BRIDGING THE CAUSATION GAP: EXAMINING THE UTILITY OF THE RES IPSA LOQUITUR DOCTRINE FOR AI-RELATED HARMS UNDER US TORT LAW

Submitted in partial fulfilment of the requirements of the Bachelor of Laws Degree, Strathmore University Law School

By

Raqda Sayidali Moalim

145682

Table of Contents

VT OMNES VNVM SINT

Acknowledgement

Finishing this dissertation would not have been possible without the help of the people in my life who care to see me succeed and push me to be my best.

I would like to first thank my supervisor, Cecil Abungu, who made this process as simple and straightforward as it could be. His guidance through this time has been insightful, useful and much-needed. This dissertation would not be even half what it is without his help and support.

I would also like to thank my family, for being as supportive and understanding as they could be. My brother for making me laugh and putting me at ease when things got too challenging.

Last but not least, I would like to shout out my friends and fellow supervisees for the emotional support, the laughs and making the tough times more bearable.

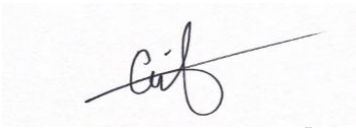
Declaration

I, RAQDA SAYIDALI MOALIM, do hereby declare that this research is my original work and that to the best of my knowledge and belief, it has not been previously, in its entirety or in part, been submitted to any other university for a degree or diploma. Other works cited or referred to are accordingly acknowledged.

Signed: 

Date:

This dissertation has been submitted for examination with my approval as University Supervisor.



Signed:

Abstract

The increasing reliance on artificial intelligence (AI) systems has introduced complex legal challenges, particularly in determining liability for AI-caused harms. The opacity, autonomy, and unpredictability of AI systems complicate the application of traditional tort law principles, particularly the doctrines of causation and foreseeability. The inability to trace an AI system's decision-making process exacerbates information asymmetry between AI developers and injured parties, making it difficult to establish fault under a negligence framework. This study examines whether the doctrine of Res Ipsa Loquitur can be applied to AI harms within the context of US tort law. It explores how AI systems operate and how their characteristics disrupt traditional liability doctrines of causation, both in fact and proximate cause. The study employs a qualitative research method and relies on doctrinal analysis as well as a deductive reasoning approach to reach its conclusions. The study proposes Res Ipsa Loquitur as a viable solution, arguing that in cases where plaintiffs cannot access critical evidence due to information asymmetry between themselves and defendants, then the burden should shift to the opposite side. The study finds that the doctrine is appropriate in situations involving unintentional AI harms that result from either goal misspecification, lack of robustness in the model or unrepresentative training data.

VT OMNES VNVM SINT

List of Abbreviations

AI – Artificial Intelligence

DL – Deep Learning

ML – Machine Learning

API – Application Programming Interface

List of Cases

1. Anderson v Minneapolis (1920) Supreme Court of Minnesota.
2. Byrne v Boadle (1863), United Kingdom Exchequer Court.
3. Holzhauer v. Saks 1997), Court of Appeals of Maryland.
4. Jesionowski v Boston & Maine (1946), The Supreme Court of the United States.
5. Montgomery elevator Co v Gordon (1980), Supreme Court of Colorado.
6. New York Central R.R. Co. v Grimstad (1920) United States Court of Appeals.
7. San Juan Light & Transit Co. v Requena (1912), Supreme Court of The United States.
8. Smith v Claude Neon Lights (1933), New Jersey Court of errors and Appeals.
9. Smith v Koenning (1965), Court of Civil Appeals of Texas.
10. Ulwick v DeChristopher (1991), Supreme Judicial court of Massachusetts
11. Ybarra v Spangard (1944), Supreme Court of California.

VT OMNES VNVM SINT

1.0. Introduction

1.1. Background

AI is considered complex due to its autonomy in executing instructions, achieving goals, and interacting with its environment with minimal human intervention, as well as its black-box nature and unpredictability.[1] This is because the AI systems being developed currently use a method known as deep learning. Deep learning is a type of machine learning that enables AIs to learn and make decisions on their own by processing large amounts of data. It works by using structures called artificial neural networks,[2] consisting of layers of "neurons," which are like small units that process information. When data, like an image or a piece of text, enters the network, each neuron in the first layer takes in some part of that data.[3] The neurons in one layer pass their results to the next layer, where more complex patterns or features are identified. This process continues through multiple layers—hence the term "deep" learning—until the final layer produces an output or prediction, such as recognizing an object in an image or understanding the meaning of a sentence.[4] So, when an AI has more layers, it can often perform better, but it also becomes harder to understand how it reaches its decisions or produces its outputs. This lack of clarity is referred to as the black-box problem.[5]

Unpredictability, in this study, refers to the inability to consistently forecast an AI system's final decisions or anticipate its behaviour once deployed in real-world environments. No one can know how the algorithm will relate different inputs and the kind of patterns it will pick out. Sometimes, even after deployment, an AI system may continue to learn and adjust to its environment, and it is not always clear to the developers exactly what that will look like.[6] In other cases, it can gain 'new' capabilities that it was not exactly trained to have.[7] This makes it impossible to foresee certain risks or accidents that may arise once the AI system has been deployed.

[1] Arcila B, 'AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight?' 54, *Computer Law & Security Review,* 2024, 3.

[2] Janiesch C*,* Zschech P, and Heinrich K, *'*Machine Learning and deep learning' *ArXiv*, 2021, 2 __< https://arxiv.org/pdf/2104.05314 >__ on 17 September 2024.

[3] Janiesch C *et al* 'Machine Learning and deep learning' 3.

[4] Janiesch C *et al* 'Machine Learning and deep learning' 3.

[5] Lior A 'Artificial Intelligence and Tort Law: Who should be held liable when AI causes damages?' Heinrich Böll Stiftung, __< https://il.boell.org/en/2021/12/24/artificial-intelligence-ai-tort-law-and-network-theory-who-should-be-held-liable-when-ai >__ on 19 September 2024.

[6] Llorca D *et al,* 'Liability Regimes in the Age of AI: a Use-Case Driven Analysis of the Burden of Proof' 623.

[7] Llorca D *et al,* 'Liability Regimes in the Age of AI: a Use-Case Driven Analysis of the Burden of Proof' 623.

When an individual suffers harm, they can bring a claim under multiple liability regimes depending on the injury as well as the circumstances around it. Liability can be strict (no-fault) or fault-based.[8] In US tort law, any case seeking to establish fault must present a potential tortfeasor, a breach of duty, and the causal link between the breach and the damage caused.[9] Fault is premised on a person having failed to reasonably protect against foreseeable harms from something within their control.[10] A negligence case arises when the defendant is both responsible for the harm caused and could have reasonably foreseen and prevented it.[11] As negligence law has evolved, its elements have generally been consistent and have been narrowed down to three or four elements: duty, breach, cause, and actual harm.[12]

Causation is an essential element in proving fault. It consists of two major parts; causation-in-fact and proximate cause.[13] The first requires the plaintiff to show that the chain of actual causation extended from the defendant's breach up to the harm suffered by the plaintiff.[14] One must satisfy what is known as the 'but-for' test by proving that but-for the defendant's breach, the plaintiff would not have suffered the harm for which he is seeking compensation.[15] This means that the harm would not have occurred if not for the act of the defendant or the probability of the occurrence of the injury increased because of the defendant's actions. Proximate cause, on the other hand, is established by showing that the relationship between the wrong and harm was sufficiently close rather than remote.[16] Put differently, a reasonable person in the position of the defendant would have foreseen that his actions would result in the kind of harm suffered by the plaintiff.[17] Therefore, to successfully make a prima facie case of negligence, the causation element requires that you establish that the defendant's actions both actually caused the harm and that the harm was reasonably foreseeable.[18]

[8] Lahe J 'Forms of liability in the law of delict: fault-based liability and liability without fault __< https://www.juridicainternational.eu/article_full.php?uri=2005_X_60_forms-of-liability-in-the-law-of-delict-fault-based-liability-and-liability-without-fault >__ on 1 March 2025.

[9] Steel S, 'Legal causation and AI' in Lim E and Morgan P (eds), *The Cambridge Handbook of Private Law and Artificial Intelligence*, Cambridge University Press, 2024,189, see also https://scholarship.law.edu/cgi/viewcontent.cgi?article=2095&context=scholar, 155.

[10] Smart W, Grimm C, and Hartzog W, 'An Education Theory of Fault for Autonomous Systems' 2 *Notre Dame Journal on Emerging Technologies,* 33, 2021, 35.

[11] Páez A 'Negligent Algorithmic Discrimination, 84 *Law and Contemporary Problems,* 19, 2021, 28.

[12] Scordato M 'Three Kinds of Fault: Understanding the purpose and function of causation in Tort law' 77 *University of Miami Law Review* 1, 2022,155.

[13] Scordato M 'Three Kinds of Fault: Understanding the purpose and function of causation in Tort law' 155 – 156.

[14] Ulwick v DeChristopher (1991), Supreme Judicial court of Massachusetts.

[15] Ulwick v DeChristopher (1991), Supreme Judicial court of Massachusetts.

[16] Owen D, 'The five elements of negligence' 1674.

[17] Scordato M 'Three Kinds of Fault: Understanding the purpose and function of causation in Tort law' 157.

[18] Scordato M 'Three Kinds of Fault: Understanding the purpose and function of causation in Tort law' 158.

The doctrine of Res Ipsa Loquitur in US tort law allows a plaintiff to meet their burden of proof by creating a rebuttable presumption of negligence.[19] The general consensus is that the doctrine was first established in the case of *Byrne v Boadle*. In this first instance of its use the court's decision was framed as a way to reach a fair outcome. The court presumed negligence in favour of the plaintiff, affirming that liability can exist where the only evidence is that a vessel fell on the plaintiff.[20] However, this presumption is and has been interpreted as rebuttable, allowing the defendant to prove he had not been negligent.[21] The importance of the doctrine is that where causation cannot be established by the plaintiff and the evidence of the true cause of the accident is accessible only to the defendant, then the burden shifts to him to show that he had not been negligent.[22] Such a doctrine may be deemed useful in AI harm cases where it may be difficult for the plaintiff to successfully prove the negligence of the defendant considering all the factors at play.

When an AI entity causes harm or injury, the 'black box' nature of its decision-making process complicates the assignment of liability. The issues of foreseeability, the autonomy of AI systems, and the limited human control over their potential actions make it challenging to establish a legal nexus of causation between the victim and the wrongdoer.[23] All these things taken together make it extremely difficult or even impossible to establish causation under a fault-based liability system. The causal link between the AI developer and the user who has suffered harm is interrupted by the very nature of how AI systems work. It has been argued that this could make it impossible for AI to fall under our current liability theories because they were created with human behaviour in mind and might not work well when applied to AI.[24]

[19] Tate Jr A 'Wex Malone and Res Ipsa Loquitur in Louisiana Tort Law' 44 *Louisiana Law Review* 5, 1984, 1398.
[20] *Byrne v Boadle* (1863), United Kingdom Exchequer Court; Webb G, 'The law of falling objects: Byrne v Boadle and the birth of Res Ipsa Loquitur', 4, *Stanford Law Review,* 59, 2007, 1080.
[21] Webb G, 'The law of falling objects: Byrne v Boadle and the birth of Res Ipsa Loquitur' 1082.
[22] Prosser W, 'The procedural effects of Res Ipsa Loquitur' *Minnesota Law Review,* 1936, 242 __< https://scholarship.law.umn.edu/cgi/viewcontent.cgi?params=/context/mlr/article/3526/&path_info=uc.pdf >__ on 20 September 2020.
[23] Lior A 'Artificial Intelligence and Tort Law: Who should be held liable when AI causes damages?' Heinrich Böll Stiftung, __< https://il.boell.org/en/2021/12/24/artificial-intelligence-ai-tort-law-and-network-theory-who-should-be-held-liable-when-ai >__ on 19 September 2024.
[24] Lior A and Kline T, 'Addressing AI-related Harms Through Existing Tort Doctrines' Oxford Business Law Blog, 3 July 2024, __< https://blogs.law.ox.ac.uk/oblb/blog-post/2024/07/addressing-ai-related-harms-through-existing-tort-doctrines >__ on 19th September 2024.

1.2. Problem Statement

This study will examine the extent to which the doctrine of Res Ipsa Loquitur can be used to find AI developers liable in scenarios where AI systems cause harm, within the context of United States tort Law.

1.3. Research Questions

1) What are the technical aspects of AI systems and subsequent AI harms that make AI challenge the traditional United States tort law concept of causation?
2) How has the doctrine of Res Ipsa Loquitur been invoked and applied in United States courts and how have the different requirements under the doctrine been interpreted by United States courts and different scholars?
3) To what extent would the doctrine of Res Ipsa Loquitur be an appropriate doctrine to use in addressing AI-caused harms?

1.4. Research Objectives

1) To examine technical aspects of AI systems, their subsequent harms and how they pose a challenge for the traditional US tort law concept of causation.
2) To explore the history and use of the doctrine of Res Ipsa Loquitur and to investigate how the doctrine and its requirements have been invoked and interpreted by the United States courts.
3) To propose that the doctrine of Res Ipsa Loquitur as the most appropriate doctrine in addressing AI-caused harms.

1.5. Hypothesis

My hypothesis is that the doctrine can be used to find model developers liable in instances of AI harms where the harms are accident harms, resulting from goal misspecification, lack of robustness in models and unrepresentative data, which occur at the development stage of AI systems.

1.6. Justification

The excessive difficulty in making claims when injury is suffered due to an AI system allows for AI developers to get off scot-free, banking on the impossibility of a claimant establishing causation. It is therefore important to find ways to bring claims against them by trying to level

the playing field between the victims and developers. This study will be useful to lawyers building legal arguments on behalf of people who have suffered from AI-caused harm. Specifically, those representing them in such cases can benefit from this research as it will show how these claims can be brought and provide a better chance of success. It would also be beneficial, as a starting point, for judges faced with such cases in discerning whether an invocation of Res Ipsa is appropriate in AI harm cases. Finally, it will be important to researchers in the field who are working on ways of holding AI creators accountable and working within a fault-based liability system.

1.7. Conceptual Framework: Information asymmetry

Information asymmetry refers to a situation where one party has more or better information than another.[25] This allows them to have an advantage over their counterpart. In a market situation, for instance, buyers and sellers have different information. A buyer may have information on the different items in the market and the price range for these goods but may lack the same depth of information, especially as to the quality, as the seller. This creates an obvious informational advantage.[26] This concept was first introduced by George Akerlof who argued that the buyer will use some market statistic to measure the value of a good while lacking the kind of intimate information the seller has, and the seller will usually try to take advantage of this.[27]

Although the concept of information asymmetry originated in the field of economics, it is now applied to various relationships where one party possesses more or better information than the other. A notable example is in healthcare, where medical providers usually have greater knowledge about a patient's condition than the patient. The patient also lacks the medical knowledge that healthcare professionals possess and therefore requires that these professionals act in their best interest.[28] Another relationship is between private firms and regulators, where

[25] Bergh D, Ketcher Jr D, and Boyd B, 'Information asymmetry in management research: Past accomplishments and future opportunities' 45, *Journal of Management* 1, 2019, 123.

[26] Auronen L 'Asymmetric Information: Theory and Applications' *In Seminar of Strategy and International Business as Helsinki University of Technology*, 21 May 2003, 4 __< https://edisciplinas.usp.br/pluginfile.php/7594769/mod_resource/content/0/Asymmetric%20Information-%20Theory%20and%20Applications.pdf >__ on 19 September 2024.

[27] Akerlof G 'The Market for "Lemons": Quality uncertainty and the market mechanism' 84 *The Quarterly Journal of Economics,* 3, 1970, 488.

[28] Fabes J, and Avşar T 'Information asymmetry in hospitals: Evidence of the lack of cost awareness in clinicians' 20, *Applied Health Economics and Health Policy,* 2022, 694.

private firms have better knowledge about their own practices and how they would operate without certain regulations, therefore undermining the regulations.[29]

This idea can also be extrapolated to legal cases where the defendant has better information than the person accusing them – cases such as medical malpractice claims, employment termination disputes and civil rights claims.[30] For instance, in medical malpractice cases, there are instances where it is necessary that the defendant himself establishes that the injury suffered was not due to their negligence but due to some other cause.[31] This is because there is some 'special' information known to the medical practitioner that the plaintiff cannot access or explain.

Similarly, in aviation accidents, the carrier has better information about their own negligence or diligence in a case where a passenger has suffered injury because of them. This makes it especially difficult, or even impossible, for the plaintiff to show negligence. They would need to learn the complex terminology in the industry and the complex scientific knowledge must be comprehended before even trying to read the official reports on the accident.[32] Normally, it is the defendant who knows how far they complied with the rules and how cautious they were as compared to the plaintiff.[33] These situations, in most cases, result in the shifting of the burden of proof from the plaintiff to the defendant, who is better placed to explain away their supposed fault.[34] For instance, in a lot of cases involving passenger carriers – like planes or ships – when the plaintiff is unable to meet the burden of proving specific negligence, they can invoke the doctrine of Res Ipsa Loquitur to then shift the burden to the defendant. Other instances are those of medical malpractice,[35] and civil rights claims, such as instances of discrimination.[36]

[29] Fullerton D and Wolfram C, 'Introduction and Summary to "The design and Implementation of US Climate Policy" in Fullerton D and Wolfram C (eds*) The design and Implementation of US Climate Policy,* University of Chicago Press, 2012,11.

[30] Institute of the Advancement of the American Legal System, *Fact-Based pleading: A solution hidden in plain sight,* 2010, 7.

[31] https://www.hoaglandlongo.com/blog/pre-existing-conditions-burden-shifting-in-medical-malpractice-cases

[32] McKee H, 'Aviation law –Problems in litigation arising from aircraft disasters' 37, *Notre Dame Law Review,* 2, 1961, 196-197.

[33] Sanchirico C 'A primary-activity approach to proof burdens' 37, *The Journal of Legal Studies,* 1, 2008, 275.

[34] __< https://www.alrc.gov.au/wp-content/uploads/2019/08/ir_127ch_11._burden_of_proof.pdf >__ on 20 September 2024.

[35] Thornton R, 'The limited use of inferred negligence in medical cases' 15, *Baylor University Medical Center Proceedings,* 2, 2002, 228.

[36] __< https://www.alrc.gov.au/wp-content/uploads/2019/08/ir_127ch_11._burden_of_proof.pdf >__ on 20 September 2024.

In situations of AI-caused harm, there exists information asymmetry between the injured party and the AI developers.[37] The developer has access to the datasets used, the number of training runs, the algorithms, the safety evaluations and measures, and the amount of compute used, with a depth that the users of the system do not have. Even when they publish reports on these things, they withhold certain information,[38] for purposes of protecting their product and IP. [39] This creates a situation where it is especially cumbersome for victims of AI harms to meet the usual standard of proving fault. Even though there are many complex variables in proving causation in such cases, it is made even more challenging by the information deficit suffered by the plaintiff.

This concept will be utilized as the basis for why it is important that AI developers bear the burden of showing that they were not at fault in cases where an AI causes harm and the plaintiff is not in a position to know the facts involved as there were 'unusual circumstances. This will be done by examining the challenges faced by plaintiffs in proving fault due to the clear information asymmetry which works against them. The information deficit on the part of the plaintiff makes it impossible to establish a causal link between the defendant and the harm suffered. Therefore, the information asymmetry concept will form the basis for why an invocation of Res Ipsa Loquitur is appropriate in such cases where the developer, as the party with the better information, should bear the burden of proving their diligence.

1.8. Literature Review

There exists are rich body of literature on issues of liability both regarding who should be held liable and under what liability regime. The proper liability framework for AI companies has become the subject of an ongoing and growing debate in the academic community. Some academics support a negligence approach, while others have advocated for the implementation of strict liability laws or the creation of a "safe harbour" for AI creators through ex-ante legislation. This growing discourse is a reflection of the challenges that new technologies—AI in particular—have presented to existing tort systems.

[37] Vasudevan A, 'Addressing the liability gap in AI accidents' *Center for International Governance Innovation*, 2003, 5 https://www.cigionline.org/static/documents/PB_no.177.pdf on 20 September 2024.
[38] OpenAI 'GPT-4 Technical Report' *arXiv,* 2024, https://arxiv.org/pdf/2303.08774 on 20 September 2020.
[39] Yeung K, 'Why worry about Decision-Making by Machine?' in Karen Yeung, and Martin Lodge (eds), *Algorithmic Regulation*, online ed, OUP, 2019, 25-26.

1.8.1. On the unique challenges created when the use of AI results in harm

AI harms can arise in many ways from unspecified instruction (misalignment), malicious use or some defect in the AI systems,[40] and can take many forms ranging from tangible harms such as physical harm or damage to property, to intangible ones such as discrimination, breach of privacy and other fundamental rights.[41] Normally, one can bring a case under civil liability law, and it is quite straightforward. However, when it comes to AI systems, scholars have identified certain 'complexities' that make it extremely difficult to make a claim when harm is suffered.[42] AI is opaque, as there are multiple stakeholders, and it is increasingly autonomous.[43] This makes it challenging to create a clear causal link between the damage suffered and the fault of the liable person.

According to Beatriz Arcila, opacity arises from both the black-box nature of AI systems as well as organizational structures.[44] The black-box refers to the fact that the decision-making process of an AI system is not very clear. For instance, it is hard to say that it was a specific part of the input data used that led to a specific output making it hard to understand why a decision was reached.[45] This means that the inputs are related to the outputs without showing exactly what the inner workings of the algorithm are like.[46] The decision-making process is concealed either for fear of exposing trade secrets (although even opening access might be fruitless as it would be impossible for the layman to decipher them) or 'gaming'.[47] Because it is not clear how the input resulted in the subsequent output, establishing causality becomes difficult.[48]

On the existence of multiple stake holders, Buiten et al, find that there are numerous actors involved at every stage of the development and deployment of AI systems, and it is hard to say

[40] Hendrycks D, Mazeika M, Woodside T 'An overview of catastrophic risks' *arXiv,* 2023, 2, __< https://arxiv.org/pdf/2306.12001 >__ on 2 March 2025.

[41] Hoffmann M and Frase H 'Adding structure to AI Harm' CSET, 2023,15 __< https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/ >__ on 20 September 2024.

[42] Buiten M, Streel A, and Peitz M, 'The law and economics of AI liability' 48 *Computer Law & Security Review* 2023, 5-6.

[43] Buiten M *et al,* 'The law and economics of AI liability' 5-6.

[44] Arcila B, 'AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight' 3.

[45] Yeung K, 'Why worry about Decision-Making by Machine?' 25-26.

[46] Cheng H *et al* 'Explaining Decision-Making Algorithms through UI: Strategies to help Non-expert Stakeholders' CHI Conference on Human Factors in Computing Machinery, New York City, 2 May 2019, 3.

[47] Yeung K, 'Why worry about Decision-Making by Machine?' 28.

[48] Arcila B, 'AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight' 5.

exactly where and from whom an accident originated.[49] This makes it hard to assign responsibility and raises questions of who to hold liable. For instance, the hardware being used could have been bought separately from the digital content. Training and deployment can also be done by different parties. Anyone who wants to bring a case might have to confront the hardware manufacturers, the software designers, the software developers and any other party involved at any given point.[50] This issue paired with opacity could make it impossible to satisfy the requirements of a fault-based liability system.[51]

Autonomy in AI refers to the capacity for a system to act and attain its purpose. This is possible because AI systems use Machine Learning (ML) algorithms which allow them to have 'decisional autonomy' in processing inputs and determining outputs, independent of human intervention.[52] Autonomy also means that the AI will not behave in a static manner and behaviours can arise in a nonlinear way. This affects foreseeability as a condition for legal causation.[53] This also creates a remoteness of problem where the unexpected behaviour of AI systems could have been learnt autonomously making it impossible to link it back to the defendant.[54]

1.8.2. On the use of existing liability doctrines in AI-caused harms

Many proposals have been made for using and adjusting existing liability systems such as product liability, strict liability and fault-based liability in cases involving harm by AI systems. Proponents of each liability regime have their own reasons for why it is the best fit.

The proponents of strict liability claim that it is the most appropriate regime for AI-caused harms seeing as it prevents lengthy and costly litigation by placing the blame on the party that is best placed to internalize it. It does so by avoiding getting into the problematic analysis of the 'reusable person' standard while also dealing with the issues of autonomy and opacity of AI systems.[55] It is said that taking the negligence route is too complex because it cannot

[49] Arcila B, 'AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight' 5.

[50] Buiten M *et al,* 'The law and economics of AI liability' 6.

[51] Vasudevan A, 'Addressing the liability gap in AI accidents' *Center for International Governance Innovation*, 2003, 5 https://www.cigionline.org/static/documents/PB_no.177.pdf on 20 September 2024.

[52] Yeung K, 'Why worry about Decision-Making by Machine?' 22.

[53] Buiten M *et al,* 'The law and economics of AI liability' *6*.

[54] Thomasen K 'AI and Tort Law' in Martin-Bariteau F and Scassa T (eds*) 'Artificial Intelligence and the law in Canda'* LexisNexis, 2021,13, see also, De Conca S, 'Bridging the liability gap: Why AI challenges existing rules in liability and how to design human-empowering solutions' in Custers B and Fosch-Villaronga E (eds) *Law and Artificial Intelligence: Regulating AI and applying legal AI in legal practice'* TMC Asser Press*,* 2022, 244.

[55] Cote K 'Outsmarting smart devices: Preparing for AI liability risks and regulations' 25, *San Diego International Law Journal* 101, 2024, 119-120.

sufficiently address the black-box nature of AI systems, making strict liability the better alternative.[56] Cote finds that a strict liability regime would offset the inherent dangers of AI systems by incentivizing developers to reduce harmful levels of activity knowing that they would be liable for any damage caused.[57]

Anat Lior argues for two main reasons why strict liability is the best regime to use in this case. The first makes an analogy to the principle-agent relationship and the second references George Fletcher's nonreciprocal harms framework. The first argument categorizes AI entities as agents and the principles – the humans in charge of making these systems – as liable for damages caused by their agents. The accountable party will be reduced to the human who is best able to 'control' and monitor the AI entity's activities.[58] The second argument is that, in principle, one should be able to recover for injuries caused by a risk greater than what he could have inflicted on the defendant. This means the defendant created an excessive risk of harm, disproportionate to the victim's own 'risk-creating activity'. AI systems inherently possess the ability to inflict greater harm on users than what the users can inflict on them therefore, the victims of AI-inflicted harms should be able to recover for the damages.[59]

Another major argument made for strict liability is that developing, and use of AI systems can be seen as an 'abnormally dangerous activity'.[60] The risk can be either in magnitude or due to the circumstances that surround it to impose strict liability regardless of reasonable care taken in carrying it out. The test for abnormally dangerous activity is that (1) the activity creates a foreseeable and highly significant risk of physical harm even when reasonable care is exercised by all actors, and (2) the activity is not one of common usage. Gabriel Weil argues that the training and deployment of 'frontier' AI models meet the test and as such we should adopt a liability regime. He claims that training and deploying AI models are 'clearly' not common usage due to the amount of computational required in this process. In the second part of the test, he claims that recognizing any software development project as abnormally dangerous can be possible if strict liability is applied differently.[61]

---

[56] Lior A 'AI strict liability vis-a-vis AI monopolization' 22, *The Columbia Science & Technology Law Review* 2020, 94.
[57] Cote K 'Outsmarting smart devices: Preparing for AI liability risks and regulations' 120.
[58] Lior A 'AI strict liability vis-a-vis AI monopolization' 95.
[59] Lior A 'AI strict liability vis-a-vis AI monopolization' 95.
[60] Henson R "I am become death, the destroyer of worlds": Applying strict liability to Artificial Intelligence as an abnormally dangerous activity' 96 *Temple Law Review* 3, 2024, 362.
[61] Weil G 'Tort law should be the centrepiece of AI Governance' Lawfare, 6 August 2024, __< https://www.lawfaremedia.org/article/tort-law-should-be-the-centerpiece-of-ai-

Proposals for negligence often find that while requirements such as the duty of the developer and the breach of said duty can be proved, factual and proximate causation may be difficult to prove.[62] For instance, in asking whether a reasonable person would have acted in the way that the defendant did, we can look at whether they had failed to consider relevant risks – according to industry experts – therefore leading to a breach of duty of care.[63] Another way to look at this is by examining whether the defendant complied with industry customs and safety standards. Since there are certain safety standards being used currently, we could hold defendants using these standards to show a breach of duty of care.[64] However, as most proponents of this regime and other scholars generally recognize, making a negligence claim is extremely onerous on the plaintiff.[65] This is why either strict, vicarious or product liability may seem more favourable.[66]

1.8.3. Contribution

In trying to deal with the impossibility of establishing causality in AI-caused harms, many scholars have proposed taking a strict liability approach.[67] Such an approach seems useful as it would save the transaction costs of tedious and lengthy litigation and places the burden of costs on the actor best placed to absorb them.[68] Proposals for a strict liability approach find that the gaps in a fault-based approach are too big for it to be a useful approach in dealing with AI-caused harms. This is because the opaque, unpredictable and autonomous nature of AI systems makes it difficult to establish proof of causation which is necessary in a fault-based liability system.[69]

Another proposal has been to lower the burden of proof borne by the claimant.[70] The EU AI Liability Directive proposes that courts should order defendants and third parties to give evidence that the court deems 'necessary and proportionate' to a claim for damage caused by

governance#:~:text=By%20holding%20AI%20developers%20responsible,cost%2Deffective%20risk%20mitigation%20measures >__ on 6 October 2024.

[62] Ramakrishnan K, Smith G and Downey C 'US Tort liability for large-scale Artificial Intelligence damages' Rand, 2024, 18-20 __< https://www.rand.org/content/dam/rand/pubs/research_reports/RRA3000/RRA3084-1/RAND_RRA3084-1.pdf>__ on 6 October 2024.

[63] Ramakrishnan K, Smith G and Downey C 'US Tort liability for large-scale Artificial Intelligence damages' 16.

[64] Ramakrishnan K, Smith G and Downey C 'US Tort liability for large-scale Artificial Intelligence damages' 17.

[65] Selbst A 'Negligence and AI's Human Users' 100 *Boston University Law Review* 1315-1318.

[66] Cote K 'Outsmarting smart devices: Preparing for AI liability risks and regulations' 120; *see also* Lior A 'AI strict liability vis-a-vis AI monopolization' 94.

[67] Vladeck D, 'Machines without Principals: Liability rules and Artificial intelligence' 89 *Washington Law Review* 1, 2014, 146, see also Duffy S and Hopkins J, 'Sit, Stay, Drive: The Future of Autonomous Car Liability', 16 *SMU Science & Technology Law Review* 3, 2013, 453, 471-3.

[68] Vladeck D, 'Machines without Principals: Liability rules and Artificial intelligence' 146.

[69] Buiten M *et al,* 'The law and economics of AI liability' 6.

[70] Article 2(2) *AI Liability Directive* (European Union).

a high-risk AI system.[71] This is supposed to make it less difficult for a plaintiff to identify the person who is potentially liable for the harm suffered by them.[72] The directive also provides for a 'presumption of causation'. This is a rebuttable presumption that it was the defendant's fault that harm was suffered. This applies where (1) the fault is "reasonably likely" to have "influenced" the AI's output, and (2) the "output gave rise to the damage."[73] The plaintiff is also expected to show that there was non-compliance with a certain EU or national obligation related to the harm suffered therefore leading to the damage.[74] This provision is meant to deal with the complexity of the AI environment such as the black-box problem.

This study seeks to fill the gap left by the proposals made to deal with the causation problem in AI-caused harms. So far, other liability regimes have been proposed to avoid dealing with the difficulties presented in a fault-based approach. However, this study seeks to propose a way to make it possible to bring an AI-harm case within a fault-based liability regime. It will do so by proposing the doctrine of Res Ipsa Loquitur as a way of shifting the burden of proof to the defendant. This allows for a plaintiff to still bring a case under fault-based liability but not be disadvantaged by the heavy burden of proving causality. This allows enough room for the defendant to defend against the claim (unlike strict liability) while relieving the plaintiff, fully, from having to create a clear causal link between the defendant's fault and the damage suffered.

1.9. Methodology

This study will take a qualitative approach in conducting its research. It will primarily rely on secondary data such as journal articles, blogs, reports, books and chapters in books to help it carry out its analysis in all the chapters.

The second chapter will draw extensively from journal articles, reports, working papers, books, and book chapters to provide a comprehensive analysis of the technical foundations of AI systems and the mechanisms through which AI harms emerge. Through a rigorous content analysis of existing literature, this chapter will critically examine the multifaceted interactions between AI systems and their environment, assessing how these interactions contribute to

[71] The directive uses 'High-risk' as defined in Article 6, *Artificial Intelligence Act* (European Union) see also, Article 2(2) AI *Liability Directive* (European Union).
[72] European Parliament '*Artificial Intelligence liability directive'* 2023, 6.
[73] Article 4(1) (b) & (c), *AI Liability Directive* (European Union).
[74] Article 4(1)(a) *AI Liability Directive* (European Union) see also, European Parliament '*Artificial Intelligence liability directive'* 6.

potential harm. The plethora of research on the subject will be useful in clearly mapping out the intricacies of how AI systems work, how they are trained, how they make decisions, and the consequences of these decisions. The chapter will also delve into defining and categorising AI harms and exploring different actors involved in the AI value chain.

Chapter three will explore case law from the United States courts and investigate how the doctrine of Res Ipsa has been interpreted and applied by courts. It will also rely on journal articles to explore the history of the doctrine and how it has evolved over time. This will help to analyse the purpose and function of the doctrine and allow the study to reach conclusions about when it is appropriate to apply the doctrine. The chapter will then examine case law and break down each of the three requirements of the doctrine. It will rely on journal articles, reports and American Tort law textbooks to carefully analyse each requirement and conclude on its application. The chapter will also examine the effects of the doctrine by analysing cases where there was a successful invocation of the doctrine. Generally, the study will rely on secondary sources and employ a deductive approach in reaching its final claim.

1.10. Chapter Breakdown

The study will be divided into five chapters. Chapter one will introduce the study and all its essential parts. It contains, among other things, a background, a justification for the study, the conceptual framework it will rely on as well as the research questions and objectives.

Chapter two gives an understanding of the AI systems being discussed in the study. it will explain the technical aspects of AI by explaining what deep learning is and what deep neural networks are. It will then explain the ways in which AI systems are trained and how they learn to carry out different tasks. The chapter will also characterise AI harms within the scope of the study and highlight the different ways AI harms can occur and what elements need to be present for an occurrence to be an AI harm. Finally, the chapter will explore the different parts of the AI lifecycle, and the different actors involved.

Chapter three will map out, in detail, the doctrine of Res Ipsa Loquitur. It will explore its history and its purpose as well as how it has evolved over time. The chapter will then explore each requirement of the doctrine, how courts and scholars have interpreted these requirements and how they have been applied. This will then aid the application of the doctrine to AI harms in Chapter 4.

Chapter four will propose Res Ipsa as the best-fit doctrine for taking on AI-caused harms. It will apply the doctrine to AI harms as they have been discussed in previous chapters. It will explore each requirement and how AI harms can meet the requirement. It will also argue that the function of the doctrine makes it appropriate for situations of AI harm.

2.0. How AI systems operate and the causes of AI-related harms

2.1. Introduction

The previous chapter mapped out the entire study. In this chapter the study will explore how AI systems operate and how AI-related harms occur. It will first break down how AI systems work by detailing deep learning and artificial neural networks as the building blocks for AI systems. It will then explain the different methods used to train AI systems and how they learn. The chapter will then delve into the complexities of AI harms. It will first define AI harms and what they look like. It will then classify them into either misuse harms or unintended harms and focus on the latter. It will then explain the ways that these harms arise, focusing on three broad causes. The chapter will finally look at the AI lifecycle and the different parties involved in it.

2.2. Understanding how AI systems work

To understand how AI systems work, first one must understand the technical aspects of AI systems. This means understanding the techniques that are used in developing AI systems as well as the ways they are trained to carry out different tasks.

2.2.1. Deep learning and artificial neural networks

Current AI systems rely on a technique called deep learning, a subset of machine learning (ML) that enables an AI to process massive datasets and make decisions on its own.[75] Deep learning (DL) is useful in situations where there is complex data and massive datasets.[76] To fully understand what deep learning is and how it works, there are certain terms worth knowing. First, an algorithm is a process that a computer can follow to analyse a dataset and pick out patterns in the data.[77] Another important term is a function which is a deterministic relationship that maps inputs to one or more outputs, similar to a model which is how different inputs relate,

[75] Janiesch C, Zschech P, and Heinrich K, 'Machine Learning and deep learning' *arXiv*, 2021, 2 __< https://arxiv.org/pdf/2104.05314 >__ on 15 November 2024.
[76] Kelleher J, *Deep Learning*, The MIT Press, Cambridge Massachusetts, 2019, 6.
[77] Kelleher J, *Deep Learning*, 7.

to produce the final output.[78] Finally, a neural network is a computer model, inspired by the human brain, which consists of small information processing units known as neurons.[79] These neurons are arranged in layers with the first layer being the input layer, followed by many hidden layers and an output layer at the end.[80] When data, such as an image or a text, enters the network, each neuron in the initial layer captures specific information from the input. This data is then passed from one layer to the next, with each successive layer identifying increasingly intricate patterns within the input data.[81]

The neural network has to go through a training process for it to 'learn' from the data it is given. First, training starts with a neural network whose parameters (like adjustable settings) are randomly chosen. At this stage, the model doesn't work well and makes inaccurate predictions. The training process involves feeding examples from the dataset into the network. These examples represent the behaviour you would like to see in the model. For each instance, the network's prediction is compared to the correct answer, and the parameters are adjusted to make the prediction more accurate. This process continues until the model can predict the correct outputs reliably for the given problem. When the model reaches this level of accuracy, the training is complete, and the final model is ready for use.[82]

2.2.2. Supervised, unsupervised and reinforcement learning

The different ways models can learn during training are supervised learning, unsupervised or reinforcement learning.[83] In supervised learning, the examples in the dataset are labelled with the expected output. For instance, if a dataset maps annual income and debt to a credit solvency score, the solvency score would be the target feature.[84] In reinforcement learning, the model interacts with its environment by producing actions which affect the state of the environment which then results in the model receiving rewards (or punishments). The goal is for it to learn to act in a way that maximizes the reward it receives (or minimizes on the punishments).[85] The third kind of learning, unsupervised learning, involves the model receiving inputs without target outputs or rewards from its environment. The goal is for it to build representations of the

[78] Kelleher J, *Deep Learning*, 7.
[79] Kelleher J, *Deep Learning*, 10.
[80] Kelleher J, *Deep Learning*, 68.
[81] Janiesch C *et al* 'Machine Learning and deep learning' 3.
[82] Kelleher J, *Deep Learning*, 13-14.
[83] Ghahramani Z 'Unsupervised learning' in *Summer school on Machine learning,* Springer, Berlin, 2003, 73.
[84] Kelleher J, *Deep Learning*, 26.
[85] Ghahramani Z 'Unsupervised learning' 73.

input data that can be useful in decision-making or uncover hidden patterns or data groupings without human intervention.[86] When the training process is done, the model should be able to use what it has learnt in new examples where the exact output value is not known, and it should be able to estimate this value.[87]

All of this happens in the hidden layers of the model or rather in a blackbox. This means it is practically impossible to follow from the initial input to the output given, considering there are billions of neurons and parameters involved.[88] This also makes it impossible to predict how they will perform in specific contexts.[89] It is due to the neural network's complex architecture that there is no transparency in how the model arrives at its decisions.[90] The blackbox problem makes it impossible to account for or predict the outcomes reached by an AI system. The unpredictability of AI is defined as our inability to precisely and consistently predict what actions an intelligent system will take to reach its objectives, even while we know its terminal goals.[91] This could result from what has been termed as 'emergent' behaviour which is the idea that the interaction between any complex system and its environment could lead to new and surprising behaviour.[92] All these things taken together have legal and ethical implications when we use AI systems to make decisions about people, in medical treatment and all other spheres of life.

2.2.3. Why deep learning is important

Deep learning has been deemed the Gold Standard in the field of ML. it has become the most widely used computational approach in the field, achieving outstanding results on complex cognitive tasks, matching or even outperforming humans.**[93]** DL is considered a driver of technology today and has boomed in different fields due to its learning capabilities, outperforming shallow ML models and traditional data analysis approaches.**[94]** Developments in neural network architectures, optimisation techniques and compute, including GPUs and

[86] Ghahramani Z 'Unsupervised learning' 73.
[87] Kelleher J, *Deep Learning*, 14.
[88] __< https://ourworldindata.org/grapher/artificial-intelligence-parameter-count?country=~Cognitron >__ on 15 November 2024.
[89] Janiesch C *et al* 'Machine Learning and deep learning' 8.
[90] Hussain J 'Deep Learning black box problem' unpublished, Uppsala University, Sweden, 2019, 10.
[91] Yampolskiy R 'Unpredictability of AI' *arXiv,* 2019, __< https://arxiv.org/ftp/arxiv/papers/1905/1905.13053.pdf >__ on 15 November 2024.
[92] Trusilo D 'Autonomous AI Systems in Conflict: Emergent Behavior and Its Impact on Predictability and Reliability' 22 *Journal of Military Ethics* 1, 2023, 4-5.
[93] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8#:~:text=Abstract,a%20full%20understanding%20of%20DL.
[94] Janiesch C *et al* 'Machine Learning and deep learning' 3.

TPUs, have significantly advanced the capabilities and efficiency of DL models. These advancements and others, like the learning techniques discussed above, have allowed for DL models to be used in real-time decision-making and problem-solving.[95] DL has become the cornerstone of AI and the advances seen today are due to the deep learning techniques used.

However, the advanced capabilities of DL models have raised lots of concern, especially regarding the range of harms that can emerge from these models. These harms emerge due to many different reasons, the main one being misalignment.[96] DL models are very complex and lack transparency.[97] They can also have dangerous capabilities such as power-seeking,[98] deception[99] and manipulation.[100] They can also be employed for harmful purposes.[101] All of these issues become even more concerning as AI advances and gains more capabilities, thanks to DL

2.3. Defining and categorizing AI harms

2.3.1. Characterizing AI harms

An AI-caused harm can be defined as when an entity experiences harm (or potential harm) that is directly related to the behaviour of an AI system.[102] According to the CSET AI harm framework, four elements must be present for an incident to qualify as an AI harm. First, there must be an entity that is affected. Second, there must be a harm event or harm issue. Third, the harm must be directly attributable to a consequence of behaviour. Finally, the behaviour in question must originate from an AI system.[103] A harm event is when the harm definitely occurred while a harm issue is when the harm did not occur but there is a reasonable probability that it could.[104]

---

[95] Elly B 'Advancements in Deep Learning: enhancing AI Capabilities' ResearchGate, 2024, 1.

[96] Ngo R, Cham L and Mindermann S, 'The alignment problem from a Deep Learning perspective' *arXiv,* 2024, 1-2, __< https://arxiv.org/pdf/2209.00626 >__ on 31 March 2025.

[97] Wróbel A, Janusz M, Zieliński B and Rymarczyk D, 'OMENN: One Matrix to Explain Nueral Networks' *arXiv,* 2024, 1, __< https://arxiv.org/pdf/2412.02399 >__ on 31 March 2025.

[98] Carlsmith J, 'Is power-seeking AI and existential risk?' *arXiv,* 2024, 7, __< https://arxiv.org/pdf/2206.13353 >__ on 31 March 2025.

[99] Gabriel I, Manzini A and Keeling G 'The Ethics of Advanced AI Assistants' *arXiv,* 2024, 82 __< https://arxiv.org/pdf/2404.16244 >__ on 15 November 2024.

[100] Gabriel I, Manzini A and Keeling G 'The Ethics of Advanced AI Assistants' *arXiv,* 2024, 55.

[101] Anderljung M, Hazell J 'Protecting society from AI misuse: when are restrictions on capabilities warranted?'__< https://arxiv.org/pdf/2303.09377 >__ on 31 march 2025.

[102] Hoffmann M and Frase H 'Adding structure to AI Harm' CSET, 2023,16 __< https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/ >__ on 15 November 2024.

[103] Hoffmann M and Frase H 'Adding structure to AI Harm' CSET, 2023,16 __< https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/ >__ on 15 November 2024.

[104] Hoffmann M and Frase H 'Adding structure to AI Harm' CSET, 2023,16 __< https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/ >__ on 15 November 2024.

CSET categorizes harms as tangible or intangible harms where a tangible harm is material, observable, verifiable and definitive. On the other hand, intangible harm cannot be directly observed and while the consequences of the harm can be observed, the harm itself cannot. Some examples of tangible harms include physical harm, damage to infrastructure and financial loss while intangible harms include detrimental content, differential treatment or violation of human rights.[105] The framework does not give an exact definition of what an AI system is since there is no agreement on this, and most definitions remain vague. Finally, the harm event or issue must be directly linked to an AI system. This means that while the AI does not have to be the sole cause, it must be instrumental enough that the harm would not have occurred had the system not been involved or had it behaved differently.[106]

OECD has also defined what harm caused by AI systems may look like. It differentiates events where the development and use of AI results in actual harm, an AI incident, or is potentially harmful, an AI hazard.[107] An AI incident is an event where the development or use of AI leads to, injury to the health of a person(s), disruption to the management of critical infrastructure, violations of human rights or breach of obligations under laws intended at the protection of fundamental rights or harm to property, communities or the environment.[108] An AI hazard is when an event could plausibly lead to any of these things happening and AI incidents or hazards are considered serious when they lead to extreme versions of the harms mentioned, such as death.[109] Generally, AI harms, whether it is an AI incident, that has already happened, or one that could potentially happen, must be rooted in the behaviour of AI systems. There must also exist an actual harm or a plausible harm, whether physical or non-physical. This characterisation tells us that there is a wide scope of what can be considered as an AI harm, as long as it stems from an AI system and, as long as damage or potential damage, is suffered.

[105] Hoffmann M and Frase H 'Adding structure to AI Harm' CSET, 2023,12 __< https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/ >__ on 15 November 2024.
[106] Hoffmann M and Frase H 'Adding structure to AI Harm' CSET, 2023,18 __< https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/ >__ on 15 November 2024.
[107] OECD, *Defining AI incidents and related terms,* 2024, 8.
[108] OECD, *Defining AI incidents and related terms,* 2024, 8.
[109] OECD, *Defining AI incidents and related terms,* 2024, 12.

2.3.2. Classifying AI harms

There are two major routes to AI harms. Either through misuse or through misalignment. While this is not a perfect classification, this is the general approach taken in the field and in the literature.[110]

a. Misuse harms

Misuse harms occur when the harm is intentional. An AI system can be intentionally designed to harm, like an AI-enabled weapon system. Or it can be employed by malicious actors to cause harm, by bypassing safety measures, to carry out illegal activities.[111] For instance, incidents of using advanced text, image and speech generation to create deepfakes, able to scam people, or political manipulation and phishing attacks[112] Large Language Models can used to develop malware or drug discovery models can used to design harmful toxins.[113] Misuse harms can be generated by developers or by end-users. For instance, end-users can employ the AI system to cause harm. For example, if a user creates a deepfake video to spread misinformation or bypass safety measures and use the system to create a harmful drug.[114] All of these things are possible due to different capabilities that different systems have and the more advanced a system, the more harmful uses it has. The major difference between misuse harms and misalignment harms is that misuse harms stem from a clear intention to cause harm, either from a malicious actor or from a system built specifically to cause harm.[115]

b. Misalignment harms

Misalignment refers to when an AI system behaves unexpectedly and produces unforeseen outcomes, leading to harm. When a human designer has in mind a certain objective, but the system ends up producing harmful or unexpected results, this means the system is misaligned.[116] For example, a facial recognition system that struggles to identify individuals

[110] Zwetsloot R and Dafoe A 'Thinking about risks from AI: accidents, misuse and structure' Lawfare, 11 February 2019, __<https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure >__ on 2 March 2025.

[111] Fraser H and Daniels O 'Understanding AI harms: An overview' CSET, 2023, __< https://cset.georgetown.edu/article/understanding-ai-harms-an-overview/ >__

[112] Fraser H and Daniels O 'Understanding AI harms: An overview' CSET, 2023, __< https://cset.georgetown.edu/article/understanding-ai-harms-an-overview/ >__ on 15 November 2024.

[113] Anderljung M and Hazell J 'Protecting Society from AI misuse: When are restrictions on capabilities warranted?' 2.

[114] Anderljung M and Hazell J 'Protecting Society from AI misuse: When are restrictions on capabilities warranted?' 2.

[115] Arnold Z and Toner H 'AI accidents: An emerging threat' CSET, 2021,4 __< https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Accidents-An-Emerging-Threat.pdf >__ on 2 march 2025.

[116] Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, and Mané D 'Concrete problems in AI safety' arXiv, 2016, 2 __< https://arxiv.org/pdf/1606.06565 >__ on 15 November 2024.

with darker skin tones because the data it was trained on was unrepresentative.[117] This can arise from specification issues, issues in robustness or issues with the data.[118] Misalignment harms can also be referred to as accident harms or unintentional harms but just because they have been termed as accident harms does not mean they cannot be avoided. For instance, they can be addressed by having more interpretable systems, where humans have insight into the system's behaviour or, ensuring systems are more robust.[119]

i. Goal misspecification

A specification issue arises when there is no alignment between the ideal specification and the behaviour displayed by the AI system. This is because there is no simple rule that captures perfectly what you want the AI system to do.[120] For instance, social media platforms use algorithms to keep people engaged through what they recommend, however, the kind of content that is highly engaging is usually 'harmful' content like hate speech, conspiracies and other such content.[121] When a system designer specifies the wrong objective function, then maximising on that objective will lead to harmful outcomes even if there was perfect learning by the model and there was infinite data for it to learn from.[122] The wrong objective function can, broadly, be a result of either reward hacking or negative side effects.[123] In other literature, misspecification is further divided into reward misspecification or goal misgeneralisation.[124]

Reward misspecification is when a system is accidentally being rewarded for bad behaviour. This is because it can be difficult to specify a reward that captures exactly the desired behaviour.[125] This is also known as reward hacking. The system can find unintended ways to maximise its rewards by exploiting loopholes in its objective function leading it to reach its goal but is not in line with the designer's intent.[126] For instance, if a cleaning robot is rewarded for 'cleaning' it might deliberately create a mess so that it can clean that mess and gain

[117] Gabriel I, Manzini A and Keeling G 'The Ethics of Advanced AI Assistants' 55.
[118] Arnold Z and Toner H 'AI accidents: An emerging threat' CSET, 2021,10 __< https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Accidents-An-Emerging-Threat.pdf >__
[119] https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure
[120] Arnold Z and Toner H 'AI accidents: An emerging threat' CSET, 2021,10 __< https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Accidents-An-Emerging-Threat.pdf >__ on 2 March 2025.
[121] Arnold Z and Toner H 'AI accidents: An emerging threat' CSET, 2021,11 __< https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Accidents-An-Emerging-Threat.pdf >__ on 2 March 2025.
[122] Amodei *et al*, 'Concrete Problems in AI safety' 2.
[123] Amodei *et al*, 'Concrete Problems in AI safety' 2.
[124] Bales A, D'Alessandro W, Kirk-Giannini C 'Artificial Intelligence: Arguments for catastrophic risk' 19 *Philosophy Compass* 2, 2024, 4.
[125] Bales *et al*, 'Artificial Intelligence: Arguments for catastrophic risk' 4.
[126] Amodei *et al*, 'Concrete Problems in AI safety' 7.

rewards.[127] Goal misgeneralisation, or negative side effects, comes from situations where the designer has not accounted for every possibility in the complex environments that systems are expected to work within.[128] This expresses, implicitly, indifference towards these variables in the environment and could lead to harm.[129] This means that goals developed by the system may lead to desirable outcomes in the training environment but not translate into novel situations.[130] Ultimately, the common issue in specification problems is that the reward function R and the performance function R* differ from each other.[131] Nevertheless, there are techniques for avoiding and lessening these issues and designers should be expected to be careful when it comes to specification.[132]

ii. Lack of robustness in the model

Robustness is the ability of a system, model or entity to maintain stable performance across a range of conditions, variations and challenges, showing resilience and adaptability when faced with uncertainties.[133] In Machine Learning specifically, it means the capacity of the model to maintain stable predictive performance when faced with changes in the input data.[134] For instance, systems are very likely to malfunction when they are used in environments that differ from what they were designed for or when given inputs that are different from what they were trained on.[135] When systems are introduced to environments they are unfamiliar with or were not trained in, they often proceed with the heuristics and intuitions they developed in previous training without recognizing their ignorance of the current environment and proceed confidently. This can lead to harmful or offensive outcomes and the confidence the system has in its response may mean it will not be flagged for human inspection and people could take action based on the response.[136]

[127] Amodei *et al*, 'Concrete Problems in AI safety' 7.
[128] Amodei *et al*, 'Concrete Problems in AI safety' 3 -4.
[129] Amodei *et al*, 'Concrete Problems in AI safety' 4.
[130] Bales *et al*, 'Artificial Intelligence: Arguments for catastrophic risk', 4-5.
[131] Leike J, Martic M, Krakovna V, Ortega P, Everitt T, Lefrancq A, Orseau L and Legg S, 'AI Safety Gridworlds' arXiv, 2017,3 __< https://arxiv.org/pdf/1711.09883 >__ on 2 March 2025.
[132] Amodei *et al*, 'Concrete Problems in AI safety' 7.
[133] Braiek H and Khomh F 'Machine learning robustness: a primer' *arXiv,* 2024, 2 __< https://arxiv.org/pdf/2404.00897v2 >__ on 2 March 2025.
[134] Braiek H and Khomh F 'Machine learning robustness: a primer' 2.
[135] Arnold Z and Toner H 'AI accidents: An emerging threat' CSET, 2021,6-7 __< https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Accidents-An-Emerging-Threat.pdf >__ on 2 March 2025.
[136] Amodei *et al*, 'Concrete Problems in AI safety' 16.

Robustness is important for how well a model will perform in the real world and when they are not robust enough, there can be very serious consequences. For instance, an autonomous vehicle that relies on a non-robust image recognition system could misread certain road signs or fail to recognise obstacles, leading to accidents.[137] A major challenge for robustness is distributional shift, which occurs when an AI system encounters a situation it is unfamiliar with. Another challenge is adversarial inputs, where a model can be manipulated into making incorrect decisions.[138] A robust model should be able to respond well to differences in the real environment and the training environment and it should be able to withstand adversarial input data.[139]

iii. Lack of representative data

Training data is a collection of labelled information that is used in deep learning models. Data is important for a model to carry out any number of tasks. The model must first be 'trained' on vast amounts of good-quality data for it to efficiently recognize patterns and learn from these patterns.[140] Depending on the learning mechanism employed, the data used in training may be labelled or unlabelled. In supervised learning the training is done using a labelled dataset while unsupervised learning is done on unlabelled data and the model picks out the patterns and the target on its own.[141] Essentially, both the quantity and quality of data on which a model is trained are important for the efficiency of the model.

A huge problem facing how models process data is that they will either overfit or underfit the training data. Overfitting occurs when a model, even after performing well during training, does not do well with new data. This is because the model 'fits' the training data too closely even including noise,[142] which leads it to pick out wrong patterns and makes it unable to properly generalize.[143] Underfitting, on the other hand, usually means that the wrong model has been employed for the particular problem. The model may be too simple and unable to properly create a function between the input and the output.[144] It does not 'learn' from the data

[137] Braiek H and Khomh F 'Machine learning robustness: a primer' 3.
[138] Leike *et al,* 'AI Safety Gridworlds' 9.
[139] Leike *et al,* 'AI Safety Gridworlds' 9.
[140] Aldoseri et al. 'Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges' 13, *Applied Sciences*, 12 2023, 4.
[141] Aldoseri et al. 'Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges' 6.
[142] Noise refers to unwanted or irrelevant data points, errors or inconsistencies that can negatively impact the performance and accuracy of a Machine Learning model by masking the meaningful patterns.
[143] Tortosa R 'Dataset for Artificial Intelligence' LAB University of Applied Sciences, 2020, 12.
[144] Tortosa R 'Dataset for Artificial Intelligence' 13.

as it should. Data is a key pillar in Deep learning and without good quality data, even the best algorithms fail. The mere process of collecting, cleaning and ensuring the data being used is suitable for training takes '45% or even 80-90% of the time.[145] Considering how crucial data is in how well a model functions, when data is not representative and an AI system makes decisions from insufficient or poorly curated training data, it may lead to certain harms.[146]

2.3.3. The 'too many hands problem'

AI systems have been said to have a 'too many hands' problem where many actors exist at each stage of the AI lifecycle. This means that it is really difficult to know who is liable when harm is suffered.[147] Discussions on assigning liability often lack nuance and try assigning liability to a certain organisation or looking for fault at a certain point in the lifecycle, either at development, pre-deployment, deployment or post-deployment. This fails to recognise the complex nature of the supply chain and the involvement of multiple actors at each stage who could possibly be held liable or actors whose role plays out in multiple stages.[148] The AI supply chain is complicated by the fact that there is a fragmentation of roles and no one actor has full control over the system, making accountability challenging.[149]

The stages of the AI lifecycle are divided, broadly, into the development stage, pre-deployment, deployment and post-deployment.[150] The development stage involves data collection, cleaning and annotation as well as pre-training and fine-tuning the models. At the pre-deployment stage, before the model is released, it goes through testing, evaluations and safeguards are implemented. In the deployment stage, the model is made available to users and businesses whether through model hubs or cloud-based APIs, and the developers still maintain some control over the deployed models. Finally, at the post-deployment stage, the models are fine-tuned for specific purposes and integrated into various applications. Different actors possess different parts of the information, for instance, the model providers monitor API activity while

[145] Whang et al. 'Data Collection and Quality challenges in Deep Learning: A data-centric AI perspective' *arXiv* 2022, 2 __< https://arxiv.org/pdf/2112.06409.pdf>__ on 31 March 2025.
[146] Amodei *et al*, 'Concrete Problems in AI safety' 3.
[147] Cobbe J, Veale M and Singh J 'Understanding accountability in algorithmic supply chains' *arXiv,* 2023, 1, __< https://arxiv.org/pdf/2304.14749 >__ on 2 March 2025.
[148] Cobbe J, Veale M and Singh J 'Understanding accountability in algorithmic supply chains' 2.
[149] Cobbe J, Veale M and Singh J 'Understanding accountability in algorithmic supply chains' 2-3.
[150] Sikumar M, Chang J and Chmielinski K, 'Risk mitigation strategies for the open foundation model value chain' Partnership on AI, 11 July 2024, __< https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/ >__ on 2 March 2025.

application developers are aware of usage information.[151] Also, there are different actors involved in all these stages and some of them are recurrent in different stages. For instance, model developers are active in the first three stages of the lifecycle.[152] A model developer is the party that develops the model that is used in the AI system which broadly sets out the capabilities and behaviours according to which the larger system operates.[153]

The complexity of the system and all the parts that are in play at once makes it hard to classify any party as the one that exercises control over the system at any one point. This also makes it challenging to attribute any liability to a specific party. This is why a broad understanding of the players, and the system altogether is more important than trying to zero in on one specific actor or one specific part of the value chain.[154] All these people exercise control at different or multiple points in the lifecycle and the negligent act can occur at any point and because of even multiple actors.

In all the different stages of the AI lifecycle, AI developers are the most recurrent actors who maintain substantial control over the system.[155] For instance, a developer can provide their pre-built model (already trained and fine-tuned) and host the system on their own infrastructure. They allow access through Application Programming Interfaces (APIs), through which customers can integrate the system's capabilities into their applications, while still maintaining control over the system. So, the model allows the developer's system to run on their infrastructure, under their control, even after they are deployed across many contexts and use cases.[156] This is partly due to the fact that there are high barriers to entry, which limits the number of organisations able to produce state-of-the-art AI either for their own use or to put on the market.[157] This is not to say that developers always maintain control in all stages of the

[151] Sikumar M, Chang J and Chmielinski K, 'Risk mitigation strategies for the open foundation model value chain' Partnership on AI, 11 July 2024, __< https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/ >__ on 2 March 2025.
[152] Sikumar M, Chang J and Chmielinski K, 'Risk mitigation strategies for the open foundation model value chain' Partnership on AI, 11 July 2024, __< https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/ >__ on 2 March 2025.
[153] Shavit Y, Agarwall S, Brundage M, Adler S and O'Keefe C, 'Practices for governing agentic AI systems' OpenAI, 14 December 2023, __< https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf >__ on 2 March 2024.
[154] Cobbe J, Veale M and Singh J 'Understanding accountability in algorithmic supply chains' 2.
[155] Sikumar M, Chang J and Chmielinski K, 'Risk mitigation strategies for the open foundation model value chain' Partnership on AI, 11 July 2024, __< https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/ >__ on 2 March 2025.
[156] Cobbe J, Veale M and Singh J 'Understanding accountability in algorithmic supply chains' 2.
[157] Cobbe J, Veale M and Singh J 'Understanding accountability in algorithmic supply chains' 2.

value chain. However, even when multiple entities share the role, for instance where one trains a model and the other fine-tunes it, they share the responsibilities of a developer in setting out the capabilities and behaviours of the system.[158] For this study, the focus will be on developers as explained and it will focus on them as the party with 'exclusive control', while also considering other possibilities.

2.4. Conclusion

This chapter sets out to give an understanding of how AI systems operate and how they may result in harms. This was done by exploring the technical aspects of how AI systems work. This involved an exploration of deep learning as the technique employed in most AI systems today. The chapter then delved into defining and characterising AI harms to set the scope for what AI harms mean in this study. It also categorised AI harms and explored different issues such as lack of robustness, goal misspecification and poor data. It finally explored the different actors involved in the AI lifecycle. The next chapter will explain the doctrine of Res Ipsa Loquitur in detail and set the grounds for the arguments in the fourth cpater.

VT OMNES VNVM SINT

[158] Shavit Y, Agarwall S, Brundage M, Adler S and O'Keefe C, 'Practices for governing agentic AI systems' 5.

3.0. Res Ipsa Loquitur in the context of US tort law

3.1. Introduction

The previous chapter detailed how AI systems operate, how they cause harms, how those harms manifest, and the different actors involved. The previous chapter aimed to explain the complexities that give rise to an invocation for Res Ipsa. This chapter will explore Res Ipsa in detail. It will be contextualised within the framework of US negligence law by first discussing negligence generally and its requirements. The chapter will then discuss the details of the doctrine, such as its requirements how courts have interpreted and applied them, and the different effects it has had in different cases.

3.2. Establishing fault under US tort law

To establish a prima facie case of negligence, a plaintiff must prove the existence of a duty, the failure of the defendant to exercise reasonable care, causation, and physical harm.[159] Each of these elements must be independently proven and require separate analyses. For causation, one must prove both proximate and factual cause. Factual causation requires the plaintiff to create a cause-and-effect relationship between the harm suffered and the defendant's breach of duty.[160] For proximate cause, the harm suffered must have been foreseeable to the actor at the time they made the decision.[161]

3.2.1. Cause in fact

Cause in fact requires the plaintiff to requires the plaintiff to establish that the defendant's negligent act directly resulted in the harm suffered. It must be that the loss suffered would not have occurred 'but for' that act of negligence.[162] Take for example a situation where someone walks off the sidewalk into the road and is hit by a vehicle which is being driven negligently fast. In a negligence suit, the plaintiff must show that not only was the driver speeding but that the excess speed, amounting to negligence, is what caused the harm. If it can be said that the pedestrian could still have been hit even if the car was being driven at a reasonable speed, then it was only the conduct of the driver, not the negligent aspect that caused the accident.[163] This

[159] Section 6, *Restatement (Third) of Torts: Liability for physical and emotional harm* (USA).
[160] Owen D 'The five elements of negligence' 35 *Hofstra Law Review* 4, 2007, 1674.
[161] Owen D 'The five elements of negligence' 1683.
[162] Owen D 'The five elements of negligence' 1674.
[163] Owen D 'The five elements of negligence' 1680.

means that there cannot be any counterfactuals available to show that the accident would still have occurred regardless of the actions of the defendant.[164]

Even in situations of multiple causes, courts have adopted tests to create a cause-and-effect relationship between the accident and the conduct of the defendants. In such cases, courts have adopted either the 'independently sufficient' test or the 'substantial factor test'. In the first test, the court analyses if the harm would have occurred the same or in a similar manner without the concurring action.[165] The substantial factor test asks whether the defendant's conduct was a substantial factor in bringing harm to the plaintiff. Prosser and Keeton find that this test is an improved version of the 'but for' test for a special class of cases of independent but concurring events that lead to or contribute to the harm suffered. The purpose of this test is to then exonerate certain defendants whose actions have not contributed substantially to the harm suffered.[166]

3.2.2. Proximate cause

This analysis is done to determine whether it would be fair to hold the defendant liable, even though they caused the harm.[167] The underlying idea is that the responsibility for the consequences of one's actions should be according to the quality of a person's choice. This is measured by the consequences the actor should have contemplated as plausible outcomes at the time they made a decision or rather, what they could have foreseen.[168] It is not necessary for a plaintiff to prove that the exact harm and exact chain of events as well as the exact person harmed were foreseeable, but only the general outline of harms and the general class of persons who are endangered by the defendant's actions.[169] If the harm is deemed too remote, for instance where a third party or event intervenes and distorts the 'natural consequences' of the negligent act, then there is no proximate cause.[170]

Causation, therefore, requires proof that the defendant's conduct was the exact cause of the harm suffered and that the defendant could or should have reasonably foreseen that their

[164] *New York Central R.R. Co. v Grimstad* (1920) United States Court of Appeals.
[165] *Anderson v Minneapolis* (1920) Supreme Court of Minnesota.
[166] Prosser and Keeton '*The law of torts'* 5th ed, West Publishing Co., United States of America, 1984, 267-68.
[167] Gunn A and Johnson V '*Studies in American tort law'* 2nd ed, California Academic Press, Durham, 1999,354.
[168] Owen D 'The five elements of negligence' 1683.
[169] Harper F 'The foreseeability factor in the Law of Torts' 7 *Articles by Maurer faculty* 468, 1932, 470.
[170] Owen D 'The five elements of negligence' 1684.

conduct would result in such harm.[171] While causation need not be proved with certainty but on a balance of probabilities,[172] in many instances, it has proven to be an insurmountable burden for plaintiffs due to the complex nature of certain accidents. In such cases, the plaintiff cannot access the information or evidence needed to prove causation. This has led to the development of certain arguments calling for the party with the 'best information' to bear the burden of proof.[173] This has also led to the development of the doctrine of Res Ipsa Loquitur, to try and remedy the causation problem.

3.3. Understanding the doctrine of Res Ipsa Loquitur

3.3.1. History of the doctrine and why it exists

The doctrine was first introduced into Anglo-American law in 1863 in the case of *Byrne v Boadle*. The plaintiff in the case was injured by barrels that fell from the window of the defendant's store and even though he failed to prove negligence or an intentional tort, he was successful in his suit.[174] The court stated that there are certain cases where the mere occurrence of an accident is evidence of negligence.[175] What followed the case was the attempts to establish criteria for the types of cases in which the doctrine can be applied.[176] By 1905 the doctrine had found its way to the US, where it was so frequently invoked that three considerations were proposed to limit its applicability. The considerations were that the apparatus must be such that it would not ordinarily be injurious unless there was some carelessness, both the inspection and control of the thing must have been in the exclusive control of the charged person and that the injury happened irrespective of any action by the injured party.[177]

The subsequent and most current expressions of the rule have similar elements but are phrased differently. The US Supreme Court applied the doctrine in *San Juan Light & Transit Co. v Requena*, stating that it applied in cases where injury occurred without any fault of the injured

[171] Owen D 'The five elements of negligence' 1683.
[172] Fleming J 'Probabilistic causation in tort law: A postscript' 68 *The Canadian Bar Review* 4, 1989, 136-137.
[173] Justice MacGrath M 'Res Ips loquitur – a rule, a principle, a maxim, a doctrine, a myth or a convenient label' 8 *Irish Judicial Studies Journal 2,* 2024,73.
[174] *Byrne v Boadle* (1863), United Kingdom Exchequer Court.
[175] Byrne v Boadle (1863), United Kingdom Exchequer Court.
[176] Petri J 'Res Ipsa Loquitur – the Wyoming Supreme Court relaxes the elements a plaintiff needs to establish before invoking the doctrine- Goedert v Newcastle Equipment Co.' 27 *Land & Water Law Review* 1, 1992, 209.
[177] Petri J 'Res Ipsa Loquitur – the Wyoming Supreme Court relaxes the elements a plaintiff needs to establish before invoking the doctrine- Goedert v Newcastle Equipment Co.' 209.

person, the thing causing the injury was in the exclusive control of the defendant, and the injury was one that would not ordinarily occur if the person in control used proper care. If all these things are present, it is reasonable to conclude that the injury came from a want of care on the side of the defendant.[178] As the doctrine continued to develop, the three requirements remained but the interpretation of what it means to meet the requirements remained up for debate. The 'exclusive control' requirement being the most contentious.[179] Once the plaintiff establishes the three elements, the burden of explanation falls on the defendant. It is the onus of the defendant to explain that they were not negligent, and if they are not successful, an inference can be drawn of their negligence.[180]

The reason behind the shifting of the burden of proof to the defendant in such cases is the recognition that it would be palpably unfair to require a plaintiff to prove something which is beyond his reach, and that falls uniquely within the defendant's capacity to explain.[181] Courts recognize the comparative ability between litigants to shed light on the facts of the matter, and while discovery and other pre-trial procedures exist, it is established practice that where the facts are uniquely within the knowledge of the defendant, then Res Ipsa applies. In fact, there are instances where it has been applied to advance this policy objective even as it distorted its evidentiary basis.[182] As was the case in *Byrne v Boadle*, the court was willing to open itself up to the commonsense conclusion that the barrels would not have fallen if it were not for some lack of care on the part of the store owner, even if the plaintiff could not point to any specific negligent act.[183]

3.3.2. The requirements of the doctrine

a. An inference of negligence

The accident must be one that would not ordinarily happen without negligence. This requirement embodies the doctrine of Res Ipsa since it requires one to show that the accident itself is one that would give rise to an inference of negligence.[184] The doctrine has been applied

[178] *San Juan Light & Transit Co. v Requena (1912),* Supreme Court of The United States.
[179] Petri J 'Res Ipsa Loquitur – the Wyoming Supreme Court relaxes the elements a plaintiff needs to establish before invoking the doctrine- Goedert v Newcastel Equipment Co.' 210.
[180] Guterson M 'Res Ipsa Loquitur: Application and effects' 27 *Washington Law Review 2,* 1952,148-149.
[181] Justice MacGrath M 'Res Ips loquitur – a rule, a principle, a maxim, a doctrine, a myth or a convenient label' 74.
[182] Fleming J *'The Law of Torts'* 6th ed, The Law Book Company Limited, Sydney,1983, 294.
[183] Justice MacGrath M 'Res Ips loquitur – a rule, a principle, a maxim, a doctrine, a myth or a convenient label' 69.
[184] Clark W 'The applicability of doctrine of Res Ipsa Loquitur to cases involving bursting bottles' 1951 *Washington University Law Review* 2, 1951, 218.

in a wide range of cases which fall within the ambit of this requirement.[185] This requirement has developed in many different ways, and it is not obvious what accidents would be considered 'unusual' enough that the only commonsense conclusion would be that they were occasioned by some form of negligence.[186] It takes common knowledge or the existence of experience to show that an accident is one that would not occur under ordinary circumstances.[187] For instance, in earlier cases of aviation, courts were not persuaded to allow an invocation of the doctrine since there was not enough common knowledge of its hazards.[188]

There are certain accidents where their occurrence does not always point to the existence of negligence, for instance, tumbling downstairs, an ordinary slip and fall, the tyre of a vehicle exploding or a fire of an unknown origin. They do not, on their own, justify the conclusion is the most likely explanation.[189] Even in cases where it is not 'common sense' for the judge and jury to form an opinion that it was negligence that most likely caused the harm, expert testimony can lay the foundation for a permissible inference of negligence.[190] The point of this element is that, while there could be other causes of the injury, the most likely one is negligence. The plaintiff does not have to eliminate all other possible causes or inferences, only on a balance of probabilities in favour of negligence.[191]

Generally, the courts have decided case-by-case what accidents are indicative of negligence, and which ones are not. There is no specific characteristic in the accidents that courts have previously found to imply negligence that concretises them. According to Prosser and Keeton, most of these cases have an element of drama and improbability and are 'unusual'. However, they also note that courts have also found seemingly mundane accidents to fall within the scope of this condition.[192]

b. Exclusive control

In this requirement, the instrumentality causing harm must be in the exclusive control of the defendant.[193] The point of this element is to focus the inference of negligence on the

[185] Prosser and Keeton '*The law of torts'* 246.
[186] Prosser and Keeton '*The law of torts'* 246.
[187] Prosser and Keeton '*The law of torts'* 246.
[188] Prosser and Keeton '*The law of torts'* 246.
[189] Prosser and Keeton '*The law of torts'* 246.
[190] Fleming J *'The Law of Torts'* 291.
[191] Fleming J *'The Law of Torts'* 289.
[192] Prosser and Keeton '*The law of torts'* 246.
[193] Witt J and Tani Karen *'Torts: Cases, Principles, and Institutions'* 5th ed, Center for Computer-Assisted Legal Instruction, Chicago, Illinois, 2016, 238.

defendant.[194] Courts have used this requirement to preclude Res Ipsa when parties other than the defendant can control the instrumentality causing injury. For instance, in *Holzhauer v. Saks*, the court ruled out Res Ipsa from a plaintiff who was injured by a mall escalator that suddenly halted. The court stated that there were hundreds of customers who had unlimited access to the escalator's emergency stop, and it was impossible to establish that the escalator was in the exclusive control of the defendant.[195] However, this literal requirement of control was found to be rigid and unduly restrictive. In *Jesionowski v Boston& Maine RR*, the court stated that the exclusive control requirement should not be construed so narrowly as to restrict the use of the doctrine. Rather, it should serve to eliminate the possibility that the accident was caused by a third party and not the defendant.[196]

This requirement is further complicated by the fact that some injuries occur when the instrumentality is in the control of the plaintiff. For instance, cases involving exploding bottles.[197] This raises the question of whether the defendant should have control at the time of the injury or at the time of the negligent act.[198] It should be sufficient to show that, at the time of the negligent act, the instrumentality was in the control of the defendant. If the defendant had control of all factors that caused the accident, then whether they were in actual possession at the time is the accident is immaterial.[199] Some scholars have proposed that the 'control' requirement be done away with entirely due to the absurdities it has created. Rather, we should require that 'the apparent negligent cause of the accident be such that the defendant would, more likely than not, be responsible for it.[200] In fact, in some cases, it does not make sense to use the word 'control' at all. For instance, when a plaintiff is riding a horse, they have exclusive control, but if the saddle slips off and they are injured, the defendant is still responsible.[201] Similarly, in cases of aeroplane accidents, it is the carrier who is held responsible, yet the pilot had exclusive control of the aircraft.

The discussion around exclusive control also involves questions of whether, in situations where multiple actors had control over the instrumentality, exclusive control can still be established.

[194] Clark W 'The applicability of doctrine of Res Ipsa Loquitur to cases involving bursting bottles' 219.
[195] *Holzhauer v. Saks 1997),* Court of Appeals of Maryland.
[196] *Jesionowski v. Boston & Maine* (1946), The Supreme Court of the United States.
[197] Prosser and Keeton '*The law of torts'* 249.
[198] Clark W 'The applicability of doctrine of Res Ipsa Loquitur to cases involving bursting bottles' 219.
[199] Clark W 'The applicability of doctrine of Res Ipsa Loquitur to cases involving bursting bottles' 219.
[200] Prosser and Keeton '*The law of torts'* 251; see also Fleming J *'The Law of Torts'* 292.
[201] Prosser and Keeton '*The law of torts'* 250.

Early cases refused to apply the doctrine to multiple defendants due to a strict interpretation of exclusive control. Eventually, the concurrent control concept emerged, and certain courts were willing to entertain it.[202] Similarly, within the ambit of exclusive control, is the concept of superior knowledge, which has been argued sometimes as an element on its own or that it forms part of and can be substituted for exclusive control.[203] These concepts that have grown out of exclusive control are also worth exploring to show the different considerations when it comes to exclusive control and how it has grown from an extremely rigid requirements

i. Concurrent control

Considering the aim of establishing exclusive control is to eliminate the existence of possible third parties causing the injury, its application becomes challenging when there are multiple people involved.[204] In fact, the control requirement prevented the application of the doctrine to multiple defendants until eventually, courts started to accept the concept of concurrent control. In *Smith v Claude Neon Lights*, the court held that Res Ipsa could be maintained against both a landowner and a sign company where the sign advertising the company fell from the landowner's property, injuring a passerby.[205] The application of concurrent control extended to automobile collisions and bursting bottles.[206]

The concept was fully developed in California in the case of *Ybarra v Spangard*.[207] The plaintiff suffered a shoulder injury during the course of an appendectomy, resulting from force applied between his shoulder and neck. He was allowed to use Res Ipsa against all the defendants who had any control over his body or the instrumentalities which might have caused the injury. The court stated that it would be unreasonable to expect the plaintiff to identify any one of the defendants, which included the diagnostician, the surgeon, the hospital owner, the anaesthetist and two nurses, as the one who did the negligent act when he was unconscious the entire time.[208] Most importantly, the court restated that the control requirement is one of the right to control rather than actual control.[209] This is a recognition of the fact that a strict

[202] Koren E 'Res Ipsa Loquitur and multiple defendants: Time for betrothal' 26 *Florida Law Review* 2, 1974, 315.
[203] Koren E 'Res Ipsa Loquitur and multiple defendants: Time for betrothal'
[204] Clark W 'The applicability of doctrine of Res Ipsa Loquitur to cases involving bursting, 219.
[205] *Smith v Claude Neon Lights* (1933), New Jersey Court of errors and Appeals.
[206] Koren E 'Res Ipsa Loquitur and multiple defendants: Time for betrothal' 315.
[207] *Ybarra v Spangard* (1944), Supreme Court of California.
[208] *Ybarra v Spangard* (1944), Supreme Court of California.
[209] *Ybarra v Spangard* (1944), Supreme Court of California.

application of control may lead to unfair results and modern circumstances necessitate a reconsideration of the requirement, taking into account the original goal of Res Ipsa.[210]

ii. Superior knowledge

Another important aspect of the *Ybarra* decision is the court's recognition that because the plaintiff was unconscious during the time of injury, then he would not have knowledge of the circumstances of the injury.[211] The aspect of superior knowledge has long been considered as one of the underlying purposes of the doctrine and the 'the particular force and justice of the presumption'.[212] In fact, in some jurisdictions, the plaintiff must not have any knowledge of the cause of the injury and any specific evidence of negligence known to him may preclude him from invoking the doctrine.[213] In these jurisdictions, it is said that 'he who has charge of the thing' that caused the injury either knows the cause or has the best opportunity to ascertain it.[214] The importance of superior knowledge is not agreed upon. For some commentators, it forms part of the requirement for exclusive control which can be translated to mean either the right to control or the right to manage, meaning access to better knowledge.[215]

Similarly, the courts that recognise the need to show superior knowledge have not considered it as an independent element to be determined on its own. For instance, in *Smith v Koenning* the court held that the element of superior knowledge is embodied in the exclusive control requirement and the inference of negligence requirement.[216] According to Prosser, superior knowledge should not be a consideration of its own or a controlling factor, especially when the control requirement has already been met.[217] However, he also recognizes that it has had a persuasive effect in courts and that courts frequently found that res Ipsa cannot be applied unless evidence of the true explanation of the accident is more accessible to the defendant than to the plaintiff.[218] Therefore, even with the diverging opinions on the importance of superior knowledge, it cannot be denied that it forms part of the analysis, especially under the exclusive

[210] Koren E 'Res Ipsa Loquitur and multiple tortfeasors: Time for betrothal' 317.
[211] *Ybarra v Spangard* (1944), Supreme Court of California.
[212] Koren E 'Res Ipsa Loquitur and multiple tortfeasors: Time for betrothal' 318.
[213] Koren E 'Res Ipsa Loquitur and multiple tortfeasors: Time for betrothal' 319.
[214] Levine H 'Res Ipsa Loquitur in Texas: The element of superior knowledge' 1 *St. Mary's Law Journal* 2, 1969, 210.
[215] Gunn A and Johnson V '*Studies in American tort law'*, 340.
[216] *Smith v Koenning* (1965), Court of Civil Appeals of Texas.
[217] Prosser and Keeton '*The law of torts'* 254.
[218] Prosser and Keeton '*The law of torts'* 254.

control requirement.[219] The doctrine loses meaning if a dynamic of knowledge asymmetry does not exist between the defendant and the plaintiff. Why else would the court relieve the burden from the plaintiff and place it on the defendant?

c. The plaintiff's actions

The accident must not have been due to any voluntary action of the plaintiff.[220] This requirement further supports the exclusive control requirement, whose aim is to eliminate other possible third parties. Consequently, the possibility of the plaintiff being responsible must be eliminated.[221] The element should not be misconstrued to mean that the plaintiff must have been completely inactive, only that an inference of the plaintiff's own responsibility is removed.[222] For instance, a plaintiff merely ordinarily riding does not make him responsible for any injury suffered.[223] The requirement was initially interpreted to mean that any contributory negligence on the part of the plaintiff would mean that they cannot invoke the doctrine. However, in *Montgomery Elevator Co v Gordon*, the court used the concept of comparative negligence, weighing the negligence of the plaintiff against that of the defendant. The court stated that the plaintiff is expected to show that the defendant's inferred negligence is more probable than not, a cause (not *the* cause) of the injury.[224]

3.3.3. The effects of the doctrine

The procedural effect of a successful invocation of the doctrine has also been the root of a lot of debate. Generally, courts have given it one of three effects: an inference of negligence which the jury may or may not draw; a rebuttable presumption of negligence; or a presumption of negligence as well as shifting the ultimate burden of proof to the defendant, who is required to prove on a preponderance of evidence that the injury was not due to their negligence.[225] Ultimately, the doctrine affords reasonable evidence of negligence in the absence of an explanation.[226] In most instances, courts regard the doctrine as a form of circumstantial

[219] Koren E 'Res Ipsa Loquitur and multiple tortfeasors: Time for betrothal' 318.
[220] Shulman H, Fleming J Jr, Gray O and Gifford D '*Law of Torts: Cases and Materials'* 4th ed, Foundation Press, New York, 2003, 269.
[221] Prosser and Keeton '*The law of torts'* 254.
[222] Prosser and Keeton '*The law of torts'* 254.
[223] Shulman *et al*, '*Law of Torts: Cases and Materials'* 268.
[224] *Montgomery elevator Co v Gordon* (1980), Supreme Court of Colorado; *see also* Gunn A and Johnson V '*Studies in American tort law'* 342.
[225] Shulman *et al*, '*Law of Torts: Cases and Materials'* 272.
[226] Shulman *et al*, '*Law of Torts: Cases and Materials'* 272.

evidence that strongly suggests the defendant's negligence as the most plausible explanation for the accident. However, the inference drawn from this remains within the discretion of the jury, meaning that while the defendant is not necessarily required to introduce any countervailing evidence, they risk an unfavourable verdict if they fail to do so.[227]

Nevertheless, there are some courts which give the doctrine a greater procedural effect by allowing it to shift the burden of producing evidence to the defendant or by treating it as a presumption that necessitates the defendant's response. [228] This effect has been more common in cases involving injuries to passengers at the hands of carriers who then bore the burden of proof on the issue of negligence. This position survived due to the conscious policy of requiring the defendant to provide evidence or to pay, and in most cases, they were forced to pay.[229] The presumption is rebuttable as soon as the defendant presents any evidence that contravenes the presumption.[230]

It has been argued that shifting the burden of proof is only appropriate in certain instances. Cases involving carriers and certain medical accidents, where some special responsibility of the defendant towards the plaintiff justifies requiring him to exonerate himself by a preponderance of evidence.[231] Considering the purpose of the doctrine is to fill the informational gap between the parties or to induce the defendant to present evidence known to him and not the plaintiff, then the most useful effect of the doctrine is to create a rebuttable presumption or to shift the burden of proof.[232] Such an effect is also useful in cases of multiple defendants, in which case any innocent defendant could present evidence to exonerate him rather than being subject to an adverse inference from the jury based solely on the evidence of the plaintiff that the accident is one that would not normally occur without some negligence.[233]

[227] Prosser and Keeton '*The law of torts'* 258.
[228] Prosser and Keeton '*The law of torts'* 258.
[229] Prosser and Keeton '*The law of torts'* 258.
[230] Shulman *et al*, '*Law of Torts: Cases and Materials'* 272.
[231] Prosser and Keeton '*The law of torts'* 259.
[232] Kahn J 'Res Ipsa Loquitur: Reducing confusion of creating bias?' *Penn State Law eLibrary*, 2020, 257 __< https://elibrary.law.psu.edu/cgi/viewcontent.cgi?article=1396&context=fac_works#page89 >__ on 2 March 2025.
[233] Koren E 'Res Ipsa Loquitur and multiple tortfeasors: Time for betrothal' 319-320.

3.4. Conclusion

This chapter sets out to clearly define the scope of Res Ipsa and how it is applied. It gives a history of the doctrine and why it exists. It also details the requirements that must be met to successfully invoke the doctrine. It explains how courts have interpreted and applied each requirement. Its aim was to ease into the practical application of the doctrine, which is what the next chapter does.

4.0 How Res Ipsa could fill the causation gap in AI harms

4.1. Introduction

The previous chapter detailed the doctrine of Res Ipsa, its conditions and how these conditions have been interpreted under different circumstances and in different US jurisdictions. This chapter will apply the elements of Res Ipsa to cases involving AI harms. It will make arguments for why AI harms meet the requirements according to how courts have applied them. It will apply each requirement and argue how AI harms fit that requirement, relying on the findings in the previous chapter.

4.2. Applying the conditions of Res Ipsa to AI harms

The three conditions of the application of res Ipsa, as already discussed, are that the accident is one that would not normally occur without negligence, that that instrumentality is within the exclusive control of the defendant and that it was not caused by the plaintiff's own negligence.[234] AI harms arguably meet these three requirements and make an invocation of the doctrine appropriate. For the first condition, the study will argue that AI accidents fall in the class of accidents whose occurrence reflects a want of care. In the second requirement, the study will argue that AI model developers have exclusive control. Considering how contentious this requirement is, the study will also consider other possible parties that can be said to hold exclusive control. For the third and final condition, the study will consider different possible plaintiffs who can claim that the accident was not due to their negligence. The study will use an example of an autonomous vehicle and another of a medical diagnostics system to make its arguments.

4.2.1 That the accident is one that would not normally occur without negligence

Under this requirement, the occurrence must be one that ordinarily does not happen without negligence. Put differently, the event must be that in light of ordinary experience it gives rise to an inference that someone must have been negligent.[235] As discussed in the third chapter, the accident on its own gives rise to an inference of negligence. This means that certain types of accidents, like a bursting bottle, an escalator halting suddenly or falling barrels, reflect the

[234] Shulman *et al*, '*Law of Torts: Cases and Materials'* 265.
[235] Prosser and Keeton '*The law of torts'* 244.

presence of negligence, even if the plaintiff does not present actual evidence of this negligence.[236] The ordinary experience of drinking soda from a bottle, taking an escalator or walking under a store, when care has been taken, is that one would not experience harm. Similarly, when using an autonomous vehicle, the ordinary experience should be that it does not collide with other vehicles, run a red light or cause any other harm and that you get to your destination safely.

As discussed in the third chapter, the courts have found a wide range of accidents to meet this requirement. This does not mean that all accidents justify a conclusion that there was negligence. For instance, accidents that frequently occur without anyone's fault like a simple tumble downstairs, a fall when alighting from a stationary bus or a tyre blowing up, do not themselves justify a conclusion of negligence as the most likely explanation.[237] For AI accidents which stem from goal misspecification, lack of robustness or poor data, there is an indication of negligence. If the necessary care was taken at the development stage, where training happens, then the accident could be avoided. Although it can be rebutted by the defendant that they took reasonable care, at face value, we can conclude that negligence is the most likely explanation. The onus shifts to the defendant to show they were not negligent. In fact, there are ways to avoid these issues and have been extensively discussed in the field.[238] So, when an autonomous vehicle, whose aim is to drive better than human drivers and reduce risk on the road,[239] suddenly collides with another vehicle or steers into the wrong side of traffic, these are unusual occurrences that then reflect negligence. The requirement does not give a conclusive indication of negligence, only that it is the most likely explanation. While other explanations may exist, it is enough that the most likely explanation is one of negligence. Not all AI harms justify a conclusion of negligence. For instance, harms that result from misuse of the system do not imply negligence on the part of the developer. However, all misalignment harms should fall within the umbrella of this condition. This is because where AI systems are involved, there is no equivalent of a simple tumble downstairs, a fall when alighting from a stationary bus or a tyre blowing up. The harms discussed in the second chapter cannot be brushed off as occurring without anyone's fault. The aim of this requirement is not for the plaintiff to rule out all other possible explanations, only to show that on a balance of

[236] *Byrne v Boadle* (1863), United Kingdom Exchequer Court.
[237] Prosser and Keeton '*The law of torts'* 246.
[238] Amodei *et al,* 'Concrete Problems in AI safety' 1-2.
[239] Vladeck D 'Machines without principals: liability rules and Artificial Intelligence' 146.

probabilities, negligence is the most probable explanation for the harm suffered. It becomes the burden of the defendant to explain or present counterfactuals and show that it was not negligence that led to the harm suffered. For instance, being able to show that they used reasonably representative data, since the training data is never really perfectly representative of real life. Similarly, goal specification is an extremely difficult task in the development of models and still poses a major challenge in the field.[240] However, there are machine learning techniques to overcome some of these challenges[241] and it is the burden of the defendant to show their diligence in this regard.

This requirement can be complicated when it comes to AI harms because it is still difficult for a plaintiff to claim that the harm suffered was an unintentional harm caused specifically by, for instance, lack of robustness in the model. It is extremely difficult for a plaintiff to know this much about the cause of the harm they have suffered. This is why they seek to invoke the doctrine of res Ipsa loquitur in the first place. In fact, certain courts have barred the use of the doctrine when a plaintiff has been able to show the direct cause of their injury, when they have proof of a specific act of negligence or when they have established a prima facie case of negligence.[242] So, without having to show any specific act of negligence, a harm caused by the behaviour of an AI system should create an inference on negligence, as it would not occur under ordinary circumstances. The defendant should then be afforded the opportunity to rebut the inference.

4.2.2. That the instrumentality was exclusively within the control of the defendant

Courts have interpreted exclusive control in various ways and have taken a less restrictive approach in cases involving accidents caused by defective devices, vehicles or apparatus, recognising that the true cause of the harm is more accessible to the defendant than the plaintiff. They have moved from defining exclusive control as physical control over the instrumentality to being in a position of ultimate responsibility over the instrumentality when the alleged negligence took place. As discussed in the third chapter, exclusive control can take different forms and can include concurrent control or entails having superior knowledge. In determining

[240] Terry M, Kulkarni C, Watterberg M, Dixon L and Morris M, 'Interactive AI alignment: specification, process and evaluation alignment' *arXiv,* 2024, 3 __< https://arxiv.org/pdf/2311.00710 >__ on 2 March 2025.
[241] Amodei *et al*, 'Concrete Problems in AI safety' 4-11.
[242] St. John's Law Review, *Res Ipsa Loquitur: The extent to which the plaintiff may establish negligence'* 1968, 422.

who possesses exclusive control of AI systems, there are two ways to look at it. First, the developer, as the party that maintains control over the model in all parts of the AI lifecycle, can be said to have exclusive control. The second possibility is that control is shared by different parties who maintain control at different points. In cases where multiple entities play different roles, enough to constitute control, then they can be said to have concurrent control and are all liable.

In the first scenario, AI developers have exclusive control over the instrumentality causing harm. Specifically, when the harm is a misalignment harm resulting from some negligence during the development stage of the system. Exclusive control does not always mean physical control at the time of the accident but control at the point of the negligent act leading to harm. Developers exercise control at the time of the negligent act. For example, if an autonomous vehicle causes an accident and harm is suffered, it is not the owner of the vehicle who is responsible. Assuming there was fog, and this caused the vehicle to malfunction because it had never interacted with foggy environments, this is a robustness issue. The vehicle malfunctioned because it did not know how to behave in this unfamiliar environment. This also means that the accident happened as a result of something in the control of the developer. Although the developer was not in the vehicle or directly instructing it, they still have the necessary control to then hold them responsible.

In the complicated scenario involving multiple actors who exercise control, it should be attributed to the party that was ultimately responsible for the AI system at the time of the negligent act. This encompasses, for instance, the developer responsible for training the model, the deployer who fine-tunes it further, or the end-user who assigns it its domain-specific objectives, depending on where the harmful decision originated. When the AI system undergoes significant modifications by a downstream actor, such as a system developer who fine-tunes the model in a way that may directly contribute to harmful outcomes, then the system developer can be said to have exercised exclusive control at the time that the negligent act occurred. Similarly, if an end-user adopts the model via an API and assigns it domain-specific goals that then lead to harm, then they may be considered as the party with exclusive control at the time of the negligent act that caused the harm. Exclusive control over an AI system should be attributed to any individual or entity that, prior to the AI causing harm, was in a position to define its objectives, influence its robustness or train it on potentially unrepresentative data.

In a situation where a system developer allows their foundation model[243] to be built on by another developer who then deploys the system to end-users, control can be said to have moved down the chain. So, the owner of the foundation model cedes control to the deployer, who fine-tunes it for a specific application, for instance, medical diagnostics. Then, the deployer can sell the specific application to a private entity like a hospital, which then uses the system to diagnose its patients. If a patient is misdiagnosed, any one of these actors can be said to have had control over the instrumentality. Either they had concurrent control, or we pinpoint the exact actor whose negligence led to the misdiagnosis. In the first instance, the foundation model provider could have made their model available via an API through which the deployer was able to fine-tune for a specific purpose. The developer could have maintained control over the foundation model even after deployment,[244] but the deployer is the one who fine-tuned it and released it. In this case, both of these actors can be said to have concurrent control over the instrumentality. It would then be up to them as defendants to show that they did not exercise that control negligently.

In the second instance, we can pinpoint the model developer as the person with the necessary control if the misdiagnosis stemmed from an issue with the foundation model. When foundation models are the base of different systems, any issues with the foundation model could impact applications downstream.[245] The deployer could have fine-tuned the model well enough to carry out the specific task of giving diagnoses, but because of the flaw at the foundation model level, the system gave erroneous diagnoses. This means that at the point of the negligent act leading to the harm suffered, the developer had exclusive control over the instrumentality. However, this exercise might prove too difficult for a plaintiff. It can be really difficult to show the exact root of an error, whether it was from the foundation model or the fine-tuned system. It is because plaintiffs lack access to such knowledge that the doctrine of res Ipsa is invoked. When different actors are implicated at the same time, courts have used the concept of concurrent control. This is especially useful in situations where it is impossible for

[243] A foundation model is an AI model designed to produce a wide and general variety of outputs, see Jones E 'What is a foundation model?' Ada Lovelace Institute, 17 July 2023, https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/ on 31 March 2025.

[244] This is the practice by providers see Cobbe J, Veale M and Singh J 'Understanding accountability in algorithmic supply chains' 2.

[245] Jones E 'What is a foundation model?' Ada Lovelace Institute, 17 July 2023, https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/ on 31 March 2025.

a plaintiff to attribute exclusive control to any one party. In the case that pioneered this concept, the court defined exclusive control as the right to control and considering the case involved a surgery procedure, there were multiple individuals deemed to have exercised the right to control.

It is worth noting that, just like in the previous condition, if a defendant is said to have exclusive control over the instrumentality, they can rebut this or show that they did not exercise their control negligently. It may seem extreme that multiple parties can be said to have control, and so the plaintiff can bring a claim against multiple actors in the AI value chain but considering the complex network of actors, it is not extreme but necessary.

4.2.3. It was not due to any voluntary action of the plaintiff

The AI harms explored in this study, stemming from failures in goal specification, lack of robustness or unrepresentative data, do not reflect any negligence on the part of the plaintiff. Under this requirement, it does not have to be that the plaintiff had no involvement or that they were completely inactive, just that they were not the root cause of the alleged negligence. The ease of meeting this requirement lies in who the plaintiff is due to the complex nature of the different actors involved in the AI lifecycle.

The affected person,[246] for instance, borrowing from the medical diagnostics system, a patient who has been misdiagnosed by the system, cannot be said to have been negligent themselves. As discussed in the second chapter, accident harms usually stem from errors at the development stage, such as misspecified goals, lack of robustness and poor training data. These people are not in any way involved in these processes and often lack detailed knowledge about the exact cause of the harm suffered. Application users, like the hospital in the example, can also be said not to have contributed to the harm suffered. For instance, if the hospital sues the deployer because the system has consistently misdiagnosed its patients. There are two ways this can play out. The first is that there was no voluntary action by the hospital that contributed to the harm, and the deployer is liable (assuming the previous conditions have been met). Conversely, a voluntary action by the hospital could be carelessly using the system. For instance, if doctors

[246] Affected persons can be understood as those who use the service or are affected by the outcomes of the system. For example, patients, job applicants or people whose lives can be impacted as a result of model use (like losing their jobs) see Jones E 'What is a foundation model?' Ada Lovelace Institute, 17 July 2023, https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/ on 31 March 2025.

do not double-check the diagnosis they receive. If they use the system to diagnose cancer but the system's appropriate application is to diagnose sickle-cell anaemia, this precludes them from invoking the doctrine.

4.2.4. The function of Res Ipsa justifies its use in cases of AI harms

The doctrine serves to prevent the injustice that could arise if plaintiffs were required to establish the precise cause of harm they have suffered and the defendant's specific role in cases where the facts are inaccessible to them. This principle is designed to fill the evidentiary gap that exists when plaintiffs are unable to determine the exact cause of their injury due to an informational deficit. It has been widely applied in cases of medical malpractice, aviation accidents and discrimination cases, where victims lack the access to evidence needed to prove fault. Given the nature of AI harms, the difficulty in proving causation and the fact that AI developers and deployers possess superior information, res Ipsa should be applied to allow plaintiffs to not only seek redress but to be compensated. The very rationale that underlies the doctrine aligns with the challenges posed by AI harms, which result from complex, opaque and autonomous decision-making processes that plaintiffs cannot fully comprehend or investigate. AI systems introduce precisely the kind of evidentiary obstacle that the doctrine intends to overcome, making its invocation not only appropriate but necessary to ensure fairness in AI-related litigation.

4.3. Conclusion

This chapter set out to apply the requirements of Res Ipsa to AI harms and argue exactly how this should be done. It argued that for each of the requirements, AI harms can fit into the category of cases where an invocation of the doctrine was allowed. It argued that AI harms are harms that would not ordinarily occur without the presence of negligence. It also broadly defined who can be said to have the 'exclusive' control of the instrumentality causing harm. It then made an argument that no voluntary act of the plaintiff can be said to be the cause of the harm suffered. In this requirement, the study considered different types of plaintiffs and how they can be said not to have negligently caused their own injury. Finally, it argued why the functions of the doctrine and the nature of AI harms make the application of the doctrine both appropriate and necessary. The next chapter will conclude the study and give recommendations based on the findings in the study.

5.0. Conclusion and Recommendations

5.1. Conclusion

This study set out to determine the applicability of the doctrine of Res Ipsa Loquitur to situations involving AI harms. Considering the complexity of proving causation in AI harms, this study proposes a way to invoke the doctrine of Res Ipsa to shift the burden of proof from the plaintiff to the defendant, who is most likely better placed to explain the injury suffered. The study began, first, by explaining how AI systems work to clearly bring out the blackbox problem and the challenges it poses to traditional tort law conception of causation. It clearly explains the technical aspects of current AI systems and the techniques employed to develop them and train them. It also details what AI harms look like, how they are categorised and how they occur. This creates an understanding of the complexity of the system.

The study, in chapter 3, details the doctrine of Res Ipsa loquitur. The doctrine is quite complex and has developed over a long time but remains elusive. It explores the different requirements of the doctrine: that the accident is one that would not ordinarily occur without negligence, that the defendant has exclusive control of the instrumentality, and that the injury was not due to the plaintiff's own actions. The requirements have themselves evolved over time, especially the second one, which has always been the hardest to prove. This allowed the study to have a full grasp of the doctrine and what its most appropriate use looks like, as well as what the most useful effect of the doctrine is.

In chapter 4, the study finally makes a case for how and why the doctrine can be applied in AI harm cases. It finds that the doctrine is appropriate in situations involving unintentional AI harms that result from either goal misspecification, lack of robustness in the model or unrepresentative training data. This is because it is such harms that meet the first requirement of the doctrine: that the accident creates an inference of negligence. The chapter then goes on to identify the person with whom exclusive control lies. It finds that a broad definition of control and the person who exercises control is necessary in AI-harm situations due to the many parties involved in all stages of the AI lifecycle. Finally, it explores how different potential plaintiffs may have to prove that it was not their own negligence that caused them to suffer injury, as per the final requirement. All of this taken together, along with the arguments made for why the function of the doctrine is most appropriate in AI harms, the study finds that the

doctrine of res Ipsa is not only applicable to AI harms cases but necessary to reach a fair outcome.

5.2. Recommendations

Traditional requirements of making a successful negligence case have proven to be incompatible with the nature of AI and its subsequent harms. This means that if we continue to operate within that system and AI continues to be developed how it is today, victims of AI harms may face a lot of injustice and may never get compensation for the harm they suffer. Therefore, this study recommends that:

5.2.1. Recommendation to judges

Judges should approach the interpretation and application of the doctrine of Res Ipsa in a manner that is liberal enough to allow the doctrine to operate how it should. The most well-known application of the doctrine in *Byrne v Boadle* shows the necessity of applying the doctrine in a manner that does not constrict it. Other cases that have moved the doctrine forward and helped it mature have also chosen to apply the requirements of the doctrine in a manner that advances cases involving unusual accidents which could not be explained by plaintiffs. In the recent past, the doctrine has mainly lost its teeth and has faded away in terms of relevance. This could be because of the difficulty in successfully invoking it. The doctrine came to exist for a reason and regardless of all its controversies and confusions, the thing that remains clear is that it seeks to allow plaintiffs to have a chance at a successful suit.

VT OMNES VNVM SINT

5.2.2. Recommendations to other researchers

The doctrine of Res Ipsa has not been explored enough as a useful doctrine in the discussion of AI and liability. While the literature on assigning liability and all the different liability regimes that can be used in AI harms is vast, the exploration of this doctrine has been very minimal. This study only provides a broad idea of how it can be useful and what its invocation would need. However, more studies could find sharper and more specific cases in which the doctrine can be used. It would also be worth studying what the ideal effects of a successful invocation of the doctrine might be. There are three main effects of the doctrine: an inference of negligence which the jury may or may not draw; a rebuttable presumption of negligence; or a presumption of negligence as well as shifting the ultimate burden of proof to the defendant, who is required to prove on a preponderance of evidence that the injury was not due to their negligence.

Each of these effects has different implications when it comes to a successful invocation of the doctrine in AI harms. for instance, while a positive inference of negligence may be problematic and could even hinder the ability of AI companies to develop and deploy their models, a presumption of negligence may be less problematic and allow the companies to explain themselves. However, the presumption could possibly be easily rebutted, defeating the entire purpose of the doctrine in the first place. A rebuttal of any sort on any of the evidence means that the court will not draw an inference of negligence. Extensive research on this and an exploration of these ideas in depth can prove to be very useful in the broader discussion of AI and liability.

5.2.3. Recommendation to Policymakers

The study found that there are gaps in assigning AI liability that ought to be filled by policymakers. For instance, there is a lack of accountability for AI harms because of the too many hands problem explained in the second chapter. This is because the different actors are not formally recognized, and they do not bear any individual responsibility. Policymakers can clearly define each actor (this has so far been done by academics or different institutions), the exact role that actor plays, what their obligations are and then what they can be held accountable for. For example, clearly defining a developer as the entity that develops the foundation model and sets out the broad capabilities and behaviours of the AI systems. The developer can be obligated to ensure their models are as transparent and efficient as possible. Then, they can be accountable for any downstream problems resulting from failure to meet their obligations.

Bibliography

Books

1. Prosser and Keeton '*The law of torts'* 5th ed, West Publishing Co., United States of America, 1984,
2. Shulman H, Fleming J Jr, Gray O and Gifford D '*Law of Torts: Cases and Materials'* 4th ed, Foundation Press, New York, 2003
3. Gunn A and Johnson V '*Studies in American tort law'* 2nd ed, California Academic Press, Durham, 1999
4. Kelleher J, *Deep Learning*, The MIT Press, Cambridge Massachusetts, 2019.
5. Witt J and Tani Karen *'Torts: Cases, Principles, and Institutions'* 5th ed, Center for Computer-Assisted Legal Instruction, Chicago, Illinois, 2016.
6. Fleming J *'The Law of Torts'* 6th ed, The Law Book Company Limited, Sydney,1983.

Chapter in Books

1. Steel S, 'Legal causation and AI' in Lim E and Morgan P (eds), *The Cambridge Handbook of Private Law and Artificial Intelligence*, Cambridge University Press, 2024.
2. Fullerton D and Wolfram C, 'Introduction and Summary to "The design and Implementation of US Climate Policy" in Fullerton D and Wolfram C (eds*) The design and Implementation of US Climate Policy,* University of Chicago Press, 2012.
3. Yeung K, 'Why worry about Decision-Making by Machine?' in Karen Yeung, and Martin Lodge (eds), *Algorithmic Regulation*, online ed, OUP, 2019.
4. De Conca S, 'Bridging the liability gap: Why AI challenges existing rules in liability and how to design human-empowering solutions' in Custers B and Fosch-Villaronga E (eds) *Law and Artificial Intelligence: Regulating AI and applying legal AI in legal practice'* TMC Asser Press, 2022.

Journal Articles

1. Arcila B, 'AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight?' 54, *Computer Law & Security Review,* 2024,
2. Smart W, Grimm C, and Hartzog W, 'An Education Theory of Fault for Autonomous Systems' 2 *Notre Dame Journal on Emerging Technologies,* 33.

3. Páez A ‘Negligent Algorithmic Discrimination, 84 *Law and Contemporary Problems,* 19, 2021.
4. Scordato M ‘Three Kinds of Fault: Understanding the purpose and function of causation in Tort law’ 77 *University of Miami Law Review* 1, 2022.
5. Sanchirico C ‘A primary-activity approach to proof burdens’ 37, *The Journal of Legal Studies,* 1, 2008.
6. Tate Jr A ‘Wex Malone and Res Ipsa Loquitur in Louisiana Tort Law’ 44 *Louisiana Law Review* 5, 1984.
7. Owen D ‘The five elements of negligence’ 35 *Hofstra Law Review* 4, 2007.
8. Prosser W, ‘The procedural effects of Res Ipsa Loquitur’ *Minnesota Law Review,* 1936.
9. Bergh D, Ketcher Jr D, and Boyd B, ‘Information asymmetry in management research: Past accomplishments and future opportunities’ 45, *Journal of Management* 1, 2019,
10. Akerlof G ‘The Market for “Lemons”: Quality uncertainty and the market mechanism’ 84 *The Quarterly Journal of Economics,* 3, 1970.
11. Fabes J, and Avşar T ‘Information asymmetry in hospitals: Evidence of the lack of cost awareness in clinicians’ 20, *Applied Health Economics and Health Policy,* 2022.
12. McKee H, ‘Aviation law –Problems in litigation arising from aircraft disasters’ 37, *Notre Dame Law Review,* 2, 1961
13. Cote K ‘Outsmarting smart devices: Preparing for AI liability risks and regulations’ 25, *San Diego International Law Journal* 101, 2024.
14. Lior A ‘AI strict liability vis-a-vis AI monopolization’ 22, *The Columbia Science & Technology Law Review* 2020.
15. Henson R “I am become death, the destroyer of worlds”: Applying strict liability to Artificial Intelligence as an abnormally dangerous activity’ 96 *Temple Law Review* 3, 2024.
16. Vladeck D, ‘Machines without Principals: Liability Rules and Artificial intelligence’ 89 *Washington Law Review* 1, 2014.
17. Duffy S and Hopkins J, ‘Sit, Stay, Drive: The Future of Autonomous Car Liability’, 16 *SMU Science & Technology Law Review* 3, 2013.
18. Trusilo D ‘Autonomous AI Systems in Conflict: Emergent Behavior and Its Impact on Predictability and Reliability’ 22 *Journal of Military Ethics* 1, 2023.
19. Bales A, D’Alessandro W, Kirk-Giannini C ‘Artificial Intelligence: Arguments for catastrophic risk’ 19 *Philosophy Compass* 2, 2024.

20. Aldoseri et al. ‘Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges’ 13, *Applied Sciences*, 12 2023.
21. Harper F ‘The foreseeability factor in the Law of Torts’ 7 *Articles by Maurer faculty* 468, 1932.
22. Fleming J ‘Probabilistic causation in tort law: A postscript’ 68 *The Canadian Bar Review* 4, 1989.
23. Justice MacGrath M ‘Res Ips Loquitur – a rule, a principle, a maxim, a doctrine, a myth or a convenient label’ 8 *Irish Judicial Studies Journal 2*, 2024.
24. Petri J ‘Res Ipsa Loquitur – the Wyoming Supreme Court relaxes the elements a plaintiff needs to establish before invoking the doctrine- Goedert v Newcastle Equipment Co.’ 27 *Land & Water Law Review* 1, 1992.
25. Guterson M ‘Res Ipsa Loquitur: Application and effects’ 27 *Washington Law Review 2*, 1952.
26. Clark W ‘The applicability of doctrine of Res Ipsa Loquitur to cases involving bursting bottles’ 1951 *Washington University Law Review* 2, 1951.
27. Koren E ‘Res Ipsa Loquitur and multiple defendants: Time for betrothal’ 26 *Florida Law Review* 2, 1974.
28. Levine H ‘Res Ipsa Loquitur in Texas: The element of superior knowledge’ 1 *St. Mary’s Law Journal* 2, 1969.
29. St. John’s Law Review, *Res Ipsa Loquitur: The extent to which the plaintiff may establish negligence’* 1968.

Thesis

1. Hussain J ‘Deep Learning black box problem’ unpublished, Uppsala University, Sweden, 2019.
2.

Internet sources

1. Arnold Z and Toner H ‘AI accidents: An emerging threat’ CSET, 2021,4 __< https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Accidents-An-Emerging-Threat.pdf

2. Lior A 'Artificial Intelligence and Tort Law: Who should be held liable when AI causes damages?' Heinrich Böll Stiftung, https://il.boell.org/en/2021/12/24/artificial-intelligence-ai-tort-law-and-network-theory-who-should-be-held-liable-when-ai.
3. Lahe J 'Forms of liability in the law of delict: fault-based liability and liability without fault 2005 https://www.juridicainternational.eu/article_full.php?uri=2005_X_60_forms-of-liability-in-the-law-of-delict-fault-based-liability-and-liability-without-fault
4. Lior A and Kline T, 'Addressing AI-related Harms Through Existing Tort Doctrines' Oxford Business Law Blog, 3 July 2024, https://blogs.law.ox.ac.uk/oblb/blog-post/2024/07/addressing-ai-related-harms-through-existing-tort-doctrines
5. Auronen L 'Asymmetric Information: Theory and Applications' *In Seminar of Strategy and International Business as Helsinki University of Technology*, 21 May 2003, 4 __< https://edisciplinas.usp.br/pluginfile.php/7594769/mod_resource/content/0/Asymmetric%20Information-%20Theory%20and%20Applications.pdf
6. https://www.alrc.gov.au/wp-content/uploads/2019/08/ir_127ch_11._burden_of_proof.pdf
7. Vasudevan A, 'Addressing the liability gap in AI accidents' *Center for International Governance Innovation*, 2003, 5 __https://www.cigionline.org/static/documents/PB_no.177.pdf
8. Hoffmann M and Frase H 'Adding structure to AI Harm' CSET, 2023,15 __< https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/
9. Vasudevan A, 'Addressing the liability gap in AI accidents' *Center for International Governance Innovation*, 2003, 5 __https://www.cigionline.org/static/documents/PB_no.177.pdf
10. Weil G 'Tort law should be the centrepiece of AI Governance' Lawfare, 6 August 2024, __< https://www.lawfaremedia.org/article/tort-law-should-be-the-centerpiece-of-ai-governance#:~:text=By%20holding%20AI%20developers%20responsible,cost%2Defective%20risk%20mitigation%20measures
11. Ramakrishnan K, Smith G and Downey C 'US Tort liability for large-scale Artificial Intelligence damages' Rand, 2024, __< https://www.rand.org/content/dam/rand/pubs/research_reports/RRA3000/RRA3084-1/RAND_RRA3084-1.pdf
12. Janiesch C, Zschech P, and Heinrich K, 'Machine Learning and deep learning' *ArXiv*, 2021, https://arxiv.org/pdf/2104.05314

13. Llorca D *et al,* 'Liability Regimes in the Age of AI: a Use-Case Driven Analysis of the Burden of Proof', *arXiv,*2023 *https://arxiv.org/abs/2211.01817*
14. Hendrycks D, Mazeika M, Woodside T 'An overview of catastrophic risks' *arXiv,* 2023, https://arxiv.org/pdf/2306.12001
15. janiesch C, Zschech P, and Heinrich K, 'Machine Learning and deep learning' *arXiv,* 2021, 2 __< https://arxiv.org/pdf/2104.05314
16. https://ourworldindata.org/grapher/artificial-intelligence-parameter-count?country=~Cognitron
17. Yampolskiy R 'Unpredictability of AI' *arXiv,* 2019, __< https://arxiv.org/ftp/arxiv/papers/1905/1905.13053.pdf
18. [1] Ngo R, Cham L and Mindermann S, 'The alignment problem from a Deep Learning perspective' *arXiv,* 2024, https://arxiv.org/pdf/2209.00626
19. Wróbel A, Janusz M, Zieliński B and Rymarczyk D, 'OMENN: One Matrix to Explain Nueral Networks' *arXiv,* 2024, 1, __< https://arxiv.org/pdf/2412.02399
20. Gabriel I, Manzini A and Keeling G 'The Ethics of Advanced AI Assistants' *arXiv,* 2024, __< https://arxiv.org/pdf/2404.16244
21. Carlsmith J, 'Is power-seeking AI and existential risk?' *arXiv,* 2024, __< https://arxiv.org/pdf/2206.13353
22. Hoffmann M and Frase H 'Adding structure to AI Harm' CSET, 2023,16 __< https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/
23. Zwetsloot R and Dafoe A 'Thinking about risks from AI: accidents, misuse and structure' Lawfare, 11 February 2019, __<https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure
24. [1] Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, and Mané D 'Concrete problems in AI safety' arXiv, 2016, __< https://arxiv.org/pdf/1606.06565
25. Leike J, Martic M, Krakovna V, Ortega P, Everitt T, Lefrancq A, Orseau L and Legg S, 'AI Safety Gridworlds' arXiv, 2017, __< https://arxiv.org/pdf/1711.09883
26. Braiek H and Khomh F 'Machine learning robustness: a primer' *arXiv,* 2024, 2 __< https://arxiv.org/pdf/2404.00897v2
27. Whang et al. 'Data Collection and Quality challenges in Deep Learning: A data-centric AI perspective' *arXiv* 2022, 2 https://arxiv.org/pdf/2112.06409.pdf>__
28. Cobbe J, Veale M and Singh J 'Understanding accountability in algorithmic supply chains' *arXiv,* 2023, https://arxiv.org/pdf/2304.14749

29. Sikumar M, Chang J and Chmielinski K, 'Risk mitigation strategies for the open foundation model value chain' Partnership on AI, 11 July 2024, __< https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/
30. Shavit Y, Agarwall S, Brundage M, Adler S and O'Keefe C, 'Practices for governing agentic AI systems' OpenAI, 14 December 2023, __< https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf
31. Kahn J 'Res Ipsa Loquitur: Reducing confusion of creating bias?' *Penn State Law eLibrary*, 2020, __< https://elibrary.law.psu.edu/cgi/viewcontent.cgi?article=1396&context=fac_works#page89
32. Terry M, Kulkarni C, Watterberg M, Dixon L and Morris M, 'Interactive AI alignment: specification, process and evaluation alignment' *arXiv*, 2024, 3 __< https://arxiv.org/pdf/2311.00710
33. OpenAI 'GPT-4 Technical Report' *arXiv*, 2024, __https://arxiv.org/pdf/2303.08774

Other sources

1. Restatement (Third) of Torts: Liability for physical and emotional harm (USA).
2. OECD, *Defining AI incidents and related terms*, 2024.
3. Elly B 'Advancements in Deep Learning: enhancing AI Capabilities' ResearchGate, 2024.
4. Ghahramani Z 'Unsupervised learning' in *Summer school on Machine learning*, Springer, Berlin, 2003.
5. Cheng H *et al* 'Explaining Decision-Making Algorithms through UI: Strategies to help Non-expert Stakeholders' CHI Conference on Human Factors in Computing Machinery, New York City, 2 May 2019.
6. Institute of the Advancement of the American Legal System, *Fact-Based pleading: A solution hidden in plain sight*, 2010.