Artificial Intelligence and Law (2023) 31:133–167
https://doi.org/10.1007/s10506-022-09308-9

ORIGINAL RESEARCH

Black is the new orange: how to determine AI liability

Paulo Henrique Padovan[1] · **Clarice Marinho Martins**[2,3] · **Chris Reed**[4]

Accepted: 28 December 2021 / Published online: 15 January 2022

Abstract

Autonomous artificial intelligence (AI) systems can lead to unpredictable behavior causing loss or damage to individuals. Intricate questions must be resolved to establish how courts determine liability. Until recently, understanding the inner workings of "black boxes" has been exceedingly difficult; however, the use of Explainable Artificial Intelligence (XAI) would help simplify the complex problems that can occur with autonomous AI systems. In this context, this article seeks to provide technical explanations that can be given by XAI, and to show how suitable explanations for liability can be reached in court. It provides an analysis of whether existing liability frameworks, in both civil and common law tort systems, with the support of XAI, can address legal concerns related to AI. Lastly, it claims their further development and adoption should allow AI liability cases to be decided under current legal and regulatory rules until new liability regimes for AI are enacted.

Keywords Explainable artificial intelligence (XAI) · Explainability · Liability

✉ Clarice Marinho Martins
clarice.marinho@unicap.br

Paulo Henrique Padovan
php@cin.ufpe.br; phpadovan@gmail.com

Chris Reed
chris.reed@qmul.ac.uk

[1] Centro de Informática, Universidade Federal de Pernambuco UFPE, Av. Jornalista Anibal Fernandes, s/n, Cidade Universitária (Campus Recife), Recife, PE, Brasil 50740-560

[2] Communication Department, Catholic University of Pernambuco (UNICAP), Rua do Príncipe, 526, Boa Vista, Recife, PE, Brasil 50050-900

[3] Visiting Scholar 2019/2020 at the Centre for Commercial Law Studies, Queen Mary University of London, London, UK

[4] Centre for Commercial Law Studies, Queen Mary University of London, 67-69 Lincoln's Inn Fields, London WC2A 3JB, England, UK

1 Introduction

In the past, artificial intelligence (AI) appeared across the science fiction genre in books and films; now, real AI systems are present across all walks of life. AI technology has embedded itself into public and private spheres and online media, as well as games, domestic robots, apps, and self-driving vehicles, among many other products and services. Broadly speaking, some people understand AI as a technological way to reconstruct human intelligence through machines, thus raising no unusual liability issues. Others still believe it is dangerous and that caution is vital. Regardless of the framework approach that we follow, AI fosters much enthusiasm at the prospect of a "perfect" or longer life, as well as fear of a world dominated by uncontrollable human-like figures, sparking the end of humanity.

Despite many AI-related myths in circulation, we have arrived at a time when scientists are able to replicate some human activities, namely, neuronal behavior, leading to "thinking" machines that perform human-like decisions. Indeed, AI is an umbrella term for different kinds of applications and is defined by the 2018 Report Artificial Intelligence for Europe as.

> Systems that display intelligent behaviour by analysing their environment and taking actions—with some degree of autonomy—to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g., voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g., advanced robots, autonomous cars, drones or Internet of Things applications).

AI technology may bring benefits for personal, social, and economic growth; nevertheless, its accelerating progress also brings significant risk. Regarding the relationship between AI and society, questions have been raised regarding a reasonable application of legal liability regimes in AI-related cases. In the 2018 House of Lords UK report entitled "AI in the UK: ready, willing and able?," some scholars state that liability regulations need more than simple amendments, because traditional rules are no longer adequate. On the other hand, some researchers argue traditional doctrines of tort law will prove satisfactory in addressing the new challenges.[1]

However, even if in the future, new regulation will deal with AI liability, current legal regimes must at least face the issue of proof to claim responsibility under the existing law. That is, all liability systems require proof of evidence that some elements exist in order to allocate responsibility to a person who may have caused damage to a victim. Establishing proof requires answers to detailed questions that need to be resolved to establish how courts can determine AI liability and who will be accountable for it.

AI technologies often produce unpredictable behaviors that may cause harm. How such actions can be understood from a legal perspective is a complex issue. A short time ago, fully understanding the results of some autonomous decisions made

[1] See this discussion in the report mentioned above, House of Lords (2018), and in *Robot Law* Calo et al. (2016, pp. introduction xiv/ xv, 98).

by AI systems was hardly possible. These types of systems are called "black boxes" because although their inputs and outputs can be seen, they can rarely explain how and why the systems work, in contrast to "white boxes" models that are transparent. Current technical developments can explain AI predictions to a certain extent, by opening their black boxes and elucidating at least how they deliberate. This new set of techniques, called Explainable Artificial Intelligence (XAI), is defined by Fisher (2019) as the.

> Ability of machines to explain their decisions to humans, or more generally to provide insights into the machine decision-making process: what input data features were most important for reaching the decision, what other options were considered and why they were rejected, and so on. This is a contrast to the standard (so-called "black box") approach where the AI is used as-is, without knowing why and how it does what it does.

Despite some decisions seeming rather counterintuitive and incomprehensible at first, they can help explain how AI works and mitigate some of the risks they pose to society. Certainly, explainability is key to legal and regulatory liability matters, and AI that cannot offer a sufficient explanation to satisfy these demands related to liability are likely to face problems, ranging from difficulty in securing insurance to failing to achieve social acceptance.

We claim the development of XAI has entered the stage of answering questions that may help in deciding liability-related issues that emerge from intricate AI decisions. We believe that with some adaptations of the current accident-investigation model already employed by courts and the adoption of XAI techniques as forensic tools, we can bridge the gap between legal demands and technical explanations.

In this sense, this article aims to tackle the role of liability in an age of increasingly complex AI technologies, providing an answer to the following question: How does one determine liability in cases of loss or damage as the result of an AI decision? Combining legal and technical understandings, we explore the role of XAI in deciding liability for AI systems.

This article is structured as follows: in Sect. 1, we address the issue of AI and its interaction with law. We also introduce the definition of XAI and how explainability can be key to legal and regulatory liability matters. In Sect. 2, we briefly present the liability framework in civil and common law, and after the examination of three case studies (illustrations), we conclude our question almost always has no obvious answer; therefore, an accident analysis is needed. In Sect. 3, we briefly present the main method employed in accident analysis by science and engineering, root cause analysis (RCA), and the digital forensic (DF) processes employed in the examination of digital evidence. Section 4 focus on XAI techniques. In Sect. 5, we present how to employ XAI techniques as a forensic toolset and how these results, allied with others obtained during the accident analysis, fulfill the legal requirements to determine liability. In Sect. 6, we give our conclusions to the article.

2 Legal requirements

For centuries, legal systems have been forced to face the perils of new technologies that challenge existing notions. Some of the complications derived from these inventions were technological; hence, many legal professionals have found them difficult to grasp.

Regardless, scholars and technologists have always overcome diverging realities to find solutions to deal with potential threats. Therefore, understanding the elements and requirements of liability frameworks in civil and common law tort systems is crucial and can be elucidated using three questions: *What*, *how*, and *why* did the fact happen?

2.1 The liability framework in civil and common law

Tort theories of liability differ in jurisdictions based on civil and common law. We present some aspects of liability in a civil law jurisdiction based on the Brazilian civil law system because it generally adopts the Roman-Germanic rules, and its Consumer Protection Code is very well designed. We also analyze the system adopted in England that in many ways is similar to other common law countries, in order to contrast it with the Brazilian system.

Liability is based on loss or damage that some person, activity, or property has caused: "The function of the law is to allocate responsibility for that causal element to some person, and then to assess whether liability arises based on the nature of that responsibility" (Reed et al. 2016, p.4). The traditional liability system is generally divided into two main areas: negligence and strict liability. Negligence (subjective liability) is based on fault. Strict liability (objective liability) does not require fault and is usually imposed for cases involving high levels of danger that could harm society. The elements of negligence and strict liability in both common and civil law are presented in Table 1.

In discussing the threats arising from AI autonomous systems that deal with defective products, we must consider consumer law. This is mainly based on the strict liability system in English and Brazilian legislation. In the Brazilian consumer law regime, strict liability can be applied for defective products and services.

However, England has no general category of strict liability, but instead a set of diverse categories of activity (e.g., keeping dangerous wild animals) where strict liability is imposed. In each category, all that needs to be proved is that the activity or product falls into the category and that the resulting loss is not remote. In the Brazilian system, the most usual examples of strict liability are based on the Consumer Protection Code (CDC) and the Civil Code (CC).

Note the causation element for common law is the "causal relationship between the defendant's conduct and result. Causation only applies where a result has been achieved and therefore is immaterial about inchoate offenses" (Goudkamp and Peel 2014).

To establish causation, two conditions must be satisfied: a *factual* and a *legal* cause. Initially, the *factual* cause is established when the defendant's acts contribute to the claimant's loss or damage. Usually, the method of establishing *factual* causation is the but-for test. That is, "If the result would not have happened but for a certain event then that event is a cause; contrariwise, if it would have happened anyway, the event is not a cause" (Goudkamp and Peel 2014, p.7–007). Hence, the test will verify if the damage was as a consequence of the defendant's breach of duty or not.

When multiple acts that may have caused harm to a victim take place simultaneously, the but-for test may fail to establish factual causation. In such circumstances, the court must resort to special tests, such as the Necessary Element of a Sufficient Set (NESS)[2] (Wright 1985) or Material Contribution to Injury. This latter test is adopted when the claimant's harm is "produced by a combination of the defendant's conduct and some innocent cause, factual causation will be established if the defendant's conduct made a *material contribution to the injury*" (Goudkamp and Peel 2014, p. 7–009).

Note that when we are dealing with the results of AI systems, multiple acts are usually involved. So, an investigation must be done to reconstruct the story, that is, connect all of them, and to establish the contributing factors (factual causation).

The *legal* cause is established when concluding that a defendant's act or omission, their breach of duty, was the cause of a claimant's loss or damage is possible. Regarding legal causation, we are reminded it has two parts:

> Firstly, it requires consideration of whether the damage was within the foresight of a reasonable person in the position of the defendant at the time of the breach (the remoteness test). This requires that the *type* of damage sustained by the claimant was reasonably foreseeable. The extent of the *actual* injury does not have to be foreseeable for the remoteness test to be satisfied. Defendants are liable for the full extent of the harm they cause, even where that harm is more extreme than might ordinarily have been the case due to an existing vulnerability of the claimant. Secondly, *legal causation* requires that there be an unbroken chain of causation between the defendant's negligence and the claimant's injury. (Reed et al. 2016, p.9)

When an AI is involved, validating underlaying factors, that is, determining the extent of the causes, to establish an unbroken chain of causation is necessary.

Foreseeability involves the consequences of a person's actions or inaction that cause damage to a victim, and this damage was cautiously predictable by a person who acted in certain way. Note the idea of foreseeability is vital to the English liability regime: no foreseeable risk, no liability. That said, with the evolution of XAI, many acts and facts performed by AI systems that before could be considered inexplicable and unforeseeable are now explicable and foreseeable.

[2] For Wright (1985), this test means "something is a cause if it is a 'necessary element of a set of conditions jointly sufficient for the result."

Table 1 Elements of negligence and strict liability in the UK and Brazil

	Tort	
	Negligence	Strict liability
UK (common law)	• Duty of care • Breach of that duty[a] • Causation[b] • Loss or damage	• Risk or dangerous activities or defect[d] • Loss or damage that is not remote and unforeseeable
Brazil (civil law)	• Fact or Human act • Loss or Damage • Causation[c] • Fault	• Risk or dangerous activities or defect[e] • Loss or damage • Causation

[a]Carelessness is premised on how far a risk ought to have been foreseen and guarded against

[b]The breach of duty caused the loss

[c]Establishment of a causal link between a harmful action of someone or a fact, and the damage suffered by the injured party

[d]Given characteristic in a product

[e]Given characteristic in a product or service

The remoteness test, essential to this concept, requires confirmation of whether damage suffered by the victim was cautiously foreseeable by the defendant when the breach of duty took place. The test for remoteness of damage is related to the legal cause and is applied when determining the kind of injury that occurred from a breach of duty.

In Brazilian tort law "causation" means the establishment of a causal link between the harmful action of someone or a fact or event and the damage suffered by the injured party. The civil law system in Brazil differs regarding the loss or damage being necessarily foreseeable because it does not adopt the concept of foreseeability.

For courts to determine liability, having all the liability elements fulfilled in each case is crucial. To fulfil such liability elements and establish a *suitable explanation*, we must analyze *what* happened, *how*, and *why*. In other words, to establish liability, the law demands two explanations: a *factual causation* (*what*) and a *legal causation*

- *What happened?* i.e., factual chain of events that demonstrate causation.
- *How did it happen?* i.e., unbroken chain of causation between the defendant's negligence and the claimant's injury (*legal causation*).
- *Why did it happen?* i.e., the explanations needed (*legal causation*)
    - Helps identify the breach of duty.
    - Explains the events and the chain of events—the order.
    - Helps determine the responsible party or determine the percentile of responsibility.

Negligence liability requires a demonstration of factual and legal causation. So, it demands an answer to the above questions through a post-incident analysis, that

is, accident analysis, namely, an "after the event" (ex-post) explanation to fulfil the requirements of this kind of liability.

Even in cases of product liability, when courts do not need an ex-post explanation, XAI can uncover what caused the loss. This knowledge might enable manufacturers or operators who took due care, in certain circumstances, to avoid being responsible.

2.2 How courts may determine liability (*illustrations*)

Now, we analyze a problematic case related to the situations in which a court is asked to determine liability and requests an after-the-event explanation about how the damage occurred, Illustration 1. Subsequently, we consider a case in which the court requires no after-the-event explanation to rule upon liability, Illustration 2. In this case, traditional tort theories of liability in jurisdictions based on civil and common law differ in some situations. For this reason, we conduct a comparative study between the main features of the Brazilian and the English liability regime. Also, we conduct a study in relation to the requirements of the foreseeability of risk in cases of liability based on fault (negligence). Lastly, we discuss the issue of how the law ought to react to AI, which cannot provide a suitable explanation, Illustration 3, is discussed.

The negligence system is what requires a post-incident analysis to establish the requirements of this kind of liability. First, to fulfill the liability elements, and establish a *suitable explanation*, we must analyze *what* happened, *how*, and *why*. Therefore, we need to establish *what* took place in a certain situation (i.e., find a factual chain of events that can show causation) and *why* (how/when it happened and who got it wrong).

> *Illustration 1*. Consider a vehicle with autonomous-driving software *A* travelling along a two-lane road. A human driver *H* is driving in the opposite direction. *H* spots *A* and provocatively decides to see how it will react to unusual driving behavior. *H* deviates into the lane toward *A*, planning to swerve out again quickly. However, *A* reasons it is best to avoid a collision and swerves into the opposite lane. As a result, it hits an innocent third party, *I*.

UK Common Law first requires information regarding the "duty of care" that any driver must follow. Neither drivers *A* nor *H* should have abruptly changed lanes on the road.

The second requirement concerns the "breach of that duty"; the autonomous car *A* swerved to another lane to avoid a collision with *H*. So, the AI "driver" of *A* did not breach its duty if no (less risky) alternative was possible. However, when *H*, the human driver, deliberately changed lanes, *A* was forced to swerve from their lane, causing a collision. When *A* hit *I*, the AI could not have prevented the accident, due to *H*'s initial action.

The third requirement, "causation," means *I* was injured because of *H*'s behavior. *H* could foresee that in changing lanes, they would cause an accident. When *A* changed lanes, they were simply trying to avoid an accident.

The fourth requirement, "loss or damage," is the physical injury to *I*.

Brazilian civil law first requires the "human act: —the drivers' behavior— as an explanation. The second requirement, "damage," pertains to the physical injury to *I*. The third requirement explores injuries to *I* because of *H*'s behavior—the abrupt changing of lanes. The fourth requirement is "fault." In this case, it was *H*'s fault because *A* swerved to the other lane to avoid a collision with *H*, who had deliberately changed lanes.

> *Illustration 2*. Consider a door lock that uses embedded face-recognition AI to open itself. The door lock identifies a burglar as the owner and allows them entry, and then the burglar steals property and sets the house on fire.

Now, we investigate a case in which no after-the-event explanation is needed to decide liability. So, we need to recognize which kinds of liability claims about an AI would not require such an explanation.

A strict liability system adopted for products in the English law does not require an ex-post analysis. Usually, it is enough to assert liability to know the product did not work as it *should* and the reasons why it failed or whether that failure could have been prevented do not matter. We study a defective-product case.

In the English strict liability system related to products, the first law requirement pertains to a "defect of a product." The defect here is evident: the AI misrecognized the homeowner. The element of "damage" also occurred since this situation resulted in theft and damage to property (due to fire).

Based on the elements of liability that Brazilian and English law require, we may sustain that they differ on this point. In the Brazilian system, the above case characterizes a consumer relationship and strict liability must be adopted. The first requirement, "defect of the product," from Article 12 in the Consumer Code is evident because the AI fundamentally overlooked the homeowner. The second element, the "damage," occurred due to theft and fire on the property. The third element, "causation," is clear as the defect in the facial recognizer allowed the burglar access to the home.

Both Brazilian consumer law and English common law adopt the strict liability regime for product defects. However, the Brazilian strict liability regime adopted for products differs because it demands that the "causation" element demonstrates the *causal nexus* between the defect of the AI product and the damage suffered by the consumer.[3] Therefore, the law requires an ex-post explanation.

In Brazil, XAI technical explanations are essential to fulfill the three elements of strict liability: risk or dangerous activities or defect of a product or a service; loss or damage; and causation. On the other hand, the English system has no requirement for an ex-post explanation for a product defect, because usually it is enough to know the product did not work as it should have. The reasons it failed or whether that failure could have been prevented do not matter. Notwithstanding, if these explanations

[3] The only exception to strict liability that does not demand a 'causation' element in Brazilian law is related to integral risk theory.

are possible, they may alter how the law deals with the strict liability question in the English system.

In strict liability regimes, since manufacturers can be held responsible regardless of fault, a "before the accident" (ex-ante) explanation may be useful as means of proof, for example, to demonstrate *non vitium* of an AI. Such explanations would help a manufacturer or producer demonstrate the victim or third party's fault, thus being exempted of liability. Similar reasoning can be applied in a foreseeability case; that is, apply XAI to demonstrate that the foreseeable damages known at the time were tested and avoided.

> *Illustration 3*. Consider the Uber case in Arizona in 2018. An automated test vehicle hit a pedestrian walking across a lane. The car was operated by a proprietary developmental automated driving system (ADS). A female operator occupied the driver's seat. The vehicle had been operating for 19 minutes in autonomous mode when it approached the collision site. At that time, the pedestrian began walking across an avenue that had no crosswalk, pushing a bicycle by her side. The ADS detected the pedestrian. Although it continued to track the pedestrian until the crash, it never accurately classified her as a pedestrian or predicted her path. By the time the ADS determined a collision was imminent, it was too late (NTSB 2018).

Here, we analyze the requirements of foreseeability of risk in cases of liability based on fault, such as negligence, and the role of ex-ante and ex-post explanations.

The example above highlights the intersection between the legal notion of foreseeability and the training of an AI system to account for all foreseeable outcomes. The producer would argue only subsequent developments in scientific knowledge would have enabled them to predict the defect. AI data can be used to determine a link, such as in the Uber case, but to grasp the reason behind it, or even to establish possible willful misconduct, XAI techniques (i.e., a before the event explanation) are needed.

In this case, the notion of foreseeability, essential to the English system, involves the consequences of an AI's actions or inaction that cause damage to a victim, where damage was cautiously predictable by the producer. The remoteness test, crucial to this concept, requires confirmation of whether the kind of damage the victim suffered was foreseeable by the defendant when the breach of duty took place.

The need for an ex-ante explanation derives from the normative question: "Should the AI have foreseen that risk?", vital to establishing the predictability of an act. From a technological perspective, a before-the-event explanation can predict some level of foreseeability.

However, depending on the answer obtained, we are compelled to contrast it with an ex-post explanation. It is imperative to understand the case and how the issue happened, that is, *how* and *why*, and not merely statistical information about the AI's performance. In the above case, the technical information was accurately given after the performance data were downloaded from the vehicle.

The ex-post explanation is essential when we are drawing an analogy between the conduct of an AI and a person, and not simply a correlation between statistics. When formulating the normative question, "Should the human have behaved in that

way?", we aim to juxtapose the two approaches and establish a basis for a comparative remoteness test.

Proving the method used by AI to reach a decision is particularly complex, and so early cases involving AI decisions might decide a comparison with human conduct will suffice. Those responsible for the AI would only be liable if a human, in the same circumstances with the same "knowledge" as the AI, would also have been liable. In trying to defend a decision reached by an algorithm, the defendant will likely seek to prove the outcome was within reasonably acceptable parameters. This will require consideration of the design of the algorithm itself, the data that the algorithm has been trained on, and the testing of outcomes. A careful auditing process will be critical in establishing the credibility and reliability of an AI system. What is acceptable will evolve over time, and AIs will start to be held to higher standards than humans.

To mitigate the issues, logs should provide accuracy of AI systems' outputs and performance measures, known security risks exacerbated by AI (robustness techniques), explainability of AI decisions to data subjects (explanation techniques), and human biases and discrimination in AI systems (fairness techniques).

As we probe in the analysis of the three illustrations above, utilizing an investigative methodology, for example, accident analysis, that allows us to resolve any doubts, especially the *whys*, is necessary, that is, a method that examines all underlying factors in a chain of events that ends in an accident, determines the cause (or causes) of an accident, and establishes factual and a legal causation. In Sect. 3, we examine the main method employed in accident analysis by science and engineering, RCA, and how linking it to digital forensics fulfills the law requirements.

3 Accident analysis

To decide who will be liable in cases of loss or damage as the result of an AI decision and to comply with the legal requirements outlined in Sect. 2, we must conduct an accident investigation. In fact, the methodology we discuss below is the same used by the NTSB in the investigation of the accident in Illustration 3 and it does not differ from that used in investigations in general. The reason is that an established investigative methodology that has been adapted by specific sectors, for example, aviation, nuclear plant, rail transport and medicine, to meet certain particularities of each one, without any loss.

The investigation is necessary to examine all underlying factors in a chain of events that ends in an accident. Even the most seemingly straightforward incidents, for example, Illustration 2, rarely relies in a single cause. For example, in Illustration 3, an "investigation" that concludes the ADS failed to detect the pedestrian, and goes no further, fails to find answers to several important questions: Why did the car not identify the pedestrian? Was the street lighting at fault? If so, would a lamppost solve the problem? Would a car equipped with light detection and ranging (LIDAR) have the same problem? Was it hacked? Or could the car AI not identify a pedestrian next to a bicycle?

Table 2 Four steps of accident analysis

	Step	Description
1	Fact gathering	Employs a set of forensic processes for preserving, collecting, and documenting evidence to gather all possibly relevant facts that may contribute to understanding the accident
2	Fact analysis	After the forensic process has been completed or partially delivered, the facts are assembled to illustrate the sequence of steps that led to the accident, checking for consistency and plausibility
3	Conclusion drawing	If the accident history is sufficiently informative, conclusions are drawn, supported by the evidence about which factors are linked to the circumstances and consequences of the incident. So, the causes of the incident that need to be corrected are identified
4	Countermeasures	Corrective actions or recommendations are made to prevent recurrence of a similar incident

A crucial part of the accident investigation process is the accident analysis. An accident analysis is carried out to determine the cause (or causes) of an accident. In Table 2, we present a resumed version consisting of four main steps: fact gathering, fact analysis, conclusion drawing, and countermeasures.

Accident-analysis models have two main categories: sequential accident models and systemic accident models. Sequential accident models describe an accident as a chain of discrete events that occur in a particular temporal order. It underlies traditional techniques, for example, failure modes and effects analysis (FMEA), fault tree analysis (FTA), event tree analysis, and cause-consequence analysis. It is usually employed in physical components failures and human errors investigations involving relatively simple systems.

Systemic accident models describe an accident as an occurrence that arises from interactions among system components, that is, a complex and interconnected network of events. The three main models are AcciMap, functional resonance accident model (FRAM), and Systems-Theoretic Accident Model and Processes (STAMP), usually employed in complex systems cases because it includes the principles, models, and laws necessary to understand complex interrelationships and interdependencies between components (technical, human, organizational, and management) of a complex system.

Although we choose a specific model, namely, RCA, to analyze in the next section, there is no loss of generality; that is, sequential or systemic accident models can be used in the analysis of AI systems without hindering the adoption of XAI techniques.

3.1 Root cause analysis

In science and engineering, RCA is the main method employed in accident analysis. It is a method of problem-solving used for identifying the root causes of faults or problems (Wilson 1993).

The different fields that apply RCA do not always use the same denomination, number of phases, or the same set of techniques per phases, but all have the same goal: set the root cause. Also, this method can be used ex ante, preventing problems from occurring, and ex post, to react and alleviate the effects of the problems. In Table 3 we present a version of the RCA consisting of seven steps:

We now focus our attention on step 2, Investigate the Factors, because XAI main contributions sits here. Although this step is the main source of information, its outcomes permeate the entire process of analysis (steps 3 to 6).

Information assembled about actions and conditions, grounded by evidence, is facts, and other information, for example, a condition that should exist but doesn't, is counterfactuals. Our primary focus of attention is on what happened, that is, the facts. Afterwards, we turn to why certain conditions and actions have not been met, that is, counterfactuals, because these have an indirect influence on the outcome of the problem.

Table 3 Seven steps of root cause analysis

	Step	Description
1	Scope the problem	Establish what happened, when it happened, where it happened, and who was involved, i.e., a clear definition of the problem you are investigating
2	Investigate the factors	Decide what information to collect and whom to interview. Collect physical evidence, review process and procedures, photographs, or video the scene, etc
3	Reconstruct the story	Recreate the incident showing a logical sequencing or flow. Develop a detailed timeline to clearly show what happened when
4	Establish contributing factors	Identify conditions, situations, or actions that triggered, allowed, or influenced the incident, i.e., establish causal factors
5	Validate underlaying Factors	Find root causes for each of the incident's contributing factors. The extent of the causes (physical, human, AI, etc.) must be determined
6	Plan corrective actions	Develop one or more corrective actions to eliminate or control each cause and the extent of cause
7	Report learnings	Provide a formal, permanent, auditable, defensible report of your findings

At the end of this step, the compiled information helps us reconstruct the incident. Based on the gathered facts, we establish the chain of actions, along with the factors that affected the performance of hardware, AI, and human.

For minor, simple, or direct problems, the findings from step 1 may be all that is necessary to establish a cause and recommend some action to address it, for example, Illustration 2, although another party, for example, the manufacturer, would benefit from further investigation. For more significant incidents and adverse conditions, a deeper cause analysis (systematically tracking all possible scenarios that may have produced the problem) must address the physical, human, and AI root(s) of the problem, for example, Illustrations 1 and 3.

When investigating an equipment failure, the first line of inquiry should be aimed at determining any form, fit, or function concerns that need to be fixed.[4] Next, the four possible degradation mechanisms should be identified: force, reactive environment, time, and temperature.[5] Then, the primary human–machine interfaces are evaluated, that is, humans who influenced or allowed the physical phenomena to exist, in order to pursue the human root(s) of the incident.[6] At this point, we may have an AI replacing this human (human out of the loop), for example, Illustration 1, or between the machine and the human (human in the loop), for example, Illustration 3.

Although an AI is not a person, we can analyze what it did and how, from three different angles: the task it was performing, the AI's potential to succeed at the task, and processing of job information. Because AIs are prone to make errors, engineers devise barriers or defenses to ensure safety. Thus, our next line of inspection should pursue engineered barriers (e.g., cybersecurity and robustness AI techniques), administrative defenses, oversight defenses, and cultural defenses.[7]

AI's roots of an incident usually have deeper latent roots in the business system, for example, Illustration 2. All that data collection and further analysis will not only show what the AI did that led to the incident but also the circumstances in which it deliberated. For that discovery process, we need as much evidence as possible, especially about the AI, that allows us to both corroborate the facts and investigate AI's actions. We present below a way to accomplish this task.

3.2 Scene recovery unit

To thoroughly investigate the factors associated with an incident, we use some investigation techniques such as: evidence preservation, witness recollection statements and interviewing witness. The AI counterpart for the first two techniques will be addressed by the scene recovery unit (SRU), and the last one by the XAI techniques.

Evidence preservation aims to successfully preserve, collect, and document evidence that may contribute to understanding the accident. The effectiveness of an

[4] See Bloch (2005).
[5] Ibid.
[6] See Bloch (2011).
[7] See Muschara (2007).

investigation depends on immediate preservation of the scene, the physical, human, cyber, and documental evidence related to the incident, as well as its security and custody, to prevent tampering or loss and establish accuracy and validity.

Witness recollection statements are the testimony of a witness involved in an accident about what happened before, during, and after the event. They usually contain information about what the individual saw but may contain other pertinent information not directly linked to what happened. The statement is signed by the witness, assuring its authenticity. This record is used to determine the time, place, and sequence of events also, crucial to determining the root cause.

Logging is the process of recording actions and states, that is, the feedback you get that tells you what's going on, to a repository ("log"). Enabling the automatic recording of an AI event is the closest to witness-recollection statements we have. Also, logs collect, document, and preserve evidence that may contribute to understanding the accident.

One may ask if simply having a log is sufficient to establish an explanation. Such information is not enough per se if we are seeking a full and detailed explanation, e.g., the root cause, but is an essential prerequisite for establishing whether a technological risk has materialized and can serve as plinth for our XAI techniques. For example, the various loggers present in the car in Illustration 3, here we include video cameras and other equipment's, were decisive in resolving several issues involving the behavior of the "AI driver."

Equipping an AI system with the means of recording operational information, that is, logging by design, achieves a certain degree of explainability. The main idea behind logging is the ability to compare a chronological set of inputs and outputs to provide an interpretation of what happened, that is, the chain of events that lead to an act.

Airplane black boxes (actually bright orange to aid recovery) are installed in aircrafts to facilitate aviation incident investigations. By picturing the whole accident—in detail, and pointing toward evidence at each action, consequence, and motive—justifying for each stakeholder what, how, and why things happened is easy.

To avoid dubiety with the term "black box," we call these devices (ranging from mere data transmitters to complex, sophisticated telemetric systems) scene recovery units (SRU). The SRU is responsible for collecting, storing, and communicating telematic information (inputs, outputs, states) essential in recreating an AI system's operating conditions at a given time. Its primary operation is like an airplane's flight data recorder (FDR), which preserves recent flight history, recording dozens of parameters, collected several times per second. Other information regarding the functioning of the system, not linked to AI, should be considered when describing the context (nature, magnitude, location, and timing), for example, audio and video footage, GPS coordinates, or weather condition.

SRU implementation should be conducted in such a way that no interested party can manipulate the data, while all stakeholders retain access as needed. Also, it must consider any adverse implications for the rights of others and be conducted in accordance with otherwise applicable laws, for example, General Data Protection Regulation (privacy), the European Aviation Safety Agency (domain specific), and trade secrets.

In addition, technical issues such as cost of storage and transmission, technical feasibility, and alternative means of gathering information should be considered when demanding, specifying, and designing the SRU. We believe that regulatory agencies (specific for the appropriate applications, e.g., medical devices, machinery regulation, civil aviation, motor vehicle, etc.), because they have the proper knowledge and trained technical staff, should lead the standardization and other assignments of the SRU for specific applications.

Recalling Illustration 1, a vehicle with autonomous-driving software that swerves into the opposite lane hitting an innocent third party, several forensic techniques would allow us to test numerous hypotheses. For example, by analyzing the brake marks left by the car tires and the conditions of the lane in place, we can determine whether the oncoming driver swerved because of a hole in the lane. These results would be incorporated into the list of evidence in the case and would corroborate, or not, the hypothesis.

However, most modern vehicles come equipped with a series of mechanisms and systems that are controlled and adjusted by an electronic element called an on-board computer (carputer). All telemetric information contained in the car—we can assume that at least a car piloted by AI has an electronic control unit that sends the correct commands to the actuators (suspension, transmission, etc.)—already provide enough evidence to establish how the accident happened. As such, if obtaining these data is possible, we can analyze them through digital forensics techniques, our next subject.

3.3 Digital forensic

As previously noted, interviewing a witness is the human equivalent of using some XAI techniques to "interview" AI. By interview, we mean a structured conversation in which one asks questions and the other provides answers.

Interviewing is a technique that helps us attain information, ideas, experiences, and understanding by talking with others. Seeking what others know helps us expand our comprehension of what happened, how, and (sometimes) why. The main objective of the interview, here, is to establish the context in which the accident took place (e.g., goals, focus, sequence of actions, knowledge, and situation awareness), focusing on facts and seeking to understand *why* and not just *what*. Along with the analysis of the other evidence obtained, interviews constitute the building blocks of our next accident analysis step, namely, reconstructing the story.

Ideally, the current level of human-AI communication would allow us to interview AI as Del Spooner interrogates Sonny (a USR's NS-5 robot in Asimov's *I, Robot*) or analyze them as Rick Deckard applies the Voigt-Kampff test in Rachael (a replicant in Philip K. Dick's *Do Androids Dream of Electric Sheep?*). However, we are still far from this level of interaction. For now, we can use our set of digital evidence, most from SRU, and some XAI techniques as forensic tools to analyze AI systems and fulfill our legal requirements.

In general, analyzing all data that were acquired, not only at the SRU, and evaluating them provides the digital evidence the investigation needs. As with

any investigation, we must identify data that verifies existing theories (inculpatory evidence), contradicts others (exculpatory evidence), or shows signs of tampering to hide data (Carrier 2002).

The branch of forensic science responsible for that analysis is digital forensics (DF), defined as.

> The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence derived from digital sources for the purpose of facilitation or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations. (Palmer 2001)

The different fields that apply digital forensics do not always use the same denomination, number of phases, or the same set of techniques per phases. Also, DF can be used ex ante, anticipating unauthorized actions, and ex post, to facilitate or furthering the reconstruction of events. In Table 4 we present a version of the digital evidence processing consisting of nine steps.

The reader should note the similarity between some steps of the process described above and others described in this work. These steps are an instance (for crimes that involve digital evidence) of current practices that collect physical evidence. Another valid question concerns the non-inclusion of the forensic analysis of evidence (physical or digital) in the RCA process or the accident analysis as a step.

The digital forensic analyst (DFA) works within the justice system, providing key evidence to criminal investigations (e.g., forensic reports). Police officers used to do the work of a DFA; however, with the massive adoption and evolution of technology, specialized professionals in the area have become necessary. Also, these professionals are often assisted by developers, manufacturers, engineers, and scientists in specific cases.

Forensic reports are intended to serve as a document that outlines the evidence, procedures, and analysis employed by the DFA to support their conclusions. These reports then serve as input for other investigative processes to reconstruct the story, establish or refute contributing factors, validate, or invalidate underlying factors, and help plan corrective actions.

Forensic analysis can be requested by judges or presented by the parties (claimant and defendant); therefore, it is an independent process. For this reason, we stated earlier that the accident-analysis method employed does not affect the use of XAI but does affect how the case is investigated.

Also, as mentioned previously, evidence preservation aims to successfully preserve, collect, and document evidence that may contribute to understanding the accident. However, even if we cannot fully analyze the evidence, that is, extract all the information and explanations we would like, it should be properly stored for future analysis and reference. In addition, the case one is analyzing today may turn out to be the first in a series that demonstrates, for example, the malfunction of a certain AI that equips a certain autonomous car model or the SRU data used to test hypotheses in another case.

Table 4 Nine steps of digital evidence processing

	Step	Description
1	Identification	Recognizing an incident from indicators and determining its type. This is not explicitly within the field of forensics but is significant because it impacts other steps
2	Preparation	Preparing tools, techniques, search warrants, and monitoring authorizations and management support
3	Approach strategy	Dynamically formulating an approach based on the potential impact on bystanders and the specific technology in question. The goal of the strategy should be to maximize the collection of untainted evidence while minimizing the impact on the victim
4	Preservation	Isolate, secure, and preserve the state of physical and digital evidence, including preventing people from using the digital device or allowing other electromagnetic devices to be used within an affected radius
5	Collection	Record the physical scene and duplicate digital evidence using standardized and accepted procedures
6	Examination	In-depth systematic search of evidence relating to the suspected crime. This focuses on identifying and locating potential evidence, possibly within unconventional locations. Construct detailed documentation for analysis
7	Analysis	Determine significance, reconstruct fragments of data, and draw conclusions based on evidence found. Supporting a crime theory may take several iterations of examination and analysis. The distinction of analysis is that it may not require high technical skills to perform, and thus, more people can work on this case
8	Presentation	Summarize and provide explanation of conclusions. This should be written in a layperson's terms using abstracted terminology. All abstracted terminology should reference the specific details
9	Returning evidence	Ensuring physical and digital property is returned to the proper owner, as well as determining how and what criminal evidence must be removed. Again, this is not an explicit forensics step; however, any model that seizes evidence rarely addresses this aspect

See Reith et al. (2002)

In Sect. 4, we turn to XAI techniques, their sources (data or model), types (global or local, direct vs. post hoc), and phases (ex ante or ex post). In Sect. 5, we focus on XAI ex-post use (as a forensic toolset), that is, how XAI can be incorporated into digital evidence processing, in steps 6 (examination), 7 (analysis), and 8 (presentation), and how these results, combined with others obtained during the accident analysis, fulfill the legal requirements to determine liability, closing the gap between our legal demands and XAI explanations.

4 Explainable artificial intelligence (XAI)

To investigate the consequences of autonomous AI systems that cause harm to individuals, and to face the issues related to legal liability, providing clarity as to how an AI decision was reached is important. Thus, special attention is given to the topic of explainability and the related (though distinct) topic of intelligibility. AI needs to be explainable, because explainability is a critical tool in building public trust and an understanding of the technology. By explainability, we mean a combination of the epistemological sense of "intelligibility" (i.e., how it works) and the sense of "accountability" (i.e., who is responsible). A suitable explanation must cover both.

Through XAI techniques and methods (Gunning & Aha, 2019), we can provide the necessary means to explain elements of the "reasoning" that led a machine to make a particular decision and the process therein. Please note interpretability is about being able to discern the mechanics without necessarily knowing the cause (i.e., how it works), and explainability refers to the untangling of the reasons (i.e., why it works).

So, what kinds of questions does XAI answer? Most importantly, it tries to answer "positive" questions. A positive question can be falsifiable, for example, "Did a certain condition lead the AI to misbehave?" On the other hand, a normative question might be "What should have happened?" or "Should the human have behaved in that way?" The first one should have a testable answer: yes or no. The answer to the other two hinges upon comparison: the former with a plausible scenario, the latter with a human behavior.

"Did a certain condition lead the AI to misbehave?' is one of the main reasons XAI exists. "What should have happened?" is the aim of an XAI technique called counterfactual. 'Should the human have behaved in that way?' is the objective of a study area in AI, namely, Moral AI.

One often hears "There ought to be a law on this" or "We should have an unbiased AI". These examples are political, centered upon values that are not falsifiable. Although they may be addressable, XAI aims to help us see how things are and not how they ought to be, even if it sometimes touches on the latter.

Such XAI techniques allow us different degrees of analysis. By examining the relationships between input and output variables in a functional way, local or global behaviors become apparent. For example, when calculating car insurance by changing its model, the insurance value changes proportionally (global effect), whereas changing the color of a car produces little or no change in the car's insurance value unless it's a premium paint job (local effect).

Another factor to consider is applicability. Most XAI techniques we cover can be applied to any model class, hereinafter model agnostic. However, some of these techniques have constraints on applicability and can only be applied to a specific model class to attain an explanation, hereinafter model specific.

We can classify models according to their opacity (i.e., lack of explainability). White-box (interpretable) models can be clearly explained in terms of how they behave, how they produce predictions, and the influencing variables. These models include rule-based models, decision trees, and monotonic gradient-boosting machines.

With black-box models, users can explain the input–output relationship, but the underlying reasons or processes in producing output are not available (i.e., explicitly, as in an interpretable model). Black-box models often result in increased accuracy over white-box models, but they sacrifice explainability. Black boxes are at the heart of our study going forward.

4.1 Explanation sources

In 2019, Gunning stated, "XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners."

Perhaps the most natural form of explanation is through example. In our XAI context, we focus on extracting representative specimens that grasp the inner relationships, quirks, and outer correlations of a model or prediction to get a better understanding of *why* it happened, that is, why the model behaves that way in a particular context. Table 5 illustrates the explanation tree; the leaf represents the source and type of normative question it explains. Although vital, visualization techniques have been omitted because they fit all categories.

As illustrated in Table 5, the sources of our explanations will be the data and the model. By analyzing the training data (ex-ante), we can understand the characteristics and attributes, that is, the correlated variables, of these data before starting training, identify irrelevant variables, discover or verify important relationships that machine-learning models must incorporate, and remove bias underlying the data before any modeling. For an ex-post analysis, usually via SRU data, we extract a chronological set of inputs and outputs, that is, the chain of events that lead to an act, providing a raw interpretation of what happened, and a factual set for further inquiring the model. The techniques employed in data explanation are covered in Sect. 4.1.1.

Regarding the model, we will always be concerned with understanding how it works from two points of view, global (holistic) and local (punctually), through explanations, features, and samples. By analyzing the model in training (ex-ante), we can investigate, through XAI techniques, questions about quality, safety, security, robustness, transparency, and so on. The result of these analyses will be crucial to support a foreseeability case, for example. The use of XAI techniques as a forensic toolset (ex-post use) allows us to explain their deliberations and behaviors, in

general or in specific situations, during the investigation. The techniques employed in the model explanation are covered in Sect. 4.1.2 to Sect. 4.1.4.

Figure 1 illustrates the main sources of ex-ante explanation. We can extract our explanations from the data. Once the model is trained, we obtain explanations about its operation and classify them according to their origin (direct and post hoc) and scope (global and local).

Next, we look at these categories in detail. We illustrate each of them with some of their main techniques, always addressing issues such as applicability (model specific or agnostic), intelligibility, origin, and scope.

4.1.1 Data explanation

Visualizing and understanding data are important for the model's intelligibility because they represent the data and their correlations. Therefore, understanding the content of those data helps us set reasonable expectations for the model's behavior and outputs. Most datasets are difficult to understand, and these techniques aim to mitigate some of the problems and highlight the relevance and correlations in those data.

For example, one may be quite familiar with their old family photos. Retrieving memories (i.e., knowledge) from a set of photos is simple when those photos are properly labelled and grouped with other similar photos, creating a context (cluster), but almost impossible when a box is labelled simply "old photos."

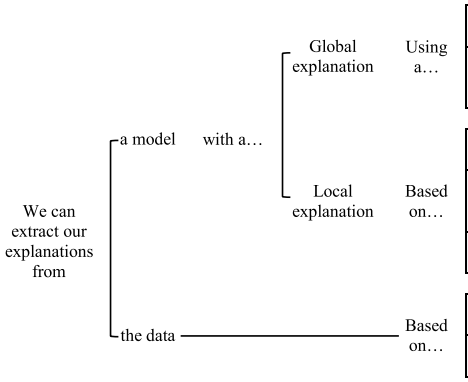
To help us illustrate important aspects of a dataset in few dimensions (i.e., project rows of a dataset from a high-dimensional space into a more visually understandable lower-dimensional space), we could use principal component analysis (PCA), usually done by performing eigenvalue decomposition of the covariance matrix of the data, T-distributed stochastic neighbor embedding (t-SNE), which is good at capturing non-linearities in data, or variational autoencoders (VAEs), a deep-learning method that learns how to encode data from a high dimension and then decode it back from a low to a high dimension. We also use correlation network graphs (i.e., 2D representation of the relationships in a dataset). Both techniques are model agnostic, providing some intelligibility, and can be used globally to see the entire dataset or to provide granular views of local portions of the dataset.

Once the model is trained, we can extract explanations about its operation, classify them according to their origin (direct and post hoc) and scope (global and local), and visualize them.

4.1.2 Direct vs. post-hoc explanation

This criterion addresses the origin of the explanation and distinguishes whether interpretability is achieved straight from the model (direct) or after training an explainable surrogate model. Direct explainability comes from models considered interpretable due to their simple structure, for example, rule-based models or decision trees.

Table 5 Explanation tree



Post-hoc explanation	Built on top of black box models to provide global understanding
Directly interpretable model	Provides safety, reliable, and compliance

Elicited explanation	Explanations elicited used to predict new samples
Features	Highlights features that are necessarily present or absent for the prediction to occur
Samples	Case-based reasoning using prototypes and criticisms

Features	Disentangled representations provide meaningful features
Samples	Case-based reasoning using prototypes and criticisms

Adapted from (Mojsilovic, 2019)

A prototype is a representative sample of data. A critique is a sample of data that is not well represented by the prototype set. Prototypes and critiques can be used independently to describe the data, create an interpretable model, or to interpret a black-box model

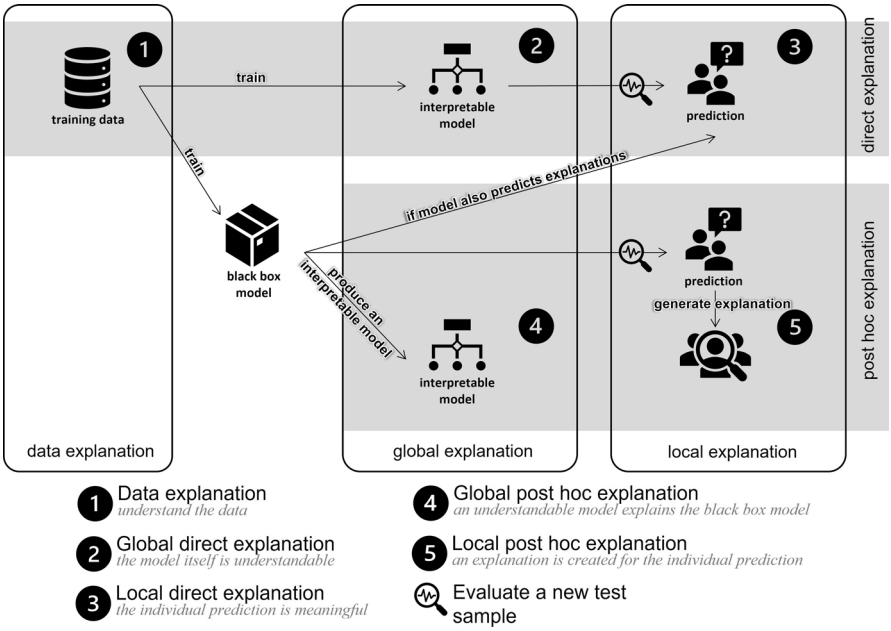


Fig. 1 Main sources and types of ex-ante explanation. Ibid

A rule-based model is composed of many simple Boolean statements that can be built using expert knowledge or learning from real data. They are model specific, provide high intelligibility, and can be used both globally and locally; they provide straightforward Boolean rules that can be easily understood by users.[8] Decision

[8] See Cohen (1995).

trees are data-derived flowcharts in which each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root to leaf represent classification rules. They are model specific, provide high intelligibility, and can be used globally; they provide a decision-structure map that can be easily visualized, interpreted, and audited.

Post-hoc interpretability refers to the craft of a white-box model (surrogate), implemented by optimizing its resemblance to its black-box model, keeping a similar performance but reducing its complexity and providing explainability, that is, making subsequent debugging, explanation, and fairness auditing tasks easier, for example, Explainable Neural Networks (XNNs).

XNN is a structured neural network designed especially to learn interpretable features. These features can be extracted from the network and the results displayed, providing an explanation of the relationship between the features and the output. XNNs are model agnostic (when used as a surrogate) and model specific (when standalone). They provide high intelligibility and can be used as surrogate models to explain other black-box models.[9]

4.1.3 Global vs. local explanation

This criterion deals with the scope of the explanation and distinguishes between interpretability to be achieved through a holistic view of the model (global) or strict observation (local). Some of the global-explainability algorithms use a carefully chosen series of local explanations to form a global explanation, because obtaining a satisfactory global explanation of a model is not always easy.

4.1.3.1 Local variable importance Local variable importance explains which input variables impacted a specific prediction. When these explanations are created automatically, they are typically called adverse-action notices or reason codes.

The best tools for the job are LIME[10] (approximate explanations using only the most important local variables) and Sharpley[11] (consistent local variable contributions to black-box model predictions). LIME trains a surrogate model by generating a new dataset out of the datapoint (currently text, image, and tabular data), resulting in a good local approximation. That is, imagine you had a loan denied by a bank; LIME would then generate a dataset with data similar to yours, that is, nearby address, similar income, and so on, and train a surrogate model that approximates the result of the bank black box. Because the trained model is explainable, you can understand the reasons behind AI denying your loan. Sharpley is better for interpreting an individual prediction, because it provides the contribution of each feature value and the amount of these contributions specifically, which is not the case for other techniques like LIME or counterfactuals.

[9] See Angelov and Soares (2019).
[10] See Ribeiro et al. (2016).
[11] See Lundberg and Lee (2017).

Both techniques are model agnostic, provide medium intelligibility (only local view), can be used locally, and enhance understanding by creating accurate explanations for each observation in a dataset.

4.1.3.2 Global variable importance Global variable importance quantifies the global contribution of each input variable to the predictions of a complex machine-learning model over an entire dataset. They state the magnitude of a variable's relationship with the response as compared with other variables used in the model. For some maximum likelihood models, global variable importance is the only commonly available measure of the relationships between input variables and the prediction target in a model.

The best tools for the job are greedy function approximation[12] (evaluates an input variable's global contribution to model predictions) and random forests[13] (creates several decision trees and measures the importance of an input by analyzing how many trees use the given input). Both techniques are model agnostic, provide medium intelligibility (only global view), and can be used globally and tell us about the most influential variables in a model and their relative rank.

4.1.4 Model visualization

Model visualization techniques provide graphical information about the behavior of almost any machine-learning model. Its main purpose is to assist the debugging of any prediction mistake that the system presents, but some become true integrated development environments (IDEs), incorporating numerous other metrics, functions, and tests. The best tools for the job are decision-tree surrogate models, ICE plots,[14] and partial dependence plots.[15]

A decision-tree surrogate model is an approximate, algorithmic-generated flowchart used to explain a black-box decision-making process. It is model agnostic, providing high intelligibility, and can be used globally.

For a local and global view, we have ICE plots and partial dependence plots, respectively. They point to how a prediction changes on certain input variables. They are both model agnostic, provide medium intelligibility, and can be used to provide local explanations. Also, partial dependence plots provide global explanations.

So far, we have seen a range of techniques that allow us to extract and visualize explanations from different models and perspectives. But XAI is not restricted to just explaining AI behavior; techniques extend to security[16] and bias issues,[17] allowing us also to answer some normative and political questions. Yet, XAI is not

[12] See Friedman (2001).
[13] See Ho (1995).
[14] See Goldstein et al. (2015).
[15] See Friedman (2001).
[16] See Gu et al. (2019) and Nicolae et al. (2018).
[17] See Verma and J Rubin (2018), d'Alessandro et al. (2017), and Friedler et al. (2016).

failproof, and emerging issues may demand new techniques, or at least the adaptation of those already in existence.

4.2 Ex-ante and Ex-post explanations

Imagine something goes wrong, such as the accident involving an autonomous vehicle in Illustrations 1 and 3. We would like to be able to analyze the accident ex post, that is, determine those involved, those responsible, aggravating and mitigating factors, chronology, and so on, as well as the AI's prior compliance with quality and safety standards, that is, the producer's ex-ante analysis approved by competent authorities.

That demands leaves us with two distinct moments when the applicability of XAI techniques and the need for data provided by the SRU will be of great value. First, in the genesis of the model (ex-ante), XAI techniques can be used to provide explanations, for example, to assess intelligibility, fairness, quality, correctness, and compliance, and the SRU must attest the entire process to provide accountability, at least data lineage. Second, in an accident investigation (ex post), XAI techniques are used in a forensic-analysis framework with data collected from the SRU, providing an ex-post explanation and assisting in the factual reconstruction of the occurrence.

Before examining XAI techniques as a forensic toolset (ex-post use), it is important to look at a simplified version of most machine-learning manufacturing processes using the CRoss-Industry Standard Process Model (CRISP-DM) methodology. Although CRISP-DM is a data-mining methodology, most developers follow it with some adaptations.[18]

According to CRISP-DM, the process involves six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment.[19]

At the business understanding phase, we focus on understanding the project objectives and requirements from a business perspective. In data understanding, we identify data-quality problems and detect interesting subsets to form hypotheses for hidden information. The data preparation phase covers all activities to construct the final dataset (data that will be fed into the model) from raw data. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modelling tools. We employ our first set of XAI techniques, data visualization, after data preparation. SRU should account not only for the data understanding and data preparation, but also the process, objectives, requirements, and people assigned at the business understanding phase, thereby establishing a data lineage.

Modelling is also known as training, and in this phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. Evaluation is the phase in which one assesses the model and reviews the steps executed to construct the model to be certain it properly achieves the objectives. Almost all our XAI techniques are employed in these two phases for the purposes

[18] See (Piatetsky-Shapiro 2007).
[19] See (Harper and Pickett 2006).

of debugging (i.e., adjusts and fixes), compliance, intelligibility, fairness, and other goals set in the business understanding phase. Deployment, the final phase, usually involves the release of the model to the customer.[20] As aforementioned, SRU should account for both phases and is an essential tool at the deployment phase for monitoring the product during its life cycle and as a data recorder.

The use of the XAI techniques shown in Fig. 1 occur in modelling and evaluation phases of CRISP-DM and aim to obtain explanations about the AI. We can adapt these techniques for ad-hoc purposes, for example, fairness or robustness, which may lead us to embrace earlier stages of the process, for example, data understanding and data preparation, to remedy any inconsistencies found. When businesses incorporate explainability and accountability into their machine-learning manufacturing processes, they foster an increase in trust in all stakeholders, resulting in higher aggregate value, better decision-making, and profitability.

Under a strict liability regime, a "before the accident" (ex-ante) explanation may be useful as a means of proof. When manufacturers are held responsible for defective products, regardless of fault, they may want to prove their AI product was not defective. Alternatively, they might want to prove that any defect did not cause loss or damage to the consumer. A similar reasoning can be applied in a foreseeability case, that is, applying XAI to demonstrate that the foreseeable damages known at the time were appropriately tested for and suitably avoided. Also, lawsuits, poor decision-making, and tarnished reputations are mitigated or avoided more easily by adopting XAI as a prophylaxis.

However, even with all the zeal employed by the manufacturer throughout its production chain, accidents are unexpected and unwanted but are part of life. When an accident investigation demands an ex-post investigation, XAI techniques might be used in a forensic-analysis framework with data collected from the SRU, providing an ex-post explanation and assisting in the factual reconstruction of the occurrence.

5 XAI techniques as a forensic toolset

After going through this arduous journey, starting from the legal requirements and settling on the AI whys, one should expect that the application of the XAI techniques will resemble the application of the Voigt-Kampff test on Rachael by Deckard, that is, a set of questions to determine her nature by measuring functions such as breathing, heart rate, and pupillary dilation in response to emotionally provocative questions. Our task is, in fact, to apply a series of tests, mostly statistical, to explain the AI data and behaviors. But our goal is to answer how and why things happened, reconstruct the story, that is, factual causation (what) and legal causation (how and why), and establish contributing and validate underlying factors.

We should have a clear definition of what we are investigating; otherwise, we could seize upon a possible scenario and then look for facts that corroborate that scenario, despite conflicting evidence. Additionally, from a plaintiff's perspective,

[20] See Chapman et al. (2000).

the explanation provided by an XAI could be paradoxical. The ex-post explanations usually demonstrate the defendant's fault, but they can also prove a different wrongdoer was responsible or that the plaintiff's carelessness caused the loss or damage, releasing the defendant from the duty to indemnify.

Because it is a specialization of digital forensic, the forensic explainable AI methodology follows the same steps described in Table 4. Next, we focus on the three steps that require more attention: examination, analysis, and presentation.

5.1 Examination

If we go back to our goals (specifically reconstruct the story), here is when we write up the script. An in-depth systematic search of evidence relating to the suspected crime is conducted, focusing on identifying and locating potential evidence, possibly within unconventional locations. As result, we construct a detailed documentation for analysis.

In Illustration 1 (a vehicle with autonomous-driving software that swerves into the opposite lane, hitting an innocent third party), if we ask *what* happened, in order to show *factual causation*, we may state that the vehicle with autonomous-driving software collided with a car and caused the accident. To establish *legal causation*, we must answer *how* and *why* it happened.

How did it happen?

(1) Autonomous vehicle *A* was travelling on a two-lane road.
(2) Another vehicle, *H*, drove in the opposite direction, passing onto the lane of the autonomous vehicle *A*.
(3) The autonomous vehicle *A* changed lanes and collided with vehicle *I*.

Imagine for a moment that we do not have any XAI techniques at our disposal—only current forensic techniques. In (1), based on the tire marks and the telemetric information contained in the cars—we can assume that at least the car piloted by AI has a form of a restraint control module —we would already extract enough evidence to establish how the accident happened.

In (2), when we reconstruct the path taken by *H*, we know if it changed direction to avoid a hole or an animal, entering the range of *A*.

In (3), we can establish that *A* chose to deviate from its route and ended up colliding with *I*. But without XAI, we cannot establish why *A* chose to change its route. Certainly, we can conjecture that it was simply an attempt to avoid colliding with *H*, but we still need to understand the root of its decision. Why did *A* not change to *H*'s lane? Did *A* see *I*? Did it decide *H* would go back to its original lane and slow down to avoid the accident? Did *A* consider it might collide with *I*? These and other questions cannot be answered without XAI techniques. Answering these questions is essential in determining the reasons behind how the accident occurred, such as, who made mistakes and why they were made, and even to be able to compare the behavior of AI with that of a human being.

We can simulate the accident with the information obtained from the SRU and establish what the car "noticed" moments before the accident, what behavior it predicted for each car, how (and based on what) it decided to change lanes, why it chose *I*'s lane, and whether it knew it might collide with *I*. Then, depending on the data recorded and recovered from the SRU, images and videos of the accident may or may not prove *H*'s claims. The XAI techniques to be used vary according to the technology used in each of the components being audited, for example, the module responsible for the recognition of the cars and other obstacles will probably be audited by an XNN technique, whereas counterfactuals and surrogate models will be used to explain the component that decided to change lanes.

As shown, we have two distinct and equally important tasks: establish what evidence we will have at our disposal for the analysis phase and craft the *evaluation script* that will be applied to the evidence.

Establishing what evidence we will use for the analysis sounds obvious if we have the SRU data in hand. However, sometimes we come across situations in which we only recover part (or none) of these data. At other times, cross-referencing information from different evidence (not necessarily digital) about the same event is interesting or even necessary. Occasionally, we come across situations in which we need to augment our data (even if we have SRU data), compile data from similar cases, or use other databases to run certain tests.

For this reason, explaining the methodology used at the data-preparation phase is important, always attentive to the principles described in steps 1 to 5 of the digital forensic process (Table 4), with special emphasis on justifying the sources, documenting modifications (data lineage), and explaining both (see item 4.1.1). Our next assignment, evaluating the script, has a large impact on these choices.

For the task of crafting the evaluation script, we need to establish which questions we want to answer, as well as outline areas to be covered and major questions to be answered and determine what information should be gained from the answers.

Grouping the questions by purpose is interesting. For example, when executing an AI recollection statement, illustrated in Fig. 2 item 1 – SRU data explanation, we obtain information about what happened before, during, and after the accident, but that data may contain other pertinent information not directly linked to what happened.

The information obtained, both about the details of the accident and information not directly linked, may be sufficient to answer other questions, raise new ones, or even be sufficient for what we intend and can analyze. This enlightenment allows for a lower and more efficient test demand and enables us to see more easily what other datasets we need to answer the rest of the questions. Then, we correlate the questions we want to ask with the tests we can run to answer them. Constructing an evaluation script may require take several iterations of data preparation and questions evaluation.

The type of question (positive or normative) we ask is also important. Recall that our techniques support both, but the assumptions are different. For example, in Illustration 1, we could ask if the autonomous car saw the third party involved (positive question) or why the autonomous car did not change to the opposite lane (normative question).

To answer the first question, we can extract our explanation from the SRU data, that is, based on the samples we found, and answer yes or no. For the second question, we can extract our answers from the AI with a broad perspective, that is, how the car generally behaves in that kind of situation, or from a more specific perspective, that is, why the car behaved that way in that particular case. Our global-explanation techniques, Fig. 2 items 2 and 4, fulfill the demand of the broad question, whereas the strict question demands the employment of a local explanation, Fig. 2 items 3 and 5, evaluated using augmented SRU data.

When eliciting and grouping questions, we should always keep in mind their objectives, for example, establish the context of the task performed by the AI and the environment in which it was performed, or understand why—not just what—happened in a certain case. This focus not only allows us to be more efficient, but also helps us establish whether we need tests that explain the AI, assess bias, and robustness or whether these explanations will require other non-XAI specific techniques.

(1) ***SRU data explanation*** – evaluates the information obtained from the SRU, showing what happened before, during, and after the accident, that is, a sort of witness recollection of the accident from the AI point of view and what it predicted in each moment.
(2) ***Global direct explanation*** – evaluates the deliberation process of an explainable model in a holistic view, for example, most influential variables and the primary behavior.
(3) ***Local direct explanation*** – evaluates an individual prediction (meaningful) or a set of data for hypothesis testing.
(4) ***Global post-hoc explanation*** – evaluates the deliberation process of a black box, through a surrogate model, from a holistic view, for example, most influential variables and the primary behavior.
(5) ***Local post-hoc explanation*** – evaluates an individual prediction or a dataset for hypothesis testing and explains them.
(6) ***Augmented SRU data*** – an enlarged copy of a sample from the SRU dataset, that is, with newly created data or slightly modified copies, for hypothesis testing.

Figure 2 illustrates the main sources of ex-post explanation. We can extract our explanations from the SRU data. Later, we obtain explanations about AI's modus operandi or specific deliberations. As such, we classify them according to their origin (direct and post hoc) and scope (global and local).

Once our systematic search for evidence is completed and we have determined which tests to perform, we can prepare the detailed documentation for analysis, namely, the evaluation script.

5.2 Analysis

In the previous stage, we were concerned with reconstructing the story; here, we are concerned with answering the *whys*, establishing contributing factors, and validating

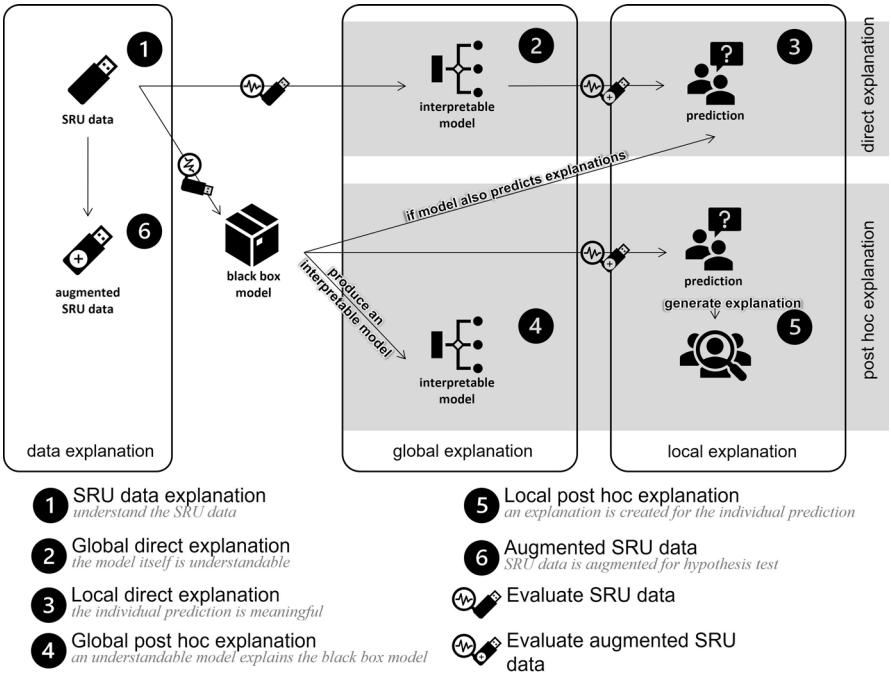


Fig. 2 Main sources and types of ex-post explanation

underlying factors. We seek to determine significance, explain the model and data, and draw conclusions based on evidence found.

Supporting an accident theory may take several iterations of examination and analysis. The distinction in this analysis is that it may not require high technical skills to perform, and thus, more people can work on this case. In fact, many of these tests can be automated.

Although we have already discussed several techniques, especially in item 4.1 and other sub-items, which can and are used in this phase, we now turn to discussing two examples. The first consists of using an XNN, to explain convolutional neural networks (CNN). The second is the employment of XAI techniques by detractors to mask dubious or discriminatory behaviors, highlighting the importance of an analyst.

In Illustration 2 (a door lock that uses embedded face recognition and recognized a burglar), one of the most popular techniques used in improving the accuracy of image classification is CNN, a neural network that has a convolution layer at the beginning instead of feeding the entire image as an array of numbers. CNN breaks up the image into several tiles, reducing them into an easier form to be processed. This breakup process is important when we are dealing with scalable or massive datasets, and crucial to avoid losing features and achieving accurate prediction. CNN then tries to predict the nature of each tile and feeds them in its convolutional layer.

When playing Pictionary, because time is critical, we start our drawings with low-level features, such as shapes, and subsequentially high-level features. That is,

we form a set of descriptions that let us classify (or guess) with confidence what we are "seeing." Similarly, when we are asked to describe someone we just met, we start with their main features, like eye and hair color, and then move on to idiosyncrasies, like slightly asymmetric ears or the person's frowns.

Likewise, CNNs are not limited to only one convolutional layer. Formally, the first layer is responsible for capturing the low-level features, for example, edges, color, orientation, and so on, with subsequent layers capturing high-level features. After going through the above process, the convolutional layers produce a general understanding of the tiles, similar to how humans would**.** Finally, these tiles are fed into a neural network that tries to predict what it depicts. This process allows the computer to parallelize the operations and detect the object regardless of where it might be located inside the image.

To show the contributing factors or validate underlying factors, we can use the XNN to explain CNN and understand what features were learned, making them explicit by feature visualization. Furthermore, using network dissection, we can link highly activated areas of CNN channels with human concepts (in this case, humans and primates). Also, we can use data-explanation techniques to measure distributions in training data—thereby discovering a possible source for the problem—as well as fairness techniques to measure bias and discrepancies.

As we've seen, specific techniques can be helpful when we run into certain problems. Next, we deal with a case with no XAI-specific technique to attack the problem, but we can still investigate and demonstrate liability.

Dieselgate is a name coined by the press for the pollutant emission testing scandal involving several car manufacturers around the world (Chappell 2015). The US Environmental Protection Agency had discovered software installed in Volkswagen vehicles that changed pollutant emission numbers only when cars were tested. Subsequent investigations uncovered the same practice in other countries and by other manufacturers, leading to one of the biggest crises in the history of the auto industry. Now, imagine the employment of XAI techniques not to explain or improve the AI, but to mask dubious or discriminatory behaviors.

Fairwashing is the practice of promoting a false perception that a machine-learning model respects some ethical values. Due to the growing importance of the concepts of fairness in machine learning, and because the right to explanation in GDPR does not give precise directives on what providing a "valid explanation" means, a legal loophole can be exploited by dishonest companies to cover up some possible unfair issues of their black-box models, by providing misleading explanations (i.e., rationalization).

Rationalization consists of finding an interpretable surrogate models approximating a black-box model b, such that s is fairer than b. To achieve fairwashing, the surrogate model obtained through rationalization could be shown to the auditor (e.g., an external dedicated entity or the users themselves) to convince him the company is "clean."[21] For this reason, regulatory agencies' auditions, independent testing, and a robust certification loop are vital to mitigate fairwashing.

[21] See Aïvodji et al. (2019).

The tools we have covered so far do not address the issue of fairness in their analysis. Some models, however,[22] provide model inspection but cannot be used for model or outcome ex-post explanations. For now, no algorithm can estimate whether an explanation is likely to be a rationalization or able to detect fairwashing by itself, and a human analysis is essential to detect an enshrouding attempt.

This example highlights the importance of the holistic process of investigation and not just the indiscriminate use of data-analysis techniques. The investigation as a whole must analyze the task the AI was performing, its potential to succeed at the task, the processing of job information, engineered barriers (e.g., cybersecurity and robustness AI techniques), administrative defenses, oversight defenses, and cultural defenses.

Now, if the analysis step just described can be depicted as a set of scientific experiment footage with a stimulating background music, the next step (presentation) is the notorious scene of the famous detective exposing the case and revealing the culprit.

5.3 Presentation

In the previous stage, we were concerned with determining significance, explaining the model and data, and drawing conclusions based on evidence found. Here, we are concerned with the *mise en scène*, that is, the arrangement of our conclusions in the reconstructed story.

Here, we summarize and provide explanation of our findings in a formal, permanent, auditable, defensible report. It should be written in a layperson's terms using abstracted terminology (referencing the specific details).

The professional responsible for the presentation must be able to determine the best way to deploy information. For example, to illustrate the accidents of Illustrations 1 and 3 highlighting the AI deliberations, developing sketches and diagrams by pinpointing key moments of the accident (locations, actions performed, witness recollections, etc.) constitutes a simple, visual, and accessible way.

Because the focus of XAI is on explaining the AI, model, and data, we generally use model visualization techniques (4.1.4.) and data visualization (4.1.1.). Three fundamental aspects must be considered when developing our findings presentation: the target audience, data/task abstraction, and encoding.

Target audience – encompasses a group of target users, their domain of interest, their questions, and their data. For instance, a judge needs answers to normative questions do determine liability, whereas an accident investigator needs evidence and explanations.

Data/task abstraction – mapping those target-audience problems and data into forms that are independent of the domain. For instance, in Illustration 3, an accident investigator might want to compare AI performance in different light-

[22] See Berendt and Preibusch (2012) and Adebayo (2016).

ing conditions, whereas a judge compares the AI performance with a human at that particular condition.

Encoding - deciding on the specific way to address the tasks previously listed. For instance, the test with different lighting conditions could be represented by a bar chart or a photo matrix with the different results illustrated, whereas the comparison of AI versus. human, a single percentile, or an "*n times better*" phrase may suffice.

At the end, all these results will be duly substantiated and packaged in a report that will be part of the digital forensic report. All these forensic findings will be incorporated into the RCA process, which in turn will generate a report. The RCA report will be incorporated into the accident report, which in turn must present, in a substantiated form, the factual and legal causation demanded by court.

The process and techniques above could explain the reasons that led the AI autonomous system to make its decision, given the account of how the accident occurred (through SRU) and why the loss or damage then happened (through XAI).

6 Conclusion

The use of these explanatory techniques would help simplify many complex problems that can occur with AI systems and autonomous decision-making, such as the problem of shared responsibility and a lack of knowledge about how AI systems make decisions and reach robust legal outcomes. Their further development and adoption should allow AI liability cases to be decided under current legal and regulatory rules, until (if it ever happens) new liability regimes for AI are enacted.

One could ask if the above explanations are sufficient to determine liability in cases of loss or damage as a result of an AI decision. In our point of view, the answer is yes. Since XAI techniques can answer *what*, *how*, and *why*, and, by answering these questions, we can establish the *factual* and *legal causation* (required by common law) and the *causal nexus* (required by the civil law), the obligations required by both legal systems to establish causation are fulfilled. Thus, courts will be able to proportionately assign liability to such failings and deal with problems of shared responsibility and a lack of knowledge about AI system decision processes.

The form may be adapted for the audience, but the narrative remains the same. The vocabulary used in the description of the case presented to a judge differs, sometimes substantially, from the vocabulary used for the jury or adopted by technicians. Legal requirements may also oblige us to follow predetermined ways of presenting information, for example, reports, transcripts, proceedings, and records. What is acceptable, or demanded, depends on the audience and finality and will evolve over time, which is an interesting and needed follow-up topic. In the early days, we will require humans to translate some XAI explanations into the required form. Whether XAI tools can be devised to produce different explanations for different audiences is potentially a new research topic.

Acknowledgements We gratefully acknowledge Dr Armando Castro's invaluable comments on the revision of this article.

References

Adebayo, J. A. (2016) FairML: ToolBox for diagnosing bias in predictive modeling (Doctoral dissertation, Massachusetts Institute of Technology)

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019) Fairwashing: the risk of rationalization. In International Conference on Machine Learning (pp. 161–170). PMLR

Angelov, P., Soares, E. (2019). Towards Explainable Deep Neural Networks (xDNN) (2019). Cornell University. ArXiv: https://arxiv.org/abs/1912.02523

Berendt, B., & Preibusch, S. (2012) Exploring discrimination: A user-centric evaluation of discrimination-aware data mining. In 2012 IEEE 12th International Conference on Data Mining Workshops (pp. 344–351). IEEE

Bloch, H. P. (2005) Successful failure analysis strategies. Reliability Advantage: Training Bulletin. Retrieved from http://www.heinzbloch.com/docs/ReliabilityAdvantage/Reliability_Advantage_Volume_3.pdf

Bloch, H. P. (2011) Structured failure analysis strategies solve pump problems. Machinery Lubrication. Retrieved from http://www.machinerylubrication.com/Read/28467/pump-failure-analysis

Calo, R., Froomkin, M. & Kerr, I. (2016) *Robot Law* Edward Elgar Edward Elgar Publishing UK

Carrier, B. (2002) Defining Digital Forensic Examination and Analysis Tools. In 2002 Digital Forensics Research Workshop

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000) CRISP-DM 1.0: Step-by-step Data Mining Guide. SPSS

Chappell, B. (2015) It Was Installed For This Purpose,' VW's U.S. CEO Tells Congress About Defeat Device. Retrieved from NPR: https://www.npr.org/sections/thetwo-way/2015/10/08/446861855/volkswagen-u-s-ceo-faces-questions-on-capitol-hill (22 May 2020)

Cohen, W., (1995) Fast effective rule induction. Proceedings of the Twelfth International Conference on International Conference on Machine Learning. Morgan Kaufmann Publishers Inc Elsevier 115–123

d'Alessandro B, O'Neil C, LaGatta T (2017) Conscientious classification: a data scientist's guide to discrimination-aware classification. Big Data 5:120–134. https://doi.org/10.1089/big.2016.0048

Fisher, D. (2019). Explainable AI: Addressing Trust, Utility, Liability. (31 May 2019). aitrends The Business and Technology of Enterprise AI. https://www.aitrends.com/explainable-ai/a-chief-ai-officer-on-explainable-ai-addressing-trust-utility-liability

Friedler, S., Scheidegger, C. & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. Cornell University. ArXiv: https://arxiv.org/abs/1609.07236

Friedman, J. H. (2001) Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5) https://projecteuclid.org/euclid.aos/1013203451

Goldstein A, Kepelner A, Bleich J, Pitkin E (2015) Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Statistics 24(1):44–65

Goudkamp, J. & Peel, W.E. (2014) *Tort Winfield & Jolowicz.* Sweet & Maxwell

Gunning D, Aha D (2019) DARPA's Explainable Artificial Intelligence (XAI) Program. AI Mag. https://doi.org/10.1609/aimag.v40i2.2850

Gunning, D. (2019) DARPA's explainable artificial intelligence (XAI) program. Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI, p. 47

Harper G, Pickett S (2006) Methods for mining HTS data. Drug Discovery Today 11(15–16):694

Ho, T.K., (1995) Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, IEEE, 278 – 282

House of Lords, Select Committee on Artificial Intelligence, AI in the UK: ready, willing and able? (2018) (Report of Session 2017–19)

Lundberg SM, Lee S (2017) A unified approach to interpreting model predictions. Adv Neural Information Processing Sys NIPS' 17:465–474

Mojsilovic, A. (2019). Introducing AI Explainability 360. https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/ (4 June 2021)

Muschara, T. (2007). INPO's approach to human performance in the United States commercial nuclear power industry. IEEE Xplore Digital Library. Retrieved from http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4413179&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4413179

National Transportation Safety Board (NTSB). (March 18, 2018) (Preliminary report highway: HWY18MH0102018) <https://www.ntsb.gov/pages/default.aspx (24 February 2020)

Nicolae, M. & Sinn, M. (2018). Adversarial Robustness Toolbox v1.0.0. Cornell University https://arxiv.org/abs/1807.01069 (24 May 2020)

Palmer, G. (2001). A Road Map for Digital Forensic Research. Technical Report DTR-T001–01, DFRWS, Report From the First Digital Forensic Research Workshop (DFRWS).

Piatetsky-Shapiro, G. (2007). Methodology Poll. (2007). KDnuggets. https://www.kdnuggets.com/polls/2007/data_mining_methodology.htm (14 June 2020)

Reed, C., Kennedy, E. & Silva, S. (2016). Responsibility, Autonomy and Accountability: legal liability for machine learning. Queen Mary University of London, School of Law Legal Studies Research Paper No. 243/2016

Reith M, Carr C, Gunsch G (2002) An Examination of Digital Forensic Models. Int J Digital Evidence 1:3

Ribeiro, M. T., Singh, S., Guestrin, C. (2016) Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining KDD 16: 1135–1144

Verma, S. & Rubin, J. (2018) Fairness Definitions Explained. ACM/IEEE International Workshop on Software Fairness Gothenburg: IEEE, p.1

Wilson PF, Dell LD, Anderson GF (1993) Root Cause Analysis: A Tool for Total Quality Management. ASQ Quality Press, Milwaukee, Wisconsin

Wright RW (1985) Causation in tort law. Calif Law Rev 73(6):1735–1828. https://doi.org/10.2307/3480373