Simulating Autonomous Vehicles for XAI

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

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Arran Scaife

November 1, 2022

Technical Team Members:
Arran Scaife
Victor Lou
Kayla Boggess

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.

ADVISORS

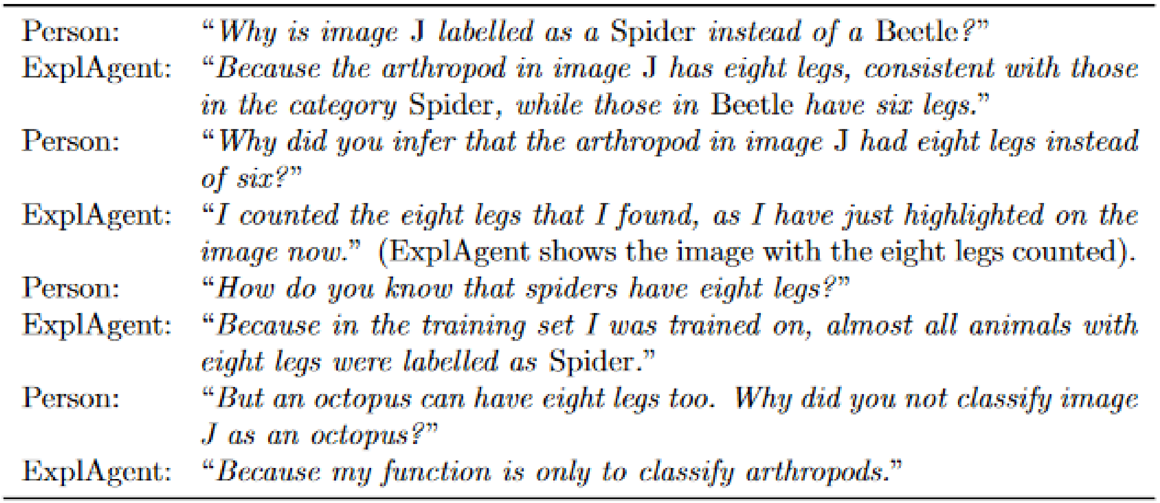
Travis Elliot, Department of Engineering and Society

Lu Feng, Department of Computer Science and Engineering

Introducing Legal Applications of XAI

In court, a judge and jury work to find those responsible through testimony, allowing them to deduce whom the responsibility falls upon. In car accidents, if the case is between the drivers, we call this fault, whereas if the case is between a driver and the manufacturer, we call this product liability established by the process of tort litigation. Here, lawyers present situational factors such as drugs, turn signals, distractions, etc. to establish cause and therefore fault for the accident. On the other hand, in the case of tort litigation, lawyers look at the effective safety of the car to determine whether the company demonstrated negligence or not. Often, the basis of these faults derives from common sense deductions about human decision-making; lawyers present evidence that excuses or persecutes clients' actions, which the jury affirms or denies.

When we introduce self-driving cars, this process becomes near impossible, as the lawyers, judge, and jury require advanced knowledge of artificial intelligence (AI) to infer the factors affecting the self-driving car's decisions. As a result, lawyers often don't have substantial evidence to defend or prosecute clients or manufacturers in accidents involving self-driving cars. To bridge this gap between lawyers and AI, our capstone, CS 4991, is aiming to automatically explain self-driving cars' decision-making process by providing human-readable factors affecting the AI's pre-crash choices via explainable AI (XAI), a system designed to communicate the internal rationalizations used by AIs to guide their decisions, as elementarily seen in Figure 1.



Person: “*Why is image* J *labelled as a* Spider *instead of a* Beetle*?*”

ExplAgent: “*Because the arthropod in image* J *has eight legs, consistent with those in the category* Spider*, while those in* Beetle *have six legs.*”

Person: “*Why did you infer that the arthropod in image* J *had eight legs instead of six?*”

ExplAgent: “*I counted the eight legs that I found, as I have just highlighted on the image now.*” (ExplAgent shows the image with the eight legs counted).

Person: “*How do you know that spiders have eight legs?*”

ExplAgent: “*Because in the training set I was trained on, almost all animals with eight legs were labelled as* Spider.”

Person: “*But an octopus can have eight legs too. Why did you not classify image J as an octopus?*”

ExplAgent: “*Because my function is only to classify arthropods.*”

Figure 1. Example Explanation Dialogue between a Person and an Explanation Agent. (Miller, 2019, p. 8)

Most of AI involved in self driving tends to use deep neural nets (DNNs), which are a complicated series of numerical weights that can self-adjust to produce a desired action. Here, this internal AI maps visualizations of the road to throttle and steering as appropriate. Contrasting this, older developments of AI used human generated features that were independently weighted to determine the action. These features’ weights gave an explicit explanation for the AI’s decision. This transparency starkly contrasts modern DNNs, since “the features are no longer manually specified, [so] it is difficult to know what features are used by the deep learnt model for making predictions” (Atkinson et al., 2020, p. 26).

This becomes a significant issue in product liability claims, as “To make out a product liability claim for a safety defect with respect to an allegedly defective AI system, it would be necessary to produce an explanation that showed how safe or unsafe the system is”. In this respect, explainable AI (XAI) applied to DNNs has been shown to be a prospective solution, as "[its] key role… [would be] to show that a harm would have been substantially less likely to occur had the defendant exercised due care and precaution" (Fraser et al., 2022, p. 189). Using AI to solve legal issues is not as uncommon as most think, as AI is used in forecasting crashes, identifying and enforcing traffic violations at intersections, recognizing and identifying faces for intelligence agencies, assisting in interpreting radiological images for medical examiners in court, analyzing DNA on crime scenes, gun identification from gunshots, alongside many other examples (Rigano, 2019). This paper will first demonstrate how my capstone plans to implement XAI for the purposes of explaining self-driving cars' decisions and will then elaborate how product liability and AI break down legal systems via Susan Leigh Star's framework on infrastructure.

Our XAI Implementation

To the driver and even AI experts, the transformation behind DNNs is currently a "black box", where more abstract, human-understandable reasoning behind the transformation is unknown. Hence, when the AI makes a decision, the resulting action it takes is taken for granted. Many experts have tried to solve this issue though developments in saliency maps, which generate visual localizations of images to highlight key areas important to the AI. But these can easily fail to discern identifiable causes for an AI's decision, as the decision is not strictly tied to identifiable objects, as seen in Figure 2. These maps can even simply look like white noise, giving rise to no human-readable explanation. Newer novel approaches of GradCAM, a method generating heatmaps "showing important regions as localized by [a DNN] model", have improved over saliency maps by demonstrating improved time efficiency, but have also been shown to still generate distorted heat-maps that do not discern particular entities of interest to the AI (Mankodiya et al., 2022, p. 11). Other models, like LRP or a human observer, require the internal values of the model (Godel et al., 2016, p. 12), which can be impossible to obtain in a legal setting due to a company desire to protect proprietary software.

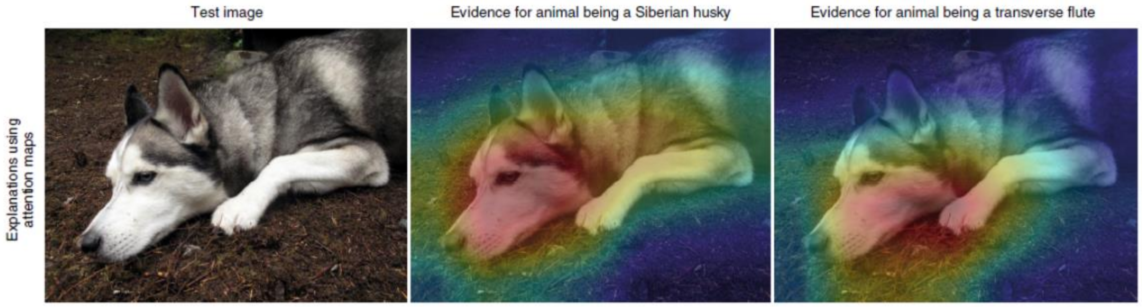
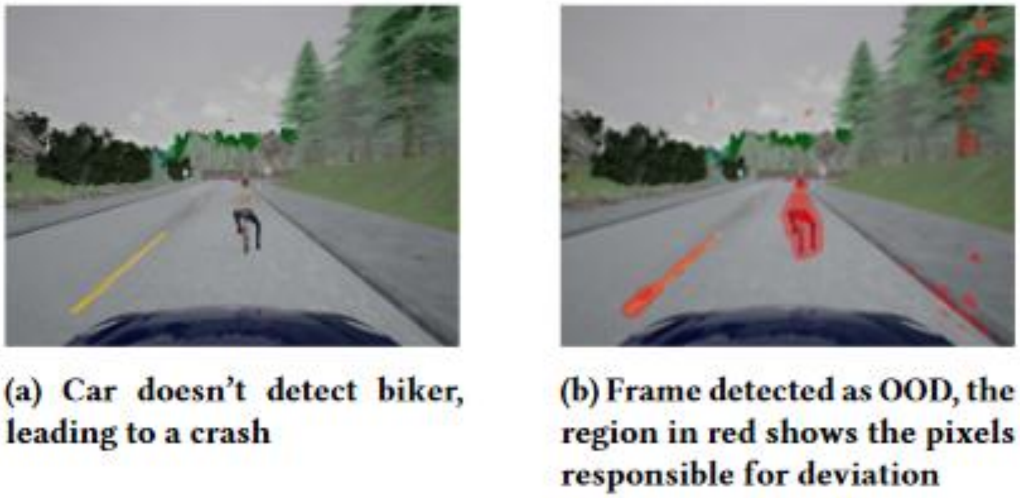

Test image
Evidence for animal being a Siberian husky
Evidence for animal being a transverse flute
Explanations using attention maps

Figure 2. Saliency does not explain anything except where the network is looking. We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. (Gohel et al., 2021, p. 3)

Thus, we want to automatically generate a handful of factors the AI finds important. We can do this using the results of novel research demonstrated in Figure 3. Here, the AI is highlighting fragments of the original image it finds most important based on which segments of the image it predicts is out of the norm of the training data. Since this is used on a neural net, which are also used in self-driving cars, we can apply this similarly. This will explain which fragments of the cars' perception are most important when making a decision like speeding up in the case of a biker. These actions will thus have explanations relative to the car's internal AI. Towards this, we are first constructing legally ambiguous scenarios in a car simulation tool called CARLA, then implementing the paper's proposed solution of detecting distribution shifts in data (Yang et al., 2022) to explain the car's decision process, ranking factors relative to the car's decisions in order of importance. In this way, we are explaining the self-driving car's faulty decisions by highlighting key areas that the AI hasn't been trained on. Building on this, we will utilize SafeBench, a software designed to dynamically generate scenarios in CARLA, to automate the generation of simulation scenarios and will look towards improved semantic understanding of said highlights by implementing a semantic embedding, where test outputs are compared to data variances which then become semantic dictionary mappings to explain such a difference. This can then leverage human domain knowledge to evaluate complicated highlights and allow semantic human understanding of the AI's rationalization (Zhang et al., 2019). We will likely also investigate LIME, an alternative natural language processing method that claims to provide visual explanations through text of any classifier (Ribeiro et al., 2016). Finally, novel developments in GradCAM have shown that utilizing temporal information through Gated Recurrent Units (GRUs) increases the accuracy of the XAI model such that GRUs will likely be an additional improvement to our implemented XAI (Karim et al., 2022).



(a) Car doesn't detect biker, leading to a crash

(b) Frame detected as OOD, the region in red shows the pixels responsible for deviation

Figure 3. Deviation from training data leads to a crash with biker as front object. Training data only had cars as front objects. Method detects deviations from in-distribution data for detecting such out-of-distribution data (OODs). (Yang et al., 2022, p. 226)

By providing internal, human-understandable embeddings of AI, we reduce the black box nature of AI such that there exist factual explanations for lawyers to utilize in tort litigation and crash fault. AI differentiates itself from standard products through this property, so alleviating it of this flaw allows judges and juries to evaluate the significance of negligence more objectively in either drivers or manufacturers.

Legal Hurdles Without XAI

When a car accident occurs, fault falls on the driver who caused the accident, but this becomes more ambiguous when it involves a self-driving car. With improvements to driving automation, we’re increasingly approaching full self-driving, as demonstrated by Hyundai’s NEXO currently being tested as a nearly fully autonomous vehicle, capable of driving in most conditions without human intervention (Sachdev, 2021). When we achieve full self-driving, since the driver isn’t technically driving, we might put blame on the manufacturer, but even they likely won’t understand why the crash happened due to the previously mentioned black box nature. For this reason, we must investigate the current legal framework to determine said fault.

In most cases of determining fault, many states employ either the “But For” or “Substantial Factor” test. In the “But For” test, lawyers ask, “Would the crash have happened if the driver hadn’t acted carelessly?” (Uslawessentials, 2015), and in the “Substantial Factor” test, lawyers ask what are the most important factors that caused the crash (Uslawessentials, 2015). The answers to these questions help determine the causation of the crash and therefore fault, but, when applied to self-driving cars, we can’t answer these questions; the inherent black boxing of the AI means we don’t know if the car is acting carelessly or what factors are causing the AI to crash. This further makes product liability claims near impossible, as they require factual explanations demonstrating negligence, which is not present if there exist no explanations for the AI’s decisions.

In many cases involving complex tort liability, lawyers tend to present expert witnesses to substitute their ignorance, as proprietary company information makes presenting the exact technical details competitively disadvantageous. In 2022, this was applied against the hotel booking company, Trivago, where an AI expert provided evidence that the AI system ranking hotels was not consistent with what should be expected from an AI ranking, leading to Trivago misleading customers on hotel prices (Fraser et al., 2022). In this case, the expert relied on actions that AI took that represented negligence, but this may not always be the case, as “the black box problem precludes anyone, including original programmers, from testifying with certainty as to the machine’s rationale… [where] the lack of an expert witness equivalent for AI further highlights the inadequacies of the legal process for handling machine [error]” (Jorstad et al., 2020, p. 15). Further, AI is incrementally outperforming humans, demonstrated by the recent development of GradCAM, which can predict crashes about 5 seconds before humans on average when compared to human attention maps generated by fixation points as shown in Figure 4 (Karim et al., 2022), which is causing a societal rise of self-driving cars and thus an equivalent increase in demand for expert AI witnesses. Since “[we can] expect [the autonomous car market] to increase tenfold in the next 5–7 years” (Kopestinsky, 2022), this will become an increasingly unreasonable solution.

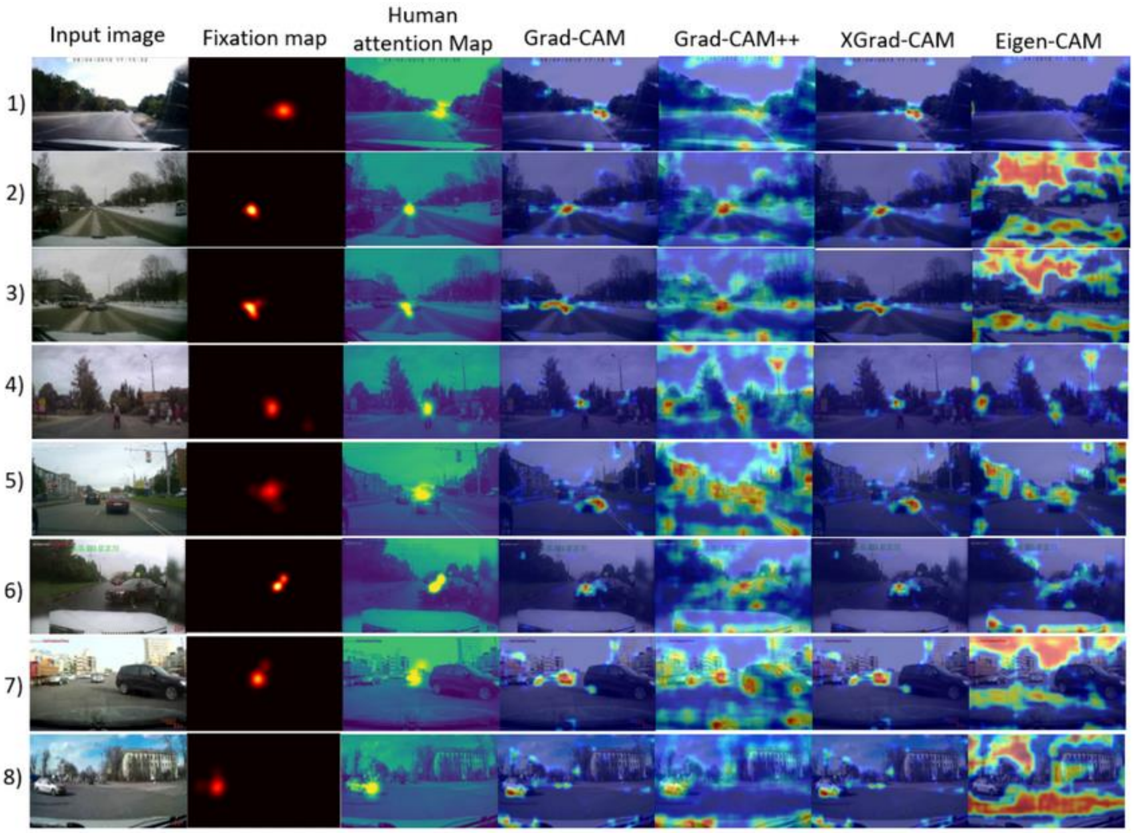


Figure 4. Visualizing the explanation generated by different XAI methods. (Karim et al., 2022, p. 9)

Despite this, many computer scientists and software engineers alike seem to maintain the master narrative of closed software. Closed software is software designed to prevent user tampering and decrease software transparency, further worsening the inherent black box nature of AI. Because of this alienation, non-Computer Science (CS) experts, such as lawyers, drivers, juries, and judges, interpret the AI as “intentional agents and apply the conceptual framework and psychological mechanisms of human behavior explanation to them”, misattributing the failure of simple tasks, like hitting a biker, as intentional, which reinforces mistrust in AI systems

(Graaf et al., 2017, p. 19). Given this, current XAI methods fail to improve this mistrust, as they are targeted towards developers of AI by not giving human-readable information. Instead, they provide maps and raw data engineers use to improve the AI models (Glomsrud et al., 2019). In the context of legally persecuting AI, current AI methods can be shown to contain three underlying issues of infrastructure described by Star's *The Ethnography of Infrastructure*. Firstly, AI is transparent in its use "…in the sense that it does not have to be reinvented… for each task, … invisibly supporting [the user]" (Star, 1999, p. 381). But, when legal negligence needs to be demonstrated, this transparency hides the internal decisions and, in extension, the manufacturer's design choices that are necessary in determining fault. Secondly, AI's operation and the software it's encapsulated in is learned as part of the CS community, where "[CS experts] acquire a naturalized familiarity with [the AI]" that outsiders don't have, leading CS-experts to overlook naturalized aspects such as DNN bias and adversarial data. (Star, 1999, p. 381). This is worsened by current XAI methods that are targeted towards developers, as this non-human readable format improves developer's AI intuition while alienating non-CS experts. In legal settings, without these experts, judges and juries are led to either no conclusion or, in the worst case, false conclusions of AI behavior after improperly personifying the AI's decision. Finally, autonomous driving becomes visible upon breakdown. When an autonomous vehicle (AV) makes a poor decision that leads to injury, the injuries are often fatal or severely debilitating often caused by driver's trust in the AV to drive safely. Star defines this property as when "the normally invisible quality of [safe driving] becomes visible when it [crashes]" (Star, 1999, p. 382). In legal settings, the damages these crashes cause when they do happen are expensive such that the outcome of the case has drastic consequences for the victims. Thus, this invisible quality of AVs becomes most visible to not only the victims, but also the court, as facts of the AI's decisions necessary to prove negligence become masked behind its black box nature.

Research Question and Methods

I am interested in how the legal system struggles to determine liability, accountability, and fault in the consequences of AI decisions. As self-driving and AI usage in general continues to grow, the legal system will be increasingly taxed with cases which, if not handled now, will result in unfair trials due to significant lacks in evidence. Additionally, an inherent lack of evidence in AI implicates an unpunishable ability to cause intentional harm through AI, as lack of explanation on the AI's decision-making can lead to the dismissal of tort litigation cases. This demonstrates an in inherent difficulty in finding "a responsible party, as so many different entities—software developers, hardware engineers, designers, and corporations—go into the creation of AI systems" (Sullivan et al., 2019) such that not only do the actions of the AI need to be considered, but the internal rationalizations need to be evaluated as well to reflect the actions of all parties.

To highlight these most prominent difficulties in AI tort litigation and fault, we will conduct four legal case studies involving an unintended consequence of AI actions and examine their legal consequences and/or rulings. These cases will be verified to not be ongoing such that they have definitive rulings to ensure all facets of the case are covered in court and the facts of the cases will be extracted from legal databases. The first case will be overviewing two car crashes involving autonomous vehicles. The first crash will be an accident caused by a self-driving Uber test car, where the autonomous vehicle struck and killed a pedestrian, Elaine Herzberg, despite the pedestrian being visible for multiple seconds before impact. The second crash for this case will be a Tesla Model S car crash with a semi-truck that killed the driver, Joshua Brown. The car failed to detect the truck for multiple seconds, similar to the first crash, and the car crashed into the side of the semi-truck at 74mph. The next case we will analyze will be a legal case between Trivago, a hotel booking website, and the Australian Competition and Consumer Commission (ACCC) for misleading advertising practices with the claim that Trivago misled consumer by representing hotel search results as impartial while simultaneously biasing it AI-driven search results towards hotels the paid Trivago higher fees for referrals. For the third case, we will analyze the usage of the AI algorithm Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) in 2012, which was a system used in judicial rulings to determine a defendant’s chance of reoffending and, thus, help judges determine defendants’ sentences. In 2016 though, it was found to be racially biased, giving double the sentences for black defendants compared to white defendants. Finally, the last case will investigate the copyright behind the creation of the "Portrait of Edmond de Belamy", which as an AI generated painting by an AI system called “The Next Rambrandt”.

The first case will be used as an example of how the legal system often fails to recognize the paradoxical nature of encouraging AI users to trust it while simultaneously using legal disclaimers to scapegoat liability for potential AI errors or harm caused by AI systems. The second case will highlight how the black-box nature of AI can enable its fraudulent usage and that the legal system has a resulting difficulty in determining whether malicious actors are at play. The third case will show how even the integration of AI into legal systems presents hurdles, where failures in such systems can have drastic and irreversible effects if not presently monitored. Finally, the last case will exemplify how the use of AI in creating art blurs the lines between human creativity and AI-generated works, thus posing new challenges for copyright and intellectual property laws as we refuse to consider AI as accountable actors.

These cases will be analyzed under Susan Leigh Star's framework by first identifying the AI system and the legal framework they're placed in as distinct infrastructures, which will be done by presenting each of the characteristics of infrastructure. This will help us define these systems as infrastructure so we can then study them. We do this under three analyses. The first is identifying the master and non-master narratives, where we'll question which assumptions the AI system and legal system make and which they do not. Next, we will surface invisible work by looking at what labor became visible when the AI failed to perform, or the legal system was obstructed. Finally, for each case, we will investigate paradoxes of infrastructure, which will be shown by small changes causing significant technical challenges for the AI system and/or the legal systems they're placed in.

Finally, the analysis of each case will then attempt to understand the AI and legal systems presented. This will be done by viewing both as constructed material artifacts designed for a pragmatic use to humans, a trace of record of activities designed to maintain an imprint of events such that it becomes primarily an information collecting device, or as a representation of the world designed to model the world in some abstract manner.

Concluding XAI Legal Benefits

In cases involving AI and tort litigation or fault, lawyers tend to be presented with a significant lack of evidence resulting from current AI's opaque nature. This presents significant legal hurdles, as both product liability and fault require factual explanations demonstrating harm for prosecution. Expert AI witnesses are current supplements for this, but the growing presence of AI will eventually outpace the quantity of AI experts.

Thus, integrating XAI in an auto-generative, human-readable format will better equip lawyers by providing evidence through AI decision explanations in cases involving fault and product liability. These generative descriptions of AI decisions will better inform juries and judges on the technical inner workings of AI, allowing them to stray away from applying the bias of intentionality and rather judge AI by the companies' diligence in training vs. highly implausible situations and visual difficulties causing crashes. By providing explanations to evolving AI technology alongside legal frameworks, we can expect the corner cases of AI causing legal issues to have improved fairness for all parties (Douma et. al., 2012).

With this, we expect to find that this improved XAI towards human-readability, transparency, and useability helps ameliorate the current lack of evidence involving tort litigation and fault applied to autonomous vehicles. Although "in state-of-the-art models, explainability and performance in terms of predication accuracy are often proportionally inverse" (Hacker et al., 2020, p. 418), we expect to show our model avoids this through its post-hoc descriptions, improving lawyers' usability by not requiring manufacturers to use such a method beforehand. With this, court cases will likely be less ambiguous when AI is involved, where the process will be more akin to determining negligence in non-AI systems. Finally, fundamental issues of determining negligence in AI is expected to boil down to a significant transparency involving AI, which XAI can solve by extrapolating internal features of decision-making.

References

Atkinson, K., Bench-Capon, T., & Bollegala, D. (2020). Explanation in AI and Law: Past, Present and Future. *University of Liverpool*. doi:10.1016/j.artint.2020.103387

Douma, F., Palodichuk, S., (2012). *Criminal liability issues created by autonomous vehicles*, Santa Clara Law Review, Volume 52, 1157.

Fraser, H., Simcock, R., & Snoswell, A. J. (2022). AI Opacity and Explainability in Tort Litigation. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 185–196). New York, NY, USA: Association for Computing Machinery. doi:10.1145/3531146.3533084

Glomsrud, J. A., Ødegårdstuen, A., Clair, A. L., & Smogeli, Ø. (2020). Trustworthy versus explainable AI in autonomous vessels. *Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC) 2019* (pp. 37-47). Sciendo. doi:10.2478/9788395669606-004

Gohel, P., Singh, P., & Mohanty, M. (2021, July 12). Explainable AI: current status and future directions. arXiv. doi:10.48550/arXiv.2107.07045

Graaf, M.D., & Malle, B.F. (2017). How People Explain Action (and Autonomous Intelligent Systems Should Too). *AAAI Fall Symposia.*

Hacker, P., Krestel, R., Grundmann, S., & Naumann, F. (2020). Explainable AI under contract and Tort Law: Legal incentives and technical challenges. *Artificial Intelligence and Law, 28*(4), 415-439. doi:10.1007/s10506-020-09260-6

Jorstad, K. T. (2020). Intersection of artificial intelligence and medicine: Tort liability in the technological age. *Journal of Medical Artificial Intelligence, 3*, 17-17. doi:10.21037/jmai-20-57

Karim, M. M., Li, Y., & Qin, R. (2022). Toward explainable artificial intelligence for early anticipation of traffic accidents. *Transportation Research Record: Journal of the Transportation Research Board, 2676*(6), 743-755. doi:10.1177/03611981221076121

Kopestinsky, Alex. "Self Driving Car Statistics for 2021: Policy Advice." *PolicyAdvice*, 5 Mar. 2022, https://policyadvice.net/insurance/insights/self-driving-car-statistics/.

Mankodiya, H., Jadav, D., Gupta, R., Tanwar, S., Hong, W., & Sharma, R. (2022). Od-Xai: Explainable AI-based semantic object detection for autonomous vehicles. *Applied Sciences*, *12*(11), 5310. doi:10.3390/app12115310

Miller, T. (2019). Explanation in artificial intelligence: Insights from the Social Sciences. *Artificial Intelligence*, *267*, 1-38. doi:10.1016/j.artint.2018.07.007

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of NAACL-HLT 2016 (Demonstrations)*, 97-101. doi:10.18653/v1/N16-3020

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

Sachdev, M. (2021, June 09). Explanation of the 6 levels of driving automation. Available at: https://blog.rgbsi.com/6-levels-of-driving-automation

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

Sullivan, H., & Schweikart, S. (2019). Are current tort liability doctrines adequate for addressing injury caused by Ai? *AMA Journal of Ethics*, *21*(2). doi:10.1001/amajethics.2019.160

Uslawessentials. (2015) “What's the Difference between ‘but for’ and ‘Substantial Factor’ Causation?” *Uslawessentials.* Available at: https://uslawessentials.com/2015214whats-the-difference-between-but-for-and-substantial-factor-causation/.

Yang, Y., Kaur, R., Dutta, S., & Lee, I. (2022). Interpretable Detection of Distribution Shifts in Learning Enabled Cyber-Physical Systems. *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*. doi:10.1109/iccps54341.2022.00027

Zhang, C., Shang, B., Wei, P., Li, L., Liu, Y., & Zheng, N. (2019). Building Explainable AI Evaluation for Autonomous Perception. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. doi:10.13140/RG.2.2.25648.20487