. 43, no. 3, 1999, 377–391., doi:10.1177/00027649921955326.

Tesla Autopilot. (2023, May 10). Wikipedia. Available at: https://en.wikipedia.org/wiki/Tesla_Autopilot#cite_note-487

Thomas, C., & Pontón-Núñez, A. (2022). Automating judicial discretion: How algorithmic risk assessments in pretrial adjudications violate equal protection rights on the basis of Race. *Minnesota Journal of Law & Inequality*, *40*(2), 371–407. doi:10.24926/25730037.649

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust Ai. *Philosophy & Technology*, *34*(4), 1607–1622. doi:10.1007/s13347-021-00477-0

White, M. J. (1992). Legal complexity and lawyers’ benefit from litigation. *International Review of Law and Economics*, *12*(3), 381–395. doi:10.1016/0144-8188(92)90016-k