AI Opacity and Explainability in Tort Litigation

Henry Fraser
Queensland University of Technology, Centre for Automated Decision-Making and Society, Australia and Queensland University of Technology, Digital Media Research Centre, Australia
h5.fraser@qut.edu.au

Aaron J. Snoswell
Queensland University of Technology, Centre for Automated Decision-Making and Society, Australia and Queensland University of Technology, Digital Media Research Centre, Australia
a.snoswell@qut.edu.au

Rhyle Simcock
Queensland University of Technology, Centre for Automated Decision-Making and Society, Australia and Queensland University of Technology, Digital Media Research Centre, Australia
r.simcock@qut.edu.au

ABSTRACT

A spate of recent accidents and a lawsuit involving Tesla's 'self-driving' cars highlights the growing need for meaningful accountability when harms are caused by AI systems. Tort (or civil liability) lawsuits are one important way for victims to redress such harms. The prospect of tort liability may also prompt AI developers to take better precautions against safety risks. Tort claims of all kinds will be hindered by AI opacity: the difficulty of determining how and why complex AI systems make decisions. We address this problem by formulating and evaluating several options for mitigating AI opacity that combine expert evidence, legal argumentation, civil procedure, and Explainable AI approaches. We emphasise the need for explanations of AI systems in tort litigation to be attuned to the elements of legal 'causes of action' – the specific facts that must be proven to succeed in a lawsuit. We take a recent Australian case involving explainable AI evidence as a starting point from which to map contemporary Explainable AI approaches to elements of tortious causes of action, focusing on misleading conduct, negligence, and product liability for safety defects. Our work synthesizes law, legal procedure, and computer science to provide greater clarity on the opportunities and challenges for Explainable AI in civil litigation, and may prove helpful to potential litigants, to courts, and to illuminate key targets for regulatory intervention.

CCS CONCEPTS

• **Applied computing**; • **Law, social and behavioral sciences**; • **Law**; • **Computing methodologies**; • **Machine learning**; • **Artificial intelligence**; • **Social and professional topics**; • **Computing / technology policy**;

KEYWORDS

Explainable AI, Law, Evidence, Expert Evidence, AI Opacity, Civil Procedure, Negligence, Product Liability, Autonomous Vehicle, Accidents, Causation, Damages

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533084

ACM Reference Format:
Henry Fraser, Aaron J. Snoswell, and Rhyle Simcock. 2022. AI Opacity and Explainability in Tort Litigation. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3533084

1 INTRODUCTION

1.1 Background

Explainability is now a widely accepted goal for AI systems and developers, and a key condition for AI accountability [17, 23, 26, 58]. The cross-disciplinary field of 'Explainable AI' or 'XAI' has grown in search of solutions to the problem of AI opacity, combining knowledge from computer science, psychology, law, human-computer interaction, and other fields [1, 2, 54]. One domain where AI explanation and explainability will be important is in tort litigation (a sub-category of 'civil litigation'). Tort law, which encompasses doctrines such as negligence, product liability and misleading conduct, is an important piece of the AI accountability puzzle because it allows individuals and businesses to sue developers directly for AI-related wrongs or harms. At the time of writing, for example, there is a negligence and product liability lawsuit on foot in Texas, brought by police officers harmed by a collision allegedly caused by defects in a Tesla car's 'Autopilot' system [90].

Various kinds of AI opacity, categorised by Burrell, as well as Selbst and Barocas, are likely to hinder tort lawsuits in several ways [11, 68]. *Intentional secrecy* may obstruct effective inspection of AI systems, or obscure information about their development, if AI developers assert confidentiality protection for these 'trade secrets' [11]. Plaintiffs and judges generally lack the technical expertise required meaningfully to inspect and assess the system – what Burrell calls *technical illiteracy* – especially where the statistical reasoning behind the system's operation is *non-intuitive*, running against the grain of common sense [11, 68]. Indeed, the scale, complexity and non-linear characteristics of deep learning systems built on artificial neural networks may make them *inherently inscrutable*, even to their developers [68]. Even if AI developers or deployers are not intentionally secretive, *inadequate documentation* of the design and development of the system may result in key information about the process simply being lost.

1.2 Thesis and method

To sue successfully in torts such as negligence and product liability, plaintiffs will need to overcome these kinds of opacity, and explain how the AI system harmed them. As Wachter, Mittelstadt and Russell observe, explanations of algorithmic systems need not only serve the purpose of enhancing general public understanding of how the systems function. The scope and content of explanations should be driven by the goal or action they are intended to support [74].The kinds of explanation required in tort litigation are distinctive. It will not necessarily be useful to provide a general explanation about how an AI system works, or how it arrived at a certain decision. What is required is to prove to the court that some characteristic of the system, or some conduct in developing and deploying the system, or some relationship between the two, falls short of the relevant legal standard. It is necessary to provide explanations that are salient to the 'elements' of the 'cause of action' in suit - the set of facts that must be proven to succeed.

This paper considers and evaluates options for dealing with AI opacity in tort litigation. Focusing on the torts of *negligence*, *product liability* and *misleading and deceptive conduct*, we identify the most pressing challenges for litigants and courts, and highlight intriguing opportunities for the use of XAI methodologies. (We will use the term 'XAI' as a shorthand for computer science techniques, rather than the broader aspects of explanation of AI systems). We concentrate on the laws of England and Wales ('UK'), Australia and the United States, focusing on key mid-level issues that are fairly consistent across jurisdictions, and avoiding detailed analysis of jurisdictional idiosyncrasies. We reflect on the degree of access to AI systems required for tort claims; the legal procedural means of obtaining such access; what kinds of information different technical XAI methods can provide; the practices (such as the giving of expert evidence) available to convey complex technical information to courts; and the kinds of legal creativity required to bring all these parts together in factual and legal arguments capable of persuading judges.

1.3 Structure and key findings

In part 2 we discuss related literature. In part 3 we use the recent Australian case of ACCC v Trivago, one of the world's first tort cases involving XAI [80, 100], as a launching point for discussion. We distil several lessons about combining legal procedure, expert technical XAI evidence, and parsimonious legal reasoning to overcome AI opacity in a tort claim. In part 4, we assess the extent to which the approach in *Trivago* may be generalised to claims of negligence, or product liability for design defects. In part 5, we consider relatively straightforward extrapolations of the approach in *Trivago*. In part 6, we consider what to do when explanations of the kind achieved in Trivago are not feasible, highlighting opportunities for creative legal argument using innovative, probabilistic, counterfactual legal reasoning and XAI methods. The innovative approaches we formulate will not be easy to pull off. Still, they could allow some plaintiffs to work backwards from an accident to give concrete evidence of negligence or a defect where the cause of harm is initially unknown. In that respect, the strategies we explore have the potential to make inroads on one of the thorniest problems that plaintiffs face in *any* claim for negligence or product liability for a safety defect, regardless of whether AI is involved.

2 RELATED WORK

Existing scholarship has examined the impact of opaque algorithmic systems on contemporary evidentiary and litigation practices. There is considerable emphasis on the need for explainable AI in the context both of civil [5, 13, 18, 34] and criminal [43, 49, 75] lawsuits, and judicial review of government decision-making [8, 15, 43, 72]. Several scholars have expressed concern that AI opacity may impact due process [8, 15, 16]. One primary pre-occupation is the difficulties that arise when confidentiality and trade secrets in relation to valuable algorithms, data and know-how stand in the way of proper inspection and review of automated decisions [9, 15, 22, 43, 49, 75]. Another line of work examines whether (and how) civil liability laws such as negligence and product liability will adapt to accommodate harms caused by AI systems, for instance focusing on the difficulty of attributing responsibility for harms involving the actions of unpredictable autonomous systems, and the difficulties in applying traditional concepts such as foreseeability [7, 10, 14, 31, 32, 39, 41, 47, 56, 67, 69, 73, 77]. While these scholarly contributions lay important foundations, progress towards workable solutions requires consideration of practical challenges for explainable AI in context. There is a small body of work that delves deeper into context, sketching out partial solutions to specific problems in law and litigation [18, 34, 35, 45, 55, 74]. Our motivation is to add to this important body of work, contributing practically applicable knowledge at the intersection of tort law, evidence law, civil procedure and contemporary XAI methodologies.

3 CASE STUDY: ACCC V TRIVAGO

The recent Australian case of *ACCC v Trivago* shows how XAI, expert evidence, pragmatic application of civil procedure, and creative legal argumentation can be combined in tort litigation [80, 100]. (As the judgment on appeal reproduced the trial judgment very closely, our references to *Trivago* and paragraph pinpoints refer to the first instance judgment). The case involved misleading and deceptive conduct, and deceptive misrepresentation under sections 18 and 29(1)(i) of the Australian Consumer Law. The Australian Competition and Consumer Commission ('ACCC') alleged that Trivago, a travel and accommodation comparison website, mislead consumers into believing that the Trivago website would identify the cheapest available rates for a hotel room in response to a consumer search.

A focal point for the case was Trivago's "Top Position Offer", an algorithmically generated promotional deal presented to consumers as the cheapest available offer for an identified hotel (para 29). The case turned partly on whether the Top Position Offer was determined by price, or by commissions paid to Trivago by online booking providers (para 12). Accordingly, the parties needed to explain how Trivago's ranking algorithm generated the Top Position Offer. Each party called expert witnesses to give evidence, and both experts used XAI methods to identify the relative importance of each of the factors influencing the generation of the Top Position Offer (para 106-121). While there was some disagreement as to the true weighting of each factor, the evidence was sufficient to show that the algorithm's real characteristics and operations were not consistent with Trivago's claims about the algorithm.

The case showcased a range of approaches to overcoming intentional secrecy, technical illiteracy, and inherent inscrutability. It is particularly significant as an example of a plaintiff's generating explanations of opaque AI systems that are sufficient to succeed in a lawsuit with only partial access to the system. We deal with each of these matters in turn below.

3.1 Dealing with intentional secrecy with discovery and disclosure regimes

Most AI scholars are familiar with the *Loomis* case in the US, where *intentional secrecy* obscured important information about an AI sentencing tool in the appeal of a criminal sentence [99]. Odd though it may seem, outcomes of this kind are less likely in civil cases. The process of gaining access to information from opposing parties in litigation is known as 'discovery' or (in the UK) 'disclosure and inspection' of documents (we will use 'discovery' as a catch-all). In common law jurisdictions, including the UK, Australia and the US, it is common for parties to give discovery of confidential information in civil litigation, with special confidentiality regimes set up to protect the information from public disclosure [95, 104, 105, 108, 109, 113]. However, in some jurisdictions, especially in the US, protective orders and confidentiality regimes can be particularly burdensome, making serious inspection and analysis of software unduly difficult [50, 118].

It is encouraging (though not surprising in the Australian context) that the trial judge in *Trivago* used established discovery and court practices to deal effectively with problems of *intentional secrecy*. Trivago's *Top Position Algorithm* was confidential and commercially sensitive, but the ACCC's case depended on obtaining access to it. As is common, the court balanced the parties' competing interests with a confidentiality regime that limited access to the confidential material to members of the court, the independent experts retained by both parties and related support staff [79]. Under this confidentiality regime, the court was also closed for the oral testimony of the expert witnesses and any reference to commercially sensitive information about Trivago's proprietary algorithm was redacted from court transcripts. Judgement was delivered at a level of generality that avoided any express reference to confidential aspects of Trivago's *Top Position Algorithm* (see para 94).

3.2 Dealing with technical illiteracy - expert evidence and efficient procedure

Merely accessing information about an AI system through discovery will not be helpful if the *technical illiteracy* of plaintiffs and judges prevents them from interpreting it. (Those to whom the label is applied may prefer something less condescending – *lack of expertise*, perhaps – but to avoid confusion we'll stick with Burrell's established category). Fortunately, courts and plaintiffs are not required to go it alone. Courts regularly deal with complex technical information in matters ranging from patent disputes to financial regulation to medical negligence. The conventional way to handle these kinds of issues is to obtain the assistance of 'expert witnesses', and courts have a long history of doing so [29]. Consistent with convention, the parties in the *Trivago* case called on experts to assist the court in understanding what factors influenced whether the Trivago algorithm selected an offer as the *Top Position Offer*. Procedure for the delivery of expert evidence largely followed standard court practice where one or more parties intend to call expert witnesses to give evidence on similar issues [24]. The considerable power of the *ACCC*, as Australia's national competition and consumer law regulator, enabled it to dictate the agenda. It posed 9 questions about the algorithm to its own expert witness; which then became the main focus for factual dispute. Following two separate reports from each party's independent experts, the experts conferred to provide a joint report structured around those nine questions (para 92). The experts delivered their evidence concurrently in closed a court session similarly structured around the nine questions outlined in the joint report (see paras 108-109).

3.3 XAI with constrained access to AI systems

Trivago shows that the choice of explainable AI methodology, and therefore the type of evidence provided, will be shaped by the level of access to an algorithm and its underlying documentation. Each of the independent experts approached the method of explaining the weights of the algorithm's features in different ways. The ACCC expert 'reverse engineered' the *Top Position Algorithm* based on logged data that Trivago had provided in discovery for three separate dates across four capital cities (para 109). He chose this method because he had not, at the time, received the actual 'weights' used in the Trivago algorithm. In doing so, the Trivago expert used a Gradient Boosting Model, a prediction model composed of an aggregate of weak prediction models – in this case decision trees (para 108-111). The Trivago expert, by contrast, had full access to the system, and therefore did not need to rely on reconstructing a surrogate model.

Some explainable AI methods require full access to the algorithm and data on which the system was trained. Others, as the case highlights, only require access to an algorithm's outputs or inputs. The broader point is that XAI in litigation is still possible with less-than-total access to an AI system, its documentation, and the data on which it was trained – though naturally explanations are likely to be more robust if they are based on a greater degree of access. We take up this issue and its more general implications for tort litigation in more detail in part 4 below.

3.4 A parsimonious approach to explanation confined to the issues in suit mitigates inherent inscrutability

One of the most notable things about *Trivago* is the parsimony of the trial judge's reasoning in relation to XAI evidence. Commentators often point out that XAI cannot produce complete or perfectly reliable explanations of the operations of opaque AI systems [20]. But the court managed this problem pragmatically. Disagreement between the experts as to which of their explanations was superior was ultimately neutralised by the court's insistence that the independent experts confer for a joint report. This process efficiently brought to light the key points of contention, as well as points of agreement, simplifying and expediting analysis of the issues. The joint report revealed that, regardless of the XAI method, there was clear convergence on the fact that commissions paid were at least the second most significant in determining the *Top Position Offer*, and that this offer was often not the lowest priced option (para 121). In the court's view, this was sufficient to show that the operation of the algorithm contradicted claims that the algorithm made it easy for users to find the best priced options [80, 100]. The court was willing to accept that there was more than one method for explaining Trivago's Top Position Algorithm; that none of the methods used was infallible; but that convergence of different methods on key findings was persuasive evidence. This point is critical. Courts are accustomed to reconstructing events from less-than-perfect explanations: indeed, it is one of their main functions. In litigation, it is not necessary to explain perfectly how an AI system works. The aim of explanation is rather to provide *evidence* that is relevant to the cause of action in suit, and sufficiently reliable to assist the court in making an informed judgment about 'what happened'.

4 GENERALISING FROM TRIVAGO TO OTHER TORTS

The *Trivago* case is an encouraging example of litigants and a court cooperating to overcome AI opacity in a lawsuit concerning misleading conduct and representations. The approach to XAI in *Trivago* is likely to be relevant to parallel torts for deceptive conduct in the US [110] and UK [117] as well as other civil suits where the critical factual issue is whether there has been a misrepresentation about an AI system (e.g. suits for fraud, misrepresentation, mistake, estoppel, *etc.*). Certainly, *Trivago* serves as a good blueprint for the use of discovery and good management of expert evidence to significantly mitigate problems of *intentional secrecy*, *inadequate documentation*, *non-intuitiveness* and *technical illiteracy*. With regard to *inherent inscrutability*, *Trivago* shows how important it is that practices around discovery and disclosure in civil litigation converge on a level of access to algorithms and documentation that actually permits meaningful explanation – for example via reverse engineering.

But does the *Trivago* approach generalise well to other torts, where the explanation required involves more nuance and complexity than an inquiry into whether a statement about the AI system is true or false? We explore that question below, reflecting on:

- possible levels of access to AI systems that could be granted in the context of discovery;
- the basic elements of product liability for a design defect, and negligence (two 'bread and butter' torts).

This leads into a discussion in the sections 5 and 6 of the opportunities and challenges for proving these causes of action given varying degrees of access.

4.1 Three levels of access to AI system models

We outline three different levels of access that a plaintiff could be granted to an AI system's model, shown in Table 1. There is, of course, room for discussion about how to categorise degrees of access, but our categorisation at least provides a starting point for that discussion. In the first place, 'debugging access' to a model is the most comprehensive. It entails a level of access to a model's function that the model developer would herself have during development or to identify bugs in the function of their model. For instance, the defendant could provide a virtual machine or Docker image with all supporting libraries, the development environment, and complete supporting infrastructure installed, and with the model implementation code and a checkpoint file with deployed model parameters.

Perhaps more realistically, a plaintiff could be granted 'query access' to a model, which amounts to the ability to specify arbitrary model inputs and observe the computed model outputs. The defendant could, for example, provide a compiled binary program that included an implementation of the AI system to be run locally by the plaintiff. Alternatively, the developer could provide an Application Programming Interface (API) for a server which hosts a running instance of the AI system.

Finally, the weakest form of access is 'descriptive access', where the plaintiff does not have a model instance that can be queried, as in 'debugging' and 'query' regimes. For example, if a defendant provides access to an AI system's algorithm, but does not provide a corresponding checkpoint file with the actual parameters of the model which are used at deployment time, then this amounts to descriptive access, and an expert witness will only be able to make general statements about the behaviour of the model as deployed. Descriptive access might even be confined to documentation about the data and training configuration that was used to construct the deployed AI system.

All this said, it is important to remember that AI systems consist of more than just a predictive model which is used in a vacuum. Technical elements used during training to derive a model's parameters (such as data, training configuration, loss functions *etc.*) may also prove useful to a plaintiff, to say nothing of the broader sociological factors such as the institutional, historic and cultural context in which a given system is developed. For instance, inspecting the dataset used to train an AI system might reveal systemic issues that could bear on the behaviour of the deployed system after training [63], while analysis of the sociological context of the system's development could reveal aspects of the defendant's practices and procedures that bear on questions of liability [21, 25, 57, 64]. We acknowledge that these will be important considerations in tort claims, however our focus here is on how AI explainability methods could contribute to overcoming AI model opacity, so we defer these considerations to future work.

Where a plaintiff seeks discovery of data or training configuration details, a defendant could provide labelled (input and output), or unlabeled (input only) data instance(s) to a plaintiff, which could in turn be drawn from training, testing, or validation datasets. On the other hand, a defendant may provide the input and/or output specification(s) (*i.e.* the data structures), without any specific data instances. To further complicate matters, the loss, gain, or reward function(s) optimized by a system (and associated hyper-parameters), as well as the optimizer (and associated hyper-parameters) used to train the system to its final configuration [3, 52, 62] are of similar import to the data.

The complexity of these inter-related parts leaves room for creativity in seeking discovery of AI system details on behalf of the plaintiffs. For example, a defendant might resist providing query access to their AI system's model, but if the system is available in a public fashion (for instance, a recommender engine underlying the function of generally accessible website, or a specific software version on a generally available autonomous vehicle), the plaintiff could systematically query this version of the system to gain a form of rate-limited query access to the AI system model. Recall, in *Trivago*, the decision to use a reverse engineering approach

Table 1: Levels of AI system access, and the corresponding affordances they offer a plaintiff

	Description of architecture, pseudo or actual model source-code without corresponding checkpoint files	Specify model inputs and observe model outputs	Incrementally step-through model processing, observe intermediate calculations
Debugging Access	✓	✓	✓
Query Access		✓	
Descriptive Access	✓		

by ACCC's expert was shaped by the level of system access provided under disclosure and the nature of the impugned ranking algorithm. Similarly, given discovery of certain combinations of the above elements (*e.g.* descriptive access to the model as well as labelled training data, and the loss function and optimizer used), plaintiffs might be able to re-construct a version of the AI system that mimics the behaviour of the true system as deployed.

4.2 What needs to be explained in product liability and negligence claims

The ACCC is an experienced and strategic regulator and litigant so it should come as no surprise that it framed its legal claim in Trivago in such a way as to simplify the XAI inquiry. The essential question that the XAI experts were called upon to answer was essentially a true or false question: were the value weights of the algorithm consistent with Trivago's marketing claims about how the algorithm worked? Other causes of action require explanations that go beyond Boolean queries, meaning the opacity and XAI challenges are tougher. We consider two 'bread and butter' torts: product liability claims for safety defects and negligence claims. Product liability may apply to AI embodied in physical products, but may also apply to purely virtual AI systems depending on jurisdiction [10, 36, 38, 48, 73, 81]. Negligence has a broad application to software whether physically embodied or provided as a service (one of the main reasons it is worth thinking about carefully as a mechanism for AI accountability).

4.2.1 Product liability for a safety defect. In a product liability suit for a safety defect in the UK, Australia and some US states, a plaintiff must prove they suffered harm as a result of a safety defect in the product. To show that the product is defective, the plaintiff must show that the safety of the product is not what persons are generally entitled to expect [73, 106, 107]. In other US states a key question in product liability for design defects is whether there exists a 'reasonable alternative design' [116]. The gist of the reasonable alternative design test is that the defendant could have created a modified version of the system that would been safer at a cost that is reasonable in relation to the degree of risk reduction [86]. In the UK (and throughout Europe) a product may be found defective even if there was no reasonable alternative design, and in Australia the status of the test is unclear [6, 78]. Regardless, evidence of a reasonable alternative design will generally be helpful in showing that a product does not meet the level of safety that the public is entitled to expect.

It is not, however, always straightforward to map the abstract concept of safety to some distinct feature of the system in the same way as it was possible for the ACCC to make their case merely by revealing the value weightings of the Trivago algorithm (though, as we show below, there are some circumstances where an approach of this kind may work). 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. But safety is an abstract, emergent property of a system. Perceptions of safety depend in part on social norms. Safety is hard to reduce to a Boolean true or false.

4.2.2 Negligence. If product liability involves explanation that is more complex than for misleading conduct claims, negligence is more complicated still. Negligence has 3 interacting elements. Plaintiffs in negligence must prove that:

- the defendant (the person being sued) owed the plaintiff a duty of care (*duty*);
- the defendant breached that duty by failing to exercise the degree of precaution that would have been reasonable in the circumstances (*breach*); and
- the breach of duty caused the harm to the plaintiff (*causation*) [88, 115].

The place for XAI here is in showing the causal connection between breach of duty in the development of the system and the harm caused. After all, one way to define an explanation is as an attribution of causal responsibility [42]. In the UK and Australia, plaintiffs in negligence must show, on the *balance of probabilities*, that a breach of duty was a *necessary condition* (or *sine qua non*) of the harm that they suffered [12, 82]. In other words, plaintiffs have to prove that it is *more likely than not* (more than 50% likely) that the harm would have been avoided *but for* the defendant's breach of the duty of care [87]. In the United States, some courts use variations on the "but for" test for causation, while others apply a "proximate cause" test, which treats causation as a matter of degree [93, 103]. Regardless of whether the "but for" or "proximate cause" test applies, a key role for AI inspection and explanation is to show that a harm would have been substantially less likely to occur had the defendant exercised due care and precaution.

Courts often face difficult questions about causation in circumstances where a harm is caused by the interaction of multiple risks and actors. We confine our analysis here to scenarios where it is clear that some problem with the AI system, and not some subsequent intervening factor or person, caused the harm. Our findings will nonetheless be relevant to suits (or even non-litigious investigations) about the responsibilities of AI users or third parties, as well as AI developers and procurers, because it will almost always be necessary explain the role of the AI system itself in accidents.

The task of explaining an AI system with a view to proving that it was negligently designed or developed is fraught, but in equal measure intriguing for XAI. On the one hand, assessing the reasonableness of design choices (or in product liability, whether there was a reasonable alternative design) is very difficult, even without contending with AI opacity. This assessment essentially requires courts to redesign the impugned product, making polycentric judgments involving trade-offs between multiple factors such as cost, safety, accuracy, utility and so on [12, 37, 46]. In part because of this difficulty, courts generally avoid deciding cases on design issues, and there have been relatively few successful negligent design cases [6, 12, 70].

On the other hand, the probabilistic ('balance of probabilities') and counterfactual ('but for') elements of the test for causation have a peculiar resonance with the statistical and counterfactual dimensions of XAI [54, 74]. In so far as XAI methods permit counterfactual assessments of how AI systems would behave with different data, model features or inputs, they could potentially provide evidence of causation and breach that is far more comprehensive and concrete than courts generally have in negligent design cases, or product liability cases where reasonable alternative design is in issue. We reflect below (particularly in part 6) on how this combination of challenges and opportunities plays out.

5 FEATURE-FOCUSED (TRIVAGO-LIKE) APPROACHES TO NEGLIGENCE AND PRODUCT LIABILITY WITH OPAQUE AI SYSTEMS

Let us evaluate the options available to a plaintiff in product liability or negligence for dealing with AI opacity. We consider more straightforward options here in part 5; and a more complex option in part 6. Essentially the simpler options considered here are extrapolations of the approach in *Trivago*, where the ACCC scoped the XAI inquiry in such a way as to resolve the issue in question by explaining a single set of characteristics of the algorithm: the 'weights'. The approach discussed in part 6 considers the case when this is unworkable because inscrutability is too deep, or there are multiple interacting factors to consider. Each option involves different degrees of system access; and each has different strengths and weaknesses. Generally, the simpler the option, the more it relies on expansive inference; and the more complex, the harder and more expensive it is to execute.

5.1 Avoid claims that depend on inspection of AI system

One way to avoid the difficulties with AI opacity entirely might be to choose claims that do not require inspection of the impugned AI system at all. For example, five police officers are currently suing Tesla for injuries caused by a collision that occurred when a vehicle's Autopilot system allegedly failed to detect two police cars engaged in a traffic stop with flashing emergency vehicle lights. They are suing not only for a design defect and for negligent design, but also for failures to warn of and remediate known problems, because the accident in issue was one in a series of more than a dozen that happened in similar circumstances (which are also under investigation by the US traffic safety watchdog) [111]. Focusing on failure to warn and remediate sidesteps the need to access and explain an opaque AI system.

The advantage of such an approach is, of course, that it simplifies the plaintiff's inquiry. It may not, however, always be possible or desirable to wait until several harms of the same kind have occurred to frame a product liability or negligence claim in this way. Moreover, plaintiffs may wish to enhance their chances of success in court, or gain leverage for an out-of-court settlement, by making multiple claims, including for negligent or defective design. That is what the Tesla plaintiffs have in fact done [90]. Indeed, activist plaintiffs may pursue strategic public-interest litigation with a view to incentivising greater care on the part of AI developers. Plaintiffs with goals such as these need to confront AI opacity more directly.

5.2 Rely on broad inferences based on descriptive, query, or debug-level model access

Some plaintiffs may be fortunate enough to need only descriptive access to an AI system in order to succeed with their claims. In a claim for a safety defect, for example, it may turn out that there is a straightforward relationship between a documented characteristic of the system and the safety of the system. If plaintiffs obtain descriptive access to a system and find, say, that the objective function is inherently incompatible with safety, this might be very persuasive evidence of a safety defect. Similarly, in a suit for negligent design or development of an AI system, the simplest way to prove breach of duty (but not necessarily causation) would be to obtain information about the design, development and deployment process, and a description of the functionality of the system, and identify some aspect of this process or description that is clearly inconsistent with what a reasonable developer or data scientist would do. This appears to be the approach taken by the US traffic safety regulator in its investigation of Tesla [112]. Compliance (or lack of compliance) with specific regulatory requirements, technical standards, guidelines, industry best practices or codes of conduct will all be relevant, though not necessarily determinative [59, 66].

What if the defendant does produce a description, but it is too general to admit of an inference that a particular act or omission caused the problem in issue? In such cases, the court may be persuaded to draw an inference of a safety defect (for product liability), or a breach of duty or causation (in negligence), from *inadequate documentation*, *intentional secrecy* or even the existence of opacity itself [34]. The Australian Human Rights Commission has in fact recommended the introduction of a rule for drawing such adverse inferences from unexplained AI opacity [5].

In negligence, courts may also draw an inference that breach of duty caused a harm in circumstances where the precise breach of duty and chain of causation is unclear, especially where the defendant is best placed to control the risks giving rise to that harm [91, 91]. Alternatively, courts may conclude that the nature of the accident itself is sufficient to support an inference that it was caused by breach of duty. This kind of inference is known as *res ipsa loquitur* (literally, 'the thing speaks for itself') [84]. The extent to which *res ipsa loquitur* arguments may be used to prove a negligence or product liability claim by themselves is, however, contested and differs from jurisdiction to jurisdiction [4, 85, 97].

The problem with these inferential approaches is that they require the court to make rather large leaps of reasoning, and are less likely to work where *inherent inscrutability* is deeper. It may be that an accident or safety risk inherent in a system is not reducible to a single feature or readily described set of features. Moreover, merely describing features does not tell you whether they could have been brought in line with the requisite level of safety by an exercise of due care (in negligence) or through a reasonable alternative design (in jurisdictions where that test is relevant to product liability). Questions of this kind require deeper consideration, possibly including a counterfactual element and empirical analysis, which we touch on in section 6.

5.3 Mixing explanations and inference

The next most complex type of approach – but still fundamentally a uni-dimensional one – might be to combine explanation with inference. For example, plaintiffs could seek to show, in relation to some accident or harm, that identified instances where harm occurs are not 'out of distribution' for the AI system's intended purpose. This would involve arguing that these circumstances are reasonable inputs that a developer should have foreseen, and taken precautions against; and that the fact of the accident in such circumstances by itself indicates a safety defect, or (in negligence) a breach of duty that caused an accident. We refer to such arguments as 'banality arguments' – similar to *res ipsa loquitur* arguments, but grounded in explanation of the kinds of inputs to which a system ought not to respond anomalously.

Banality arguments might be possible with descriptive access only, but would tend to require a more comprehensive body of descriptive information than the kinds of inference-based arguments described above. They would likely involve analysis of the comprehensiveness or inclusiveness of the AI system's utilised training data, and evidence to substantiate such claims might include design documentation, marketing material, or documentation describing the intended purpose(s) of the AI system. For instance in the Tesla claim, the plaintiffs specifically refer to promotional tweets from Tesla CEO Elon Musk as evidence about the claimed purposes of the 'Autopilot' mode in the Tesla vehicles [90]). Plaintiffs might also include appeals to statistical likelihood estimates of certain circumstances occurring (e.g. from census or other data sources).

While banality arguments do not ask quite so much of judges as the broader inferences described above, they are susceptible to the objection that any given accident is not in fact representative of the operation of the impugned AI system; that it was caused by something outside the AI developer defendant's control; or that the circumstances were not in fact banal or foreseeable. Without inspecting or testing the system itself, it is hard prove that a specific accident or event is really indicative of a broader problem with the safety of an AI system (for safety defect claims), or was actually caused by the defendant's breach of duty (negligence) rather than some other factor.

Where the court cannot be led to draw an inference from this kind of high-level picture of a system, it may be necessary to provide deeper explanations. As the *Trivago* case shows, reverse-engineering a proxy model from publicly available or provided data may produce useful evidence with minimal access. In the context of a negligence or product liability suit, a comparable approach might be to conduct a post-hoc global feature importance analysis. If such an analysis revealed that some safety-critical feature (such as road one- or two-way status for a vehicle routing algorithm) consistently contributes insufficiently to the model output, this would be a persuasive basis for drawing an inference that an inherent feature of the system caused the harm in issue – i.e. it would be evidence of a defect, and of a causal chain between the system's development and the harm in issue.

6 AN EMPIRICAL-COUNTERFACTUAL APPROACH TO NEGLIGENCE AND PRODUCT LIABILITY WITH OPAQUE AI SYSTEMS

What happens when the behaviour of a system, or its safety, is not readily attributable to some (relatively) straightforward characteristic inherent in the system – such as a loss function, or weights? Or what happens when it is hard to work out what design or development step caused the system to manifest that characteristic? In other words, what happens when inherent inscrutability is deeper?

One way of dealing with this problem would be to take an empirical approach to explanation that is focused on outcomes, rather than features or characteristics of the system itself. Here we present one possible method of deriving evidence useful to a claim for negligent design, or for product liability for a safety defect, illustrated in Figure 1. It involves (1) establishing a chain of events where an AI system error leads to the harm in question; then (2) using empirical XAI methods to obtain evidence of the extent to which the system creates risk of this 'error' happening (*i.e.* derive an error rate in relation to some set of inputs); and finally (3) using counterfactual XAI methods to show that some reasonable alternative steps in design or development would have reduced the risk or error rate by a legally significant degree.

What is the appeal of this empirical-counterfactual approach? Plaintiffs in suits for negligence may not necessarily be able to find clear evidence of breach of duty at the outset. Likewise, plaintiffs in product liability may not at the outset be in a position clearly to attribute an accident to an inherent defect. They may therefore be faced with the challenge of working backwards from an accident or harm itself to show that it is caused by a defect (for product liability) or a specific failure of precaution amounting to a breach of duty (for negligence) [12]. Courts generally deal with this problem by accepting impressionistic, common-sense evidence of whether a precaution would 'more probably than not' have prevented a harm, or whether a failure of precaution more probably than not caused a harm, and to draw inferences from circumstantial evidence [98]. But there is an opportunity for plaintiffs to furnish more compelling arguments about negligent or defective design by leaning in to the statistical nature of AI systems.

Stapleton (an eminent tort scholar) translated the *balance of probabilities* standard into a formula that calls for the plaintiff to show that the defendant's negligence increased the risk of the harm occurring by more than 100%, or that the risk would have been

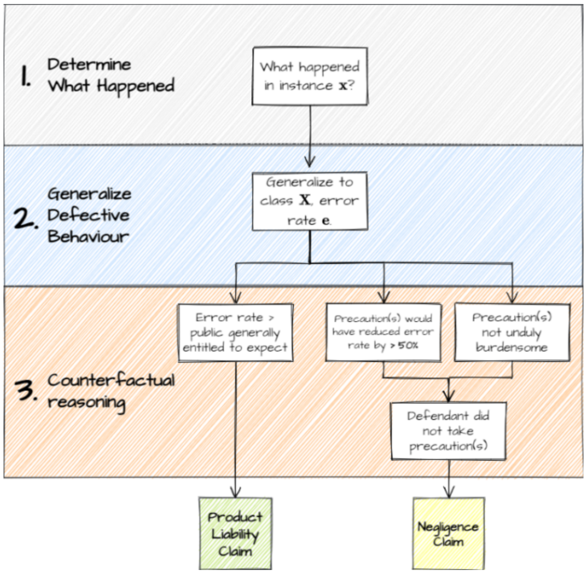
1. Determine What Happened
What happened in instance **x**?
2. Generalize Defective Behaviour
Generalize to class **X**, error rate **e**.
3. Counterfactual reasoning
Error rate > public generally entitled to expect
Precaution(s) would have reduced error rate by > 50%
Precaution(s) not unduly burdensome
Defendant did not take precaution(s)
Product Liability Claim
Negligence Claim

Figure 1: An empirical-counterfactual approach to building Negligence and Product Liability claims

reduced by more than 50% had the defendant taken the requisite degree of care – *i.e.* the negligence more than doubled the risk [71]. This approach is controversial, and Stapleton herself described the test in this way precisely because she wanted to make clear the problems with taking it seriously. One such problem is that the test requires plaintiffs accurately to ascribe a probability to the background risk in the first place, and to the risk as magnified by the lack of due care. The fact that plaintiffs are generally unable to provide such evidence is a key reason why courts take a 'common sense' [94], 'everyday' [89] approach to causation rather than adopting a 'philosophical or scientific theory of causation' [83].

None of this is to say that courts will reject good scientific or statistical evidence [27, 53, 92]. Counterfactual XAI could potentially allow a plaintiff in an AI case to provide concrete evidence of how alternative approaches to design and development might have prevented a harm. Such a plaintiff might even be at an advantage to a plaintiff in a normal negligence case, who may be unable to do much more than merely assert that things would have turned out differently had the defendant taken reasonable care (for example in designing and developing a product) [6, 46]. Counterfactual XAI methods may thus alleviate a key problem that confronts many plaintiffs in negligence.

6.1 Example – object detection problem

An example helps illustrate how such an empirical approach might work. Many robotic platforms, including autonomous vehicles, depend on object detection and recognition to avoid collisions and accidents. The Tesla suit involved a crash caused by apparent object detection failure in the presence of flashing lights. In circumstances of this kind there is a clear connection between the accuracy of an object detection or classification system and the level of safety of the system. Accuracy is something that can be determined empirically using testing and validation data, rather than by way of describing or explaining the nature of the system itself. It follows that the error rate of an object detection system makes a good target for empirical explanation.

Following the steps described above, an empirical approach to showing negligence or a safety defect in circumstances like those in the Tesla suit could perhaps then consist of: (1) establishing the nature of the error that occurred in the harmful instance x; (2) generalizing this to an error rate e over some class of inputs X; then (3)(a) arguing that this error rate is inconsistent with the level of safety the public is entitled to expect (for product liability); or (3)(b) providing evidence that the error rate would have been more than halved, and therefore the harm would *more likely than no*t have been prevented, with appropriate precaution in development of the object detection system (for negligence or a reasonable alternative design claim). See figure 1 above for a graphical representation.

6.2 Steps 1 and 2 – The empirical part

For the empirical approach to work, the plaintiff must first establish what happened in the specific instance x that caused harm (by 'instance', we refer to a configuration of inputs to the AI system). The functional or mechanistic component of this question will to a large degree be addressed through standard civil law discovery processes, with technical matter being interpreted by expert witnesses. In an ideal world, through discovery of logging information, corroboration with witness accounts, and re-construction of an event timeline, the parties would reach a point of agreement on the relevant facts, for instance that at some given point in time, one component of the AI system generated a particular output, which through a cascade of subsequent steps, led to the harm in point.

However, a difficulty will arise where the component of the AI system generating the faulty output is opaque – for instance a deep neural network. In this situation, the parties are left with a residual 'why' beyond the mechanistic 'why'. That is, what was it about the system's inputs in case x that led to the observed (and subsequently harmful) output y, and not, in fact, some other preferred non-harmful output y'? Is there, in other words, an underlying defect? This is where XAI methods – particularly those that operate locally, may be of use.

Local XAI methods seek to 'explain' a given output of an AI system model for a specific input, typically by producing a measure of the importance of the input factors (*e.g.* positive or negative correlation) determining the given output. Examples of Local XAI methods include Shapley values and SHAP force and dependence graphs [30, 51], LIME [61], and scoped rules or 'Anchors' [60]. The specific way feature importance is determined is irrelevant for our present discussion, but suffice to say different local XAI methods will vary in the requisite model access and/or dataset access. As such, the plaintiff's discovery at this point will need to be guided by their expert witness as to the most appropriate methodology, and the best way to proceed in requesting discovery of the implicated opaque component of the AI system.

For instance, if a plaintiff is granted query access to the allegedly faulty object detection and classification system, as well as logging details sufficient to re-construct the harmful input instance x, they might be able to apply local XAI methods such as SHAP or LIME, and show that the object classification system's defective output was positively correlated with the some feature (such as the presence of flashing lights on an emergency vehicle). This step clearly identifies a defective AI system input-output instance for product liability causes, and serves to identify the opaque AI system as a necessary component of the breach in a negligence cause.

Next, a plaintiff could attempt to generalize this defective behaviour to an empirical error rate across a broader class of inputs. This will require a deeper level of access to the AI system. Discovery of data instances used in training or testing the AI system model, and the use of instance-based XAI methodologies or exploratory data analysis methods may be helpful here.

Instance-based or example-based XAI methods attempt to explain characteristics of a dataset generally, or part of the decision boundary of a model, by selecting or generating data instances that are representative or non-representative (referred to as 'prototypes' and 'criticisms' respectively). These methods are most useful when the input data are of a form that humans can readily interpret (such a video footage or images). Examples of instance-based XAI methods include Case-Based-Reasoning [65], MMD-Critic [44], ProtoDash [33], and CLEAR [76]. Applying such a method to the instance where a harmful decision occurred might identify representative data that show the model also generates a harmful output in similar circumstances. Alternatively, applying such methods to the training data directly might indicate that the harmful instance x is not in-fact well represented in the training data (which might support an argument for breach of duty in negligence).

Returning to our example, if a plaintiff is granted access to a quantity of testing or validation data, they might be able to apply instance based XAI methods to unearth other data instances x' which lie near the decision boundary corresponding to the erroneous model output. Comparing these data with the harmful instance x might reveal common characteristics which could be used to guide a search for other similar data instances. The goal of this process is to establish a set of data instances X with shared characteristics (e.g. highways where an emergency vehicle is stopped in a shoulder lane) which can be used empirically to test the AI system. This set can then be tested to determine an error rate of the AI system over these inputs, e, which, as we show below can help furnish legal arguments for negligence and product liability.

6.3 Step 3 – The counterfactual part

The final step is to explain the legal significance of the information derived from steps 1 and 2. In product liability the essential point to be demonstrated is that that the identified rate of error e is greater than a rate of error e' that the public is generally entitled to expect for the class of inputs X In negligence, the point to be demonstrated is that the harm would more likely than not have been prevented had the defendant exercised due care (*e.g.* taken reasonable precautions in development). These might seem at first like the types of matters for which XAI is not well suited. Step 3 in a product liability suit depends in part on what the public thinks, rather than how the system works. Step 3 in a negligence action requires a number of extra steps in reasoning to show not only that the error rate was attributable to a failure of precaution; but also that appropriate precaution would more probably than not have prevented the particular instance of error and subsequent harm.

XAI may, however, be capable of producing uniquely clear and demonstrative evidence to that effect. If a plaintiff can show that some reasonable precaution(s) would more likely than not have prevented the harm, this is evidence of a breach of duty *and* of causation in negligence. It is also evidence of the existence of a reasonable alternative design for the purposes of product liability. This kind of reasoning is at its heart *counterfactual.* As a result, it opens the door to counterfactual XAI methods that allow plaintiffs to simulate what might have happened if such reasonable precautions had been taken.

Delving back into our example, steps 1 and 2 (described above) have furnished the plaintiff with an error rate, e for a class of inputs X. If the plaintiffs can show, using counterfactual XAI that some precaution, or set of precautions, in the design and development of the object classification system in a self-driving car would have reduced the error rate to *less than half* of the original rate e, then this means the precaution would more than halve the risk of the misclassification, with a commensurate reduction in the risk of the collision. Put another way, it means the failure to take the precaution more than doubled the risk of the accident-causing error, so that the error is *more likely than not* to be attributable to that failure. That is one way of showing 'on the balance of probabilities' that the harm would not have occurred 'but for' the defendant's failure of precaution. Since, under the empirical approach, the interrogation of the system has shown an actual error rate of e, and the plaintiff has shown that the error rate could have been substantially reduced to $< 0.5\ e$ with the counterfactual precautions, it is also straightforward to infer that the defendant did not in fact take such precautions. In jurisdictions where the proximate cause test holds sway, it may not even be necessary to show that precautions would have reduced the risk by more than half – demonstrating a smaller reduction in risk may be sufficient.

It will not suffice, though, merely to prove that the precaution would have prevented collision or reduced its risk substantially, and that the defendant did not take the precaution. The precaution must be 'reasonable', otherwise failing to take it will not be a breach of the duty of care. In determining whether a precaution is reasonable, courts consider (among other relevant factors) the probability that the harm would occur, the likely seriousness of the harm, the burden and cost of taking precautions to avoid or reduce the risk of harm, and the social utility of the activity that creates the risk of harm [96, 101, 102]. In other words, the court conducts an informal cost-benefit analysis of precautions asserted to be required to meet the duty of care [28]. In cases of negligent design, questions of reasonableness of design choices (and whether some reasonably feasible alternative would have prevented the harm) are very thorny, because design choices generally involve polycentric or multi-factor trade-offs that are difficult simply to characterise as reasonable or unreasonable [37, 46].

So how might a plaintiff show that a precaution would have more than halved (or at least substantially reduced) the error rate, *and* that it is reasonable, using XAI? The precise means of generating the explanation will depend on the AI system being discussed, and depending on the level of model access granted in discovery this could take multiple forms – for instance, with debugging access to a model, an expert witness may be able to literally train alternate versions of the AI system model of interest under varying dataset perturbations, hyper-parameter adjustments, or other architectural or design choices, comparing the performance in each case with that of the deployed system. Alternatively, under query and/or descriptive model access only, an expert witness might be able to 'reverse engineer' a surrogate model that mimics the general behaviour of the AI system, but without the defective outputs. Even if this model does not internally function the same as the deployed AI system, and even if the general performance of the surrogate is reduced compared to the deployed system, the fact that it could be generated under the constraints of inspecting an AI system in the context of litigation is at the very least a good indication that the defendant could have adopted such precautions at a reasonable cost. The proviso is of course that the surrogate model would have been attainable given the state of the art at the time the defendant developed the actual system.

This counterfactual approach bears some resemblance to the kind Wachter, Mittelstadt and Russel espoused in an influential article a few years ago. They advocated a counterfactual method of explanation that would inform subjects of an automated decision how the outcome would have changed if their circumstances or conduct were different. There is, however, a key difference in the goals of counterfactual explanations of that kind, and the kind we describe here. The objective of the counterfactuals proposed by Wachter, Mittelstadt and Russel is to explain how an automated decision might have changed given different 'nearby possible worlds' where there was something different about the *subject of an AI system's decision*. In tort suits of the kind we contemplate, by contrast, the objective is to give a sense of 'nearby possible worlds' where there was something different about the *AI system itself*, such that the system would not have caused (or would have been less likely to cause) a particular accident. In short, our approach focuses on counterfactual changes to the development of the AI system, while theirs looks at changes in inputs to the already-developed AI system.

The approach we have sketched here is not intended to be formulaic or prescriptive, but rather illustrative of the kind of creative argumentation made possible when contemporary XAI practices and legal thinking are brought to bear on opaque AI systems. Granted it has several drawbacks. From a legal standpoint, the major challenge will be in convincing a court that it really would have been reasonable for the defendant to have taken the measures identified through counterfactual XAI as risk-reducing. That depends first on the plaintiff's expert witness actually identifying such measures (which is not guaranteed). Then defendants may argue that, even though the marginal cost of some particular precaution is low relative to the quantum of risk, it would be prohibitively expensive to take equivalent precautions against every one of the very many possible mechanisms through which risks might eventuate [19, 40, 67]. Rather than being an objection to the approach we have described though, this is a more general concern about the difficulty of setting a sensible, balanced standard of 'reasonable care' for systems that are unpredictable in nature; and for which it is impossible to foresee and prevent every possible mechanism of harm. It is beyond the scope of this paper to advocate for some particular standard of care. Courts will have to develop an approach to breach of duty that is attuned to the peculiarities of AI – and they will have to make the most of the flexibility afforded by the concept of 'reasonableness'. But judges can only develop the law when there are cases before them. For that to happen there must be plaintiffs willing and able to present compelling evidence of the causal connection between decisions or omissions in the development of complex AI systems and subsequent harms. Our approach has the potential to contribute to that process.

The major technical challenge is that our method requires query access to the AI system – that is, the ability to interrogate the model under varying input configurations. If this level of access is not granted in discovery, other approaches might be needed. This type of model interrogation could also be particularly expensive or difficult in circumstances where the AI system is physically embodied (e.g. an autonomous vehicle) rather than virtual (e.g. a recommender system) – because the degree of control over the AI system's inputs is necessarily less in the former scenario. Where a plaintiff is not granted discovery access to (part of) the AI system dataset, then they might need to furnish their own dataset matching the data structure of the AI system, which would expose the line of argumentation to criticism about the representativeness of the computed error rates.

Nonetheless, if plaintiffs could pull it off, the actual *demonstration* of reasonableness of alternative development and design measures by way of counterfactual XAI methods could be superior to the impressionistic cost-benefit analysis courts often use in judging the reasonableness of the design of allegedly defective, or negligently designed, products [6].

7 CONCLUSIONS

Our analysis of opacity in civil litigation indicates that there are several key conditions without which it will be very difficult for plaintiffs meaningfully to explain AI systems for the purposes of civil litigation. At an absolute minimum, intentional secrecy and inadequate documentation cannot be permitted to stand in the way of accountability and access to justice in relation to AI-caused harms. Drawing adverse inferences from intentional secrecy and inadequate documentation would be one way for courts to incentivize better practice on the part of AI developers. If courts do not do so, legislative intervention may be warranted [5]. There is a strong case for mandatory levels of documentation and logging (Europe leads the way here [114]) since several of the options we consider above depend on detailed descriptive access to AI systems which is not possible without such documentation. It will also be important for courts to converge on an approach to discovery, disclosure and inspection that ensures, at a minimum, access to relevant documentation of the development of a system, and query access to the system subject to appropriate confidentiality regimes - as in *Trivago*. In jurisdictions where 'protection orders' and confidentiality regimes stand in the way of meaningful access of this kind [50], regulatory intervention would be justified.

More generally, commentators make much of the challenges posed by AI opacity for accountability and law. Our analysis shows that explaining opaque AI systems in the context of litigation presents not only challenges, but also exciting opportunities for creative legal and technical thinking. In presenting these opportunities to use *existing* legal and technical tools, it is not our intention to advocate for a *laissez faire* approach to AI regulation, or to oppose any particular suggestions for regulatory interventions. Nonetheless, we are pragmatists. Regulation takes time and may have powerful opponents. In the meanwhile, we should not deny redress to victims of AI-caused harm, merely on the ground that we would prefer to wait for better laws. Negligence, product liability and other torts are likely to play a significant role in AI accountability. They are the tools we have. If they are to be used to best effect, litigants need a roadmap for overcoming AI opacity. We hope in this paper to have sketched the outlines of such a roadmap.

ACKNOWLEDGMENTS

We are grateful to Professor Dan Hunter for his advice and feedback on this paper. This research was funded by the Australian Research Council Centre of Excellence for Automated Decision-Making and Society (CE200100005). Rhyle Simcock receives support through a higher degree research scholarship funded by the Digital Media Research Centre at Queensland University of Technology.

REFERENCES

[1] Abdul, Ashraf *et al*, 'Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda' in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018) 1

[2] Adadi, Amina and Mohammed Berrada, 'Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)' (2018) 6 *IEEE access* 52138

[3] Akbari, Ali *et al*, 'A Theoretical Insight Into the Effect of Loss Function for Deep Semantic-Preserving Learning' [2021] *IEEE Transactions on Neural Networks and Learning Systems* 1

[4] Atiyah, PS, 'Res Ipsa Loquitur in England and Australia' (1972) 35(4) *The Modern Law Review* 337

[5] Australian Human Rights Commission, *Human Rights and Technology* (Final Report, 2021)

[6] Barker, Kit, *The Law of Torts in Australia* (Oxford University Press, 5th ed, 2012)

[7] Benhamou, Yaniv and Justine Ferland, 'Artificial Intelligence & Damages: Assessing Liability and Calculating the Damages' (2020) forthcoming *Leading Legal Disruption: Artificial Intelligence and a Toolkit for Lawyers and the Law*

[8] Bloch-Wehba, Hannah, 'Access to Algorithms' (2019) 88(4) *Fordham Law Review* 1265

[9] Brauneis, Robert and Ellen P Goodman, 'Algorithmic Transparency for the Smart City' (2018) 20 *Yale Journal of Law and Technology* 103

[10] Buiten, Miriam, Alexandre de Streel and Martin Peitz, *EU Liability Rules for the Age of Artificial Intelligence* (Centre on Regulation in Europe, March 2021)

[11] Burrell, Jenna, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3(1) *Big Data & Society* 1

[12] Cane, Peter and James Goudkamp, *Atiyah's Accidents, Compensation and the Law* (Cambridge University Press, 9th ed, 2018) <https://www.cambridge.org/core/books/atiyahs-accidents-compensation-and-the-law/47E05F02588747B1B0AA27FD9F732391>

[13] Cauffman, Caroline, 'Robo-Liability: The European Union in Search of the Best Way to Deal with Liability for Damage Caused by Artificial Intelligence' (2018) 25(5) *Maastricht Journal of European and Comparative Law*

[14] Chinen, Mark A, 'The Co-Evolution of Autonomous Machines and Legal Responsibility' (2016) 20 *Vanderbilt Journal of Law and Technology* 338

[15] Citron, Danielle Keats, 'Technological Due Process' (2007) 85(6) *Washington University Law Review* 1249

[16] Cobbe, Jennifer, 'Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making' (2019) 39(4) *Legal Studies* 636

[17] Dawson, Dave *et al*, *Artificial Intelligence: Australia's Ethics Framework* (Discussion Paper, Data61 CSIRO, 2019)

[18] Deeks, Ashley, 'The Judicial Demand for Explainable Artificial Intelligence' (2019) 119(7) *Columbia Law Review* 1829

[19] Desai, Deven R and Joshua A Kroll, 'Trust but Verify: A Guide to Algorithms and the Law' (2017) 31(1) *Harvard Journal of Law & Technology* 1

[20] Doshi-Velez, Finale and Been Kim, 'Towards A Rigorous Science of Interpretable Machine Learning' [2017] *arXiv:1702.08608 [cs, stat]* <http://arxiv.org/abs/1702.08608>

[21] Ehsan, Upol *et al*, 'Expanding Explainability: Towards Social Transparency in AI Systems' in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021) 1

[22] Engstrom, David Freeman and Daniel E Ho, 'Algorithmic Accountability in the Administrative State' (2020) 37 *Yale J. on Reg.* 800

[23] *Ethics Guidelines for Trustworthy AI* (European Commission: High-level Expert Group on Artificial Intelligence, 8 April 2019) <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

[24] Federal Court of Australia: Case Management Handbook (2014)

[25] Felzmann, Heike *et al*, 'Transparency You Can Trust: Transparency Requirements for Artificial Intelligence between Legal Norms and Contextual Concerns' (2019) 6(1) *Big Data & Society*

[26] Fjeld, Jessica *et al*, 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI' [2020] (2020–1) *Berkman Klein Center Research Publication*

[27] Fleming, John G, 'Probabilistic Causation in Tort Law' (1989) 68(4) *The Canadian Bar Review* 661

[28] Fraser, Henry L and Jose-Miguel Bello y Villarino, 'Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation' [2021] *Working Draft* <https://papers.ssrn.com/abstract=3960461>

[29] Freckelton, Ian and Hugh Selby, *Expert Evidence: Law, Practice, Procedure and Advocacy* (Lawbook Co, 5th ed, 2013)

[30] Fryer, Daniel, Inga Strümke and Hien Nguyen, 'Shapley Values for Feature Selection: The Good, the Bad, and the Axioms' [2021] *arXiv:2102.10936 [cs, stat]* <http://arxiv.org/abs/2102.10936>

[31] Goldenfein, Jake *et al*, 'Through the Handoff Lens: Competing Visions of Autonomous Futures' (2020) 35(4) *Berkeley Technology Law Journal* 835

[32] Graham, Kyle, 'Of Frightened Horses and Autonomous Vehicles: Tort Law and Its Assimilation of Innovations' (2012) 52(4) *Santa Clara Law Review* 1241

[33] Gurumoorthy, Karthik S, Amit Dhurandhar and Guillermo Cecchi, 'Protodash: Fast Interpretable Prototype Selection' [2017] *arXiv preprint arXiv:1707.01212*

[34] Hacker, Philipp *et al*, 'Explainable AI under Contract and Tort Law: Legal Incentives and Technical Challenges' [2020] *Artificial Intelligence and Law* 1

[35] Hall, Stuart W, Amin Sakzad and Kim-Kwang Raymond Choo, 'Explainable Artificial Intelligence for Digital Forensics' n/a(n/a) *WIREs Forensic Science* e1434

[36] Hayward, Benjamin, 'What's in a Name?: Software, Digital Products, and the Sale of Goods' (2016) 38(4) *Sydney Law Review* 441

[37] Henderson, James A, 'Judicial Review of Manufacturers' Conscious Design Choices: The Limits of Adjudication' (1973) 73(8) *Columbia Law Review* 1531

[38] Howells, Geraint, Christian Twigg-Flesner and Chris Willett, 'Product Liability and Digital Products' in Tatiana-Eleni Synodinou *et al* (eds), *EU Internet Law: Regulation and Enforcement* (Springer International Publishing, 2017) 183 <https://doi.org/10.1007/978-3-319-64955-9_8>

[39] Hubbard, F Patrick, 'Allocating the Risk of Physical Injury from "Sophisticated Robots": Efficiency, Fairness, and Innovation' [2016] *Robot Law* <https://www-elgaronline-com.ezp01.library.qut.edu.au/view/edcoll/9781783476725/9781783476725.00009.xml>

[40] Hubbard, F Patrick, 'Sophisticated Robots: Balancing Liability, Regulation, and Innovation' (2014) 66 *Fla. L. Rev.* 1803

[41] Huberman, Pinchas, 'Tort Law, Corrective Justice and the Problem of Autonomous-Machine-Caused Harm' (2021) 34(1) *Canadian Journal of Law & Jurisprudence* 105

[42] Josephson, John R and Susan G Josephson, *Abductive Inference: Computation, Philosophy, Technology* (Cambridge University press, 1996) <https://www.google.com.au/books/edition/Abductive_Inference/uu6zXrogwWAC?hl=en&gbpv=1&pg=PA1&printsec=frontcover>

[43] Katyal, Sonia K, 'Private Accountability in the Age of Artificial Intelligence' (2019) 66(1) *UCLA Law Review* 54

[44] Kim, Been, Rajiv Khanna and Oluwasanmi O Koyejo, 'Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability' (2016) 29 *Advances in neural information processing systems*

[45] Krafft, Tobias D, Katharina A Zweig and Pascal D König, 'How to Regulate Algorithmic Decision-Making: A Framework of Regulatory Requirements for Different Applications' *Regulation & Governance* <http://onlinelibrary.wiley.com/doi/abs/10.1111/rego.12369>

[46] Lemer, Bruce, 'Strict Products Liability: The Problem of Improperly Designed Products' (1982) 20(2) *Osgoode Hall Law Journal* 250

[47] Lemley, Mark A and Bryan Casey, 'Remedies for Robots' (2019) 86(5) *The University of Chicago Law Review* 1311

[48] Lindsay, David F, Evana Wright and Genevieve Wilkinson, *Regulating to Protect Security & Privacy in the Internet of Things (IoT): Draft Report* (SSRN Scholarly Paper No 4052068, Social Science Research Network, 11 February 2022) <https://papers.ssrn.com/abstract=4052068>

[49] Liu, Han-Wei, Ching-Fu Lin and Yu-Jie Chen, 'Beyond State v Loomis: Artificial Intelligence, Government Algorithmization and Accountability' (2019) 27(2) *International Journal of Law and Information Technology* 122

[50] Loren, Lydia Pallas and Andy Johnson-Laird, 'Computer Software-Related Litigation: Discovery and the Overly-Protective Order' (2012) 6 *Lewis and Clark Law Review* 75

[51] Lundberg, Scott M and Su-In Lee, 'A Unified Approach to Interpreting Model Predictions' in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017) 4768

[52] Masnadi-Shirazi, Hamed and Nuno Vasconcelos, ‘On the Design of Loss Functions for Classification: Theory, Robustness to Outliers, and SavageBoost’ in *Proceedings of the 21st International Conference on Neural Information Processing Systems* (2008) 1049

[53] Mengersen, K, SA Moynihan and RL Tweedie, ‘Causality and Association: The Statistical and Legal Approaches’ (2007) 22(2) *Statistical Science* 227

[54] Miller, Tim, ‘Explanation in Artificial Intelligence: Insights from the Social Sciences’ (2019) 267 *Artificial Intelligence* 1

[55] Nutter, Patrick W, ‘Machine Learning Evidence: Admissibility and Weight’ 21 40

[56] Rachum-Twaig, Omri, ‘Whose Robot Is It Anyway?: Liability for Artificial-Intelligence-Based Robots’ (2020) 2020(4) *University of Illinois Law Review* 1141

[57] Raji, Inioluwa Deborah *et al*, ‘Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing’ <https://arxiv.org/abs/2001.00973v1>

[58] *Recommendation of the Council on Artificial Intelligence* (Organisation for Economic Cooperation and Development Legal Instruments (OECD), 5 May 2019)

[59] Reisman, Dillon *et al*, *Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability* (AI Now, April 2018)

[60] Ribeiro, Marco Tulio, Sameer Singh and Carlos Guestrin, ‘Anchors: High-Precision Model-Agnostic Explanations’ in *Proceedings of the AAAI Conference on Artificial Intelligence* (2018)

[61] Ribeiro, Marco Tulio, Sameer Singh and Carlos Guestrin, “‘Why Should i Trust You?” Explaining the Predictions of Any Classifier’ in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016) 1135

[62] Saito, Sean and Sujoy Roy, ‘Effects of Loss Functions and Target Representations on Adversarial Robustness’ [2018] *arXiv preprint arXiv:1812.00181*

[63] Sambasivan, Nithya *et al*, ‘Everyone Wants to Do the Model Work, Not the Data Work: Data Cascades in High-Stakes AI’ in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2021) 1 <https://doi.org/10.1145/3411764.3445518>

[64] Sartori, Laura and Andreas Theodorou, ‘A Sociotechnical Perspective for the Future of AI: Narratives, Inequalities, and Human Control’ (2022) 24(1) *Ethics and Information Technology* 4

[65] Schoenborn, Jakob M *et al*, ‘Explainable Case-Based Reasoning: A Survey’

[66] Selbst, Andrew, ‘An Institutional View Of Algorithmic Impact Assessments’ 35(Forthcoming) *Harvard Journal of Law & Technology*

[67] Selbst, Andrew D, ‘Negligence and AI’s Human Users’ (2020) 100 *BUL Rev.* 1315

[68] Selbst, Andrew D and Solon Barocas, ‘The Intuitive Appeal of Explainable Machines’ (2018) 87(3) *Fordham Law Review* 1085

[69] Smart, William D, Cindy Grimm and Woodrow Hartzog, ‘An Education Theory of Fault For Autonomous Systems’ (2021) 2 *Notre Dame Journal on Emerging Technologies* 33

[70] Stapleton, Jane, ‘Compensating Victims of Diseases’ (1985) 5 *Oxford J. Legal Stud.* 248

[71] Stapleton, Jane, ‘The Gist of Negligence Part II: The Relationship Between" Damage" and Causation"’ (1988) 104 *LQR* 389

[72] Tomlinson, Joe, Katy Sheridan and Adam Harkens, ‘Judicial Review Evidence in the Era of the Digital State’ [2020] *Public Law* (forthcoming)

[73] Vladeck, David C, ‘Machines Without Principals: Liability Rules and Artificial Intelligence’ (2014) 89 *Washington Law Review* 117

[74] Wachter, Sandra, Brent Mittelstadt and Chris Russell, ‘Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR’ (2018) 31 *Harvard Journal of Law & Technology* 841

[75] Wexler, Rebecca, ‘Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System’ (2018) 70(5) *Stanford Law Review* 1343

[76] White, Adam and Artur d’Avila Garcez, ‘Measurable Counterfactual Local Explanations for Any Classifier’ [2019] *arXiv preprint arXiv:1908.03020*

[77] Yoshikawa, Jin, ‘Sharing the Costs of Artificial Intelligence: Universal No-Fault Social Insurance for Personal Injuries’ (2018) 21(4) *Vanderbilt Journal of Entertainment and Technology Law* 1155

[78] *A v National Blood Authority* [2001] 3 All ER 289

[79] ACCC v Trivago, Administrative Listing: 19 September 2019, Order No VID1034/2018

[80] Australian Competition and Consumer Commission (ACCC) v Trivago NV [2020] FCA 16

[81] Australian Competition and Consumer Commission v Valve Corporation (No 3) [2016] FCA 196

[82] Barnett v Chelsea and Kensington Hospital Management Committee [1969] 1 QB 428

[83] *Bennett v Minister of Community Welfare* 176 CLR 408

[84] *Byrne v Boadle* [1863] 159 ER 299

[85] *Clark v McLennan* [1983] 1 All ER 416

[86] *Connelly v Hyundai Motor Co* [2003] 351 F.3d 535, 541 (1st Cir. 2003)

[87] *Cork v Kirby MacLean Ltd* [1952] 2 All ER 402

[88] Donoghue v Stevenson [1932] AC 562 at 580

[89] *Fairchild v Glenhaven Funeral Services Ltd* [2002] 3 All ER 305

[90] *Fields vs Tesla, Plaintiff's Original Petition* (District Court, Harris County Texas, 2021-62207 / Court 80)

[91] *Fitzgerald v Penn* [1954] 91 CLR 268

[92] *Hotson v East Berkshire Area Health Authority* (1987) 2 All ER 909

[93] *IHS Cedars Treatment Ctr of DeSoto, Tex, Inc v Mason* [2004] 143 S.W.3d 794, 800 Texas

[94] *Medlin v State Government Insurance Commission* [1995] 182 CLR 1

[95] *Oneplus v Mitsubishi* [2020] EWHC 2641

[96] Paris v Stepney Borough Council [1951] AC 367

[97] *Schellenberg v Tunnel Holdings Pty Ltd* [1999] 200 CLR 121

[98] *Stapley v Gypsum Mines Ltd* [1953] AC 663

[99] *State v Loomis* 881 NW 2d 749 (Wis 2016) 754 (US)

[100] Trivago NV v Australian Competition and Consumer Commission (ACCC) [2020] FCAFC 185

[101] US v Carroll Towing, 159 F2d 169 (2d Cir 1947)

[102] *Wyong Shire Council v Shirt* (1980) 146 CLR 40

[103] California Civil Jury Instruction (CACI) 430

[104] California Code of Civil Procedure Section 2031.010 .

[105] Civil Procedure Rules Statutory Instrument 1998 No. 3132 (L.17) (England and Wales) Rule 31.5

[106] Competition and Consumer Act (Australian Consumer Law) 2010 (Cth)

[107] *Consumer Protection Act (UK) 1987*

[108] Federal Court Rules (Australia) 2011 Rule 7.23

[109] Federal Rules of Civil Procedure (US) (2020), Rule 26-37, 45

[110] Federal Trade Commission Act, 15 U.S.C. §§41-58,

[111] National Highway Traffic Safety Administration, Safety Issue ID PE21020

[112] NHTSA, Letter to Tesla of 31 August 2021 <https://Static.Nhtsa.Gov/Odi/Inv/2021/INIM-PE21020-84913P.Pdf>Accessed 10 January 2021.

[113] Practice Direction 51U – Disclosure Pilot for the Business and Property Courts (England and Wales), Para 15

[114] Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) | Shaping Europe's Digital Future 2021

[115] Restatement (Third) of Torts: Physical and Emotional Harm (2010).

[116] Restatement (Third) of Torts: Product Liability, (1998) 2(b)).

[117] The Consumer Protection from Unfair Trading Regulations 2008, Part II

[118] Schulman, Andrew, ‘Source Code Protective Orders, From the Perspective of a Source Code Examiner’, *DisputeSoft Blog* <https://www.disputesoft.com/source-code-protective-orders-from-the-perspective-of-a-source-code-examiner/>