Forensic Science International: Digital Investigation 42 (2022) 301403

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS 2022 USA - Proceedings of the Twenty-Second Annual DFRWS USA

Explainable digital forensics AI: Towards mitigating distrust in AI-based digital forensics analysis using interpretable models

Abiodun A. Solanke

CIRSFID Alma-AI, University of Bologna, Italy

ARTICLE INFO

Article history:

Keywords:
Digital forensics AI
Evidence mining
Explainable AI
Interpretable AI
AI and Law

ABSTRACT

The present level of skepticism expressed by courts, legal practitioners, and the general public over Artificial Intelligence (AI) based digital evidence extraction techniques has been observed, and understandably so. Concerns have been raised about closed-box AI models' transparency and their suitability for use in digital evidence mining. While AI models are firmly rooted in mathematical, statistical, and computational theories, the argument has centered on their explainability and understandability, particularly in terms of how they arrive at certain conclusions. This paper examines the issues with closed-box models; the goals; and methods of explainability/interpretability. Most importantly, recommendations for interpretable AI-based digital forensics (DF) investigation are proposed.

1. Introduction

During the last two decades, machine-generated proofs have mostly taken over the function of humans in fact-finding, albeit with increased accuracy (Roth, 2015). There are considerable concerns about the legality of digital evidence or machine-generated conclusions, particularly given that these decisions can differ for the same scientific evidence, just as they do with human experts. Similarly, just as out-of-court testimony, such as hearsay (Goodison et al., 2015), machine testimonies (sources) may create closed-box[1] problems for the justice system, leading fact-finders to make incorrect/incomplete inferences (Carr, 2014; Pasquale, 2015). Although the design, input, model, and environment can all contribute to the flaws or inaccurate interpretations of a machine-driven DF analysis, the most likely causes are erroneous algorithms/code, skewed or disproportionate datasets, and defective functional components of the system (e.g., OS, distributed platforms, etc.). Humans are responsible for designing and structuring all important components of a machine (including its design, input, and operational modules), and so some scholars assert that machines' credibility is strongly reliant on humans. As a result, the true declarant[2] of any output that a machine is capable of producing is a human being (Wolfson, 2005). While the designer or operator of a machine bears some moral responsibility for the statements it makes, she is not the sole source of such statements (Roth, 2015). She is only reiterating to the audience the output that a machine generated. A machine-driven forensic investigation, like an expert opinion, is the product of "distributed cognition" between humans and technology (Dror and Mnookin, 2010). As noted previously, humans and machines are inextricably linked in a variety of ways, which impacts everything from the closed-box to determining responsibility.

AI and its inscrutability (opaqueness) remain active study areas; yet given widespread misconceptions about whether AI systems should be explainable or interpretable, the road to a unifying consensus may be longer. AI/Machine Learning (ML) powered systems have a wide variety of applications in our daily lives, with differing implications in each sector. Where judgments have a substantial impact on individuals, or where accountability, transparency, or legal compliance are required (for example, in health and law), there is a rising concern about the inexplicability of AI systems (Coyle and Weller, 2020). This has prompted calls for forensic investigation of AI systems (Baggili and Behzadan, 2020) and auditing of their application in a variety of scenarios (Schneider and Breitinger, 2020) in order to ascertain their behaviours. Intelligent systems have proven particularly useful in refuting or supporting claims in DF investigation, as they have identified or detected interesting clues that could have been missed or overlooked. There is an additional degree of complication that needs to be addressed when trying to explain a forensic investigation's

Abbreviations: DFAI, Digital Forensics AI.

E-mail address: abiodun.solanke@unibo.it.

[1] See section 4.

[2] "Declarant" is a term used in the context of hearsay as a label for the witness tendering evidence statement as truth of the matter asserted.

https://doi.org/10.1016/j.fsidi.2022.301403

findings, because the methods used to arrive at such conclusions may be questionable scientifically or insufficiently transparent. As technology has become more sophisticated, so has the crime that is committed with it, necessitating a shift from traditional methods (such as forensic tools familiar to lawyers, jurors, and others) to a more robust, and equally intelligent systems such as AI to identify potential evidence.

The primary goal of this work is to examine, first, the diverse ideas on explainability and interpretability in AI, with a specific focus on how they affect DF and evidence mined using AI algorithms. This is necessary in order to provide a solid foundation for such ambiguous ideas. To put things in the right perspective, guidance through literature will, perhaps, help to draw the right connections especially as it pertains to digital forensics AI (DFAI)[3] (Solanke and Biasiotti, 2022). Second, the many approaches and attempts to find a viable answer to the issue of closed-boxes are discussed (even though that remains elusive). Domain-specific recommendations to mitigate distrust in digital evidence mining[4] are then offered after discussions about several work-around proposed.

The key contribution of this paper are the recommendations offered for mitigating mistrust in AI-powered digital forensics investigations. Additionally, a formal pre-concept for explainable digital forensics AI is presented, as well as various relevant methods for providing understandable interpretations for AI models and their applicability to AI-based DF analysis.

The next sections discuss the concepts of explainability and interpretability; the goals and methods for interpreting AI models; and recommendations for making the application of AI in digital forensics more interpretable.

2. The concepts

The promise of AI was to enable better decision-making, as seen in some forms of medical diagnostics (De Fauw et al., 2018) or monitoring attempted financial frauds (Aziz and Dowling, 2019), but doubts have been raised about its use in critical contexts like justice and policing systems (Aziz and Dowling, 2019). There is a pressing demand to explain to audience who might be curious about how algorithmic decisions were reached. Explainable AI (XAI) (Samek et al., 2017, 2019; Pedreschi et al., 2018; Guidotti et al., 2019), is an area of research that is focused on making AI systems and the data they utilize transparent by "glass-boxing" the system's functioning components (Gross-Brown et al., 2015). In light of AI's broad use in many sectors, different explanations connote diverse meanings, and the weight of significance is assigned based on the technical requirements and the implications of the outcomes. For instance, the decision-making process of a recommender system requires little or no explanation, while questions about the decision-making mechanism of a crime prediction or recidivism algorithm will be raised. Since a wrong machine-generated decision could have serious consequences on law enforcement, and the criminal justice system as a whole, XAI holds a lot of weight. XAI idea stems from the continuous effort to minimize (or eliminate entirely) the opaqueness of AI systems through the deconstruction of complex variables while maintaining a good balance between transparency, performance, and correctness. For this, there have been arguments over whether the outcomes of a closed-box AI system should be explainable (Arrieta et al., 2020) or interpretable (Rudin, 2019); some argue instead for systems that are intelligible or responsible (Benjamins et al., 2019). However, interpretable and explainable AI, in particular, have been used interchangeably across literatures. A simple search in the Scopus[5] database highlights these misconceptions over time and the gradual shift in reasoning towards interpretability in literatures. According to the search, "interpretable AI" was more prevalent over time until 2018, before explainability started getting formalized. Interpretable AI (IAI) and XAI are now widely used in a range of fields of study, including health and decision sciences (to which, perhaps, DF belongs), in addition to the primary fields in which the concept was majorly prevalent (e.g. computer science, mathematics, engineering, social science, etc.).

To better understand these concepts, definitions and distinctions between terms may be required; thus the summary of the most widely used nomenclatures are offered below.

Explainability: relates to the idea of connecting a machine's decision-making process with human explanations that are both accurate and understandable (Guidotti et al., 2019). It embodies the notion that, AI models and their output can be rationally explained in a way that humans can accept and understand. Despite their lower performance, classical ML models are fairly easy to understand. Deep Neural Networks/Deep Learning (DNN/DL) (LeCun et al., 2015), on the other hand, performs better but is considerably more difficult to explain. AI systems that are truly explainable uses knowledge bases for data analysis and provide a technique for deconstructing the results in a way that logically justifies the interpretations of their input data (Hall et al., 2021). According to Gunning (2019), *"XAI will create a suite of machine learning techniques that enable human users to understand, appropriately trust, and effectively manage the emerging generations of artificially intelligent partners."*

Interpretability: is the ability to communicate an explanation or meaning in a way that is comprehensible (Arrieta et al., 2020). A universal definition might be impossible since interpretability is domain-specific (Ruping, 2006; Huysman et al., 2011). It is important to note, however, that interpretability in the context of machine-generated output should be regarded in terms of its conformance to structural domain knowledge; causality; or physical constraints; and, of course, sparsity (of data); which can be measured in terms of human cognitive capacity (Miller, 1956; Cowan, 2010). In addition to being able to visualize a model, an interpretable system allows users to examine and comprehend the mathematical underpinnings of how input is transferred to output (Doran et al., 2017). It conveys a sense of transparency and clarity. Interpretable consideration can help improve the implementation of an AI model in three ways: 1) ensure objectivity in decision-making; 2) ensure resilience to adversarial perturbations that could impair prediction; and 3) ensure that only correct variables are used to infer the output, i.e., assurance that true causality underpins the model reasoning (Arrieta et al., 2020). For an interpretable AI system to be effective, the predictions it makes must be understandable, its discriminating rules must be visualizable, and any circumstances that could perturb the model must be disclosed (Hall, 2018).

Understandability: or intelligibility, refers to the features of a model that allow it to be self-explanatory in terms of its operational functionality —without the need to describe its internal structure or the underlying algorithms used to process data ((Montavon et al.,

[3] 'Digital Forensics AI' herein refers to a broader concept of automated systems that encompasses the scientific and legal tools, models, methods; including evaluation, standardization, optimization, interpretability, and understandability of AI techniques (or AI-enabled tools) deployed in digital forensics domain.

[4] 'digital evidence mining' as the process of automatically identifying, detecting, extracting, and analyzing digital evidence with AI-driven techniques. Mining is borrowed from the phrase 'Data Mining'.

[5] https://www.scopus.com/home.uri. A larger body of works, however, may have more references to explainability/interpretability than the titles, abstracts, or keywords that are considered (in this study) from a single database.

2018).

Comprehensibility: is often quantified in terms of the model's complexity (Guidotti et al., 2019), which includes the model's ability to describe its learning process in a comprehensible manner (Crave, 1996; Gleicher, 2016). Comprehensibility is commonly achieved in AI by including deductive symbols in the model's output, which permits reverse engineering, and by establishing links between output features and their corresponding inputs.

Transparency: Algorithmic transparency, simulatability (i.e., the ease with which the system may be replicated), decomposability (i.e., chunking, and easy analysis of the functional components), and transparency are all characteristics that a transparent model should posses (Lipton, 2018).

All above-defined concepts are interwoven in that they emphasize the significance of AI models that are understandable, precise, and objective in their decision-making. It is easy to misinterpret the fundamental meaning of these concepts, and of course, in this paper, they are used interchangeably. Most significantly, this paper places considerable emphasis on two concepts: explainability and interpretability, and while other notions are presented, the goal is to determine which is more fundamental to DFAI.

3. Right to explanation in law and AI: a brief

It is obvious that courts do not create evidence; they are not witnesses and are not subject to the rules of evidence. Likewise, Law and case law are not evidential. The court is, nevertheless, there to uphold the law and interpret the evidence (Marcinowski, 2021). It is, therefore, the responsibility of law enforcement or forensic practitioners to identify such evidence. The commissioner must also prove (with a persuasive explanation) the validity of the procedures and approaches used to establish the presented facts. When these approaches involve implicitly complex application (e.g., a closed-box system), the prosecution and defence also have a fundamental right: *the right to explanation* (Doshi-Velez et al., 2017).

The transparency necessary to prove the veracity of the outcome of a case may be missing without explanation in a practical legal context where “justice must not only be done but also seen to be done” (Atkinson et al., 2020). The Law discipline may have been the first to grasp the importance of explaining AI systems, and it has been the driving force in that direction in recent decades. In his insightful assessment of AI from a social science perspective, Miller (2019) listed four crucial characteristics of explanations (in AI) that he claimed the majority of AI researchers are unaware of. According to the author, explanations should be:

1. *Contrastive:* reasoning is occasionally expressed as a counterfactual hypothesis; for instance, if a predictive analysis classifies certain image as containing child exploitation content (CSEM) (Islam et al., 2019), a balanced explanation for this classification will explain what influences such inference (and why not something else). An effective approach is to investigate whether hypothetical changes to cases might have affected their conclusion as presented in HYPO (Rissland and Ashley, 1987; Ashley, 1991)
2. *Selective:* typically influenced by cognitive biases —which means that an exhaustive analysis of an event's causation is rarely presented logically. Rather, on the assumption of shared background knowledge among stakeholders —which might not always be the case —a few (selective; purportedly only persuasive) causes are chosen to explain an infinite number of causal events.
3. *Rarely Probabilistic:* while truth and probability (in ratio terms) are critical in forensic science, using *“most likely”*, for example, as a semantic explanation for a causal event may be inappropriate. Thus, utilizing explanations based on probabilities or statistical correlations as a general justification for an event's occurrence may be ineffective unless accompanied with a causal explanation for why that generalization is typical.
4. *Social:* refers to the dissemination (or transmission) of knowledge via discussion or interaction. Thus, the explanation is presented in light of the explainer's beliefs about the beliefs of the audience.

Explanation as a right can be communicated through examples (Atkinson et al., 2020), i.e., it is a common law tradition to offer contrastive precedent cases (i.e., with positive and negative examples) in order to persuade jurors or judges who may favour one side over the other. The use of hypothetical features from a prior case to explain how the outcome of a case may have been different if the features had changed is an illustration of an explanation by example (Rissland and Ashley, 1987).

4. Explanations and closed-box models: some key concerns for DFAI

Within the scope of this paper, the term “closed-box” system is used in reference with DL/DNN models (not classical ML models) used in DF. While neural networks are the focus, other shallow ML models with considerable complex algorithmic structures, such as Support Vector Machine (SVM) (Cortes and Vapnik, 1995; Noble, 2006) or Random Forests (RF) ((Ho, 1995, 1998; Breiman, 2001), are also included in the closed-box category. The issues highlighted below are just few of the factors that may have exacerbated scepticism about the use of AI in digital forensics; which is largely driven by inexplicability of AI models.

“Closed-box” refers to an incomprehensible system (or algorithmic function) to humans. We employ machines, apparently, because they possess superhuman abilities to detect patterns, discriminate, and draw conclusions. Our comprehension of these processes, however, is conditional on the model's output; which we cannot follow (Yampolski, 2020). A closed-box system does not always imply inefficiency; it, more often than not, performs as intended. The concern is that if the system claims to possess reasoning abilities and the capacity to draw conclusions comparable to those of humans in a variety of contexts, it should be able to explain how it arrived at a particular conclusion. Notably, a low-fidelity explanation of a system's decision-making process lessens both the system's and the explanation's credibility with audiences in a high-stakes domain like law. The crucial point here is that explanation is just as important as the model itself, and this is an area that DFAI desperately needs to address. Adding another layer of distrust through unconscious irrational explanations, is likely to impede the full adoption of AI in DF.

A worrying trend in the explanation of closed-box systems may be the provision of explanations primarily for correctly classified labels, which could lead to misinterpretation. An excellent use case is the description of the saliency map (Li, 2002; Underwood et al., 2006; Alqaraawi et al., 2020) in an object detection/recognition task. A saliency map is a visual representation of the area of an image that is most likely to be noticed. One of its primary goal is to communicate the importance of a given pixel in an image to human visual system, and it has been a vital component in forensic image classification methods (Thakur and Jindal, 2018; Yang et al., 2021). Often, explanations for each class on a saliency map will be identical, even if they are incorrect. A recent studies on medical imaging in (Arun et al., 2021; Saporta et al., 2021), discovered that the use of saliency to interpret DNNs did not meet key critical utility and robustness requirements. This presents a significant challenge for the numerous attempts aimed at providing explanations based on important features in input samples that may have influenced a certain prediction/classification.

Research has shown that DNN models can learn counter-intuitive solutions despite their expressiveness (Szegedy et al., 2013). DL-based classifiers have shown erroneous predictions with "high confidence" when a minor but deliberate undetectable perturbation is introduced to the examples (Goodfellow et al., 2014a). Using a specific example, Goodfellow et al. (2014a) show how adversarial cases (such as noise) can disrupt a correctly classified example with a confidence level of 57%, causing the model to falsely predict with a confidence level of 99%. Consider a counterfactual claim (such as the impact of adversarial examples) made by an opposing party showing that a forensic conclusion may be incorrect, and that decisions deduced using the same technique are unreliable. Such an example is easily persuadable to a reasonably informed AI audience, let alone those that are less informed. In spite of this, more resilient deep generative models like the Generative Adversarial Network (GAN) (Goodfellow et al., 2014b, 2020) and VAE (Kingma and Welling, 2013) have emerged as a result of this adversarial discovery. GAN's game-theoretic foundation has, however, presented unique challenges to the generative model.

Analytical inaccuracies could arise if machines augment their operating parameters in unexpected ways (Roth, 2017). This could be caused by training sets with fewer samples, which are either less representative of real-world use cases or insufficient to make inferences about future observations. Incorporating too many variables in the model runs the risk of training the model to learn illogical representations. Consider, for example, a predictive crime detection algorithm[6] installed in surveillance cameras that tracks criminal movements and alerts officers before or just when crime is committed. According to reports, by analyzing crime-related samples from surveillance camera data, the algorithm learned to recognize three handshakes in succession as likely narcotic transactions. While this reasoning appears logical, it may overlook drug-related occurrences in the real world if no such pattern exists (Roth, 2017). Exemplifying with such instances in a court case (as a reason why AI-methods should not be trusted) will only serve to increase public distrust of machine-generated evidence.

5. Explainable DFAI: the goal

The resulting value of a digital forensics investigation is the evidence, which is mined (extracted, uncovered) by a forensic expert and communicated to fact finders (e.g., legal practitioners, law enforcement, organizations, etc.). The majority of evidence is presented as facts deduced from a sequence of correlations of causal relationships, which requires decoupling intricate interrelationships between multiple heterogeneous artifacts. The court or commissioning agency establishes the evidence's weight, relevance, and substance. However, it is the role of the forensic expert to provide an understandable review of the methodology and hypothetical approaches employed to achieve the conclusion. Explaining an AI-based DF analysis may require weighting, comparing, or persuading the audience via logic-based formalization of (counter) arguments (Besnard and Hunter, 2008), or simplifying the outcome by lowering the complexities.

Given the high-stakes audiences in an evidence-oriented context, for whom presentation is crucial, an explainable DFAI (xDFAI) can be referred to *"as an AI-based digital forensics method(s) that provides explicit and intelligible (as well as assessable) rationale for its functions and the specifics of its inferential reasoning."* This definition may serve as a preliminary (tentative) formalization of explainable DFAI (xDFAI), with a more refined conceptualization envisaged as research in the domain progresses. In accordance with Clancey (1983) concept of explanation (which is adaptable to xDFAI), the goal (of xDFAI) should be to explain the following: *Why did a specific fact end up being used? When a certain fact was ignored, why did that happen? Why did the investigator not come to a different conclusion?*

The evaluation of the performance and accuracy of the technique used in DFAI has received considerable attention, but less attention has been given to the interpretability of the technique(s) used. Considering the above, it may be possible to expound on the goal of xDFAI by relating it to notions that have been widely connected with XAI in research. The following general XAI objectives are expressed here in terms of goals that an xDFAI can pursue during the examination and presentation phases of derived results:

- ***Trustworthiness:*** A model's ability to act (always) as expected (or defined) in a given context is measured by its trustworthiness, which is not a guarantee that it can be explained. Model's trust builds over time as long as it behaves consistently in accordance with the stakeholder's mental model and provides accurate and verifiable predictions (Bhatt et al., 2020). Stakeholders may overlook an unexpected failure in a trusted system because it will not have a significant impact on their confidence. It is feasible, however, to "trust but verify" in the case of DFAI —where the system is expected to perform optimally at all times due to the grave repercussions of its failure.
- ***Discovering Causality:*** Causality is the process of establishing (or inferring) causal relationships between observed data (Pearl, 2020). Thus, in order to identify these relationships, an investigator must have extensive prior knowledge (or expertise) in the field and must be aware that the existence of certain relationships between data does not imply causality.

A robust xDFAI should be capable of providing intuitive evidence and explanations for causal relationships within observable artifacts, or assist in the validating the output of a causality-inference method.

- ***Reproducibility:*** The training and testing (as well as validation) phases in a model can be validated and their applicability verified. Thus, the purpose of explainability in this context should be to elucidate the model's functionality in order to ease comprehension of its constraints (or boundaries), and the seamless transfer of knowledge for reproduction (Arrieta et al., 2020). Lack of explanation could lead to erroneous assumptions about the model (Kim et al., 2017).

Indeed, in ML research, the explanations presented in the literature have influenced the improvements on state-of-the-art. Consequently, confidence in DFAI models is likely to increase when the functional parameters are explicitly elucidated and its methods widely extensively reproduced.

- ***Informativeness:*** The output of a DFAI model is almost exclusively numerical (probabilistic of some sort). It will require time and effort to draw a connection between these values and the investigative problem for which a evidence is

[6] See https://www.govtech.com/public-safety/smart-cameras-aim-to-stop-crimes-before-they-occur.html.

sought. It is critical that xDFAI describes how these values are represented and how they assist investigators in deducing the facts. Both explanation and information are complementary; neither is possible without the other. To some extent, once a model has proved its capacity to predict reliably across a range of scenarios, its credibility will be determined by the amount of information it can convey about its inferential processes and the accuracy of its output.

- ***Confidence:*** In a stable system, this is a quality that is practically synonymous with trust and believe. When reliability is demanded, confidence is relative; it is tangible (Arrieta et al., 2020). Confidence is expressible; could be conveyed by the person presenting the facts, or by the one receiving it. As with trustworthiness, confidence in a DFAI model might not easily lend itself to the notion of explainability because it is earned via operational and result consistency —not necessarily through explicitness of its operational parameters. Nonetheless, an xDFAI can be critical in providing information on the level of confidence for each modular component of the system. This way, each component of the decision-making process can be evaluated and appropriate confidence scales assigned.
- ***Algorithmic Fairness:*** In relation to the system's specified objectives, fairness could be seen as one of the aims of explainability. Fairness is considered in the legal domain in terms of adherence to ethical principles, the right to be informed, and the right to contest decisions (Goodman and Flaxman, 2016; Wachter et al., 2017). To achieve algorithmic fairness, it is necessary to draw a clear picture of the relationship between hypothetical components that may have influenced a certain decision. This includes taking into account counterfactual components. It is possible that an investigator disregard facts that contradicts her own perception. As a result, erroneous inferences may be drawn. If this (erroneous) conclusion is reached based on algorithmic analysis, it risks undermining trust in machine-generated outcomes; this should be avoided.
- ***Availability:*** This relates to accessibility and comprises examining explainability as a strategy to engage end users in the process of enhancing specific AI models (Miller et al., 2017). This means that open-sourcing and peer-reviewing a DFAI algorithm should ideally aid technical users in grasping the technique, while xDFAI will almost likely assist non-technical users in interacting with the algorithm. Thus, if a forensic expert is required to report (or testify) in a legal proceeding regarding an algorithm's decision, an easily available open-sourced and/or peer-reviewed procedure is likely to be understood and accepted.

6. Explainable DFAI: the methods

This section addresses several ways for explaining AI models. The objective is to expound on XAI and, when appropriate, establish relevant connections with xDFAI.

It has been discussed whether to oversimplify AI models in order to make them more intelligible at the expense of performance and accuracy (Shalaginov, 2017). Given that interpretability and model performance are (to a significant extent) the fundamental aims of XAI, a post-hoc explanation technique has grown in popularity. Conversely, the intrinsic approaches (not discussed in detail in this paper) that are based on simpler, self-explainable models (e.g. Decision Trees, rule-based, linear models, etc.) are possible. Fig. 1 is an illustration of the xDFAI structural model.

6.1. Post-hoc explainability approaches

To throw light on certain model, post-hoc explanations can make clearer its salient features (Ribiero et al., 2016; Lundberg and Lee, 2017; Davis et al., 2020), training points (Koh and Liang, 2017; Yeh et al., 2018), counterfactual reasoning (Wachter et al., 2018)), or decision boundaries (Ribiero et al., 2016; Lundberg and Lee, 2017) (Bhatt et al., 2020). Post-hoc techniques aim to improve the interpretability of closed-box models by a variety of means, including explanations by: *model simplification, visualization, localization, feature importance, example, and text.* This paper examines post-hoc explainability in two unique contexts: model-agnostic and model-specific.

The model-agnostic explainability, on the one hand, is built into the model's internal mechanism in a manner independent of the model's internal structure and it is implemented after the model has been trained (Molnar, 2019). Using this method, it is possible to learn more about how a model predicts outcomes (Arrieta et al., 2020). On the other hand, model-specific explainability methods are restricted, and only applicable to specific algorithm types. All intrinsic approaches are, in fact, model-specific. In this paper, model-specific methods are described from the perspective of their use in DNNs —because of their opaqueness which this work focuses on.

Brief description of post-hoc explainability methods are presented below. Additionally, within the scope of this work and in line with the context of opaque models, the emphasis is primarily on methods that are applicable to deep-layered neural networks, however, methods for shallow models (e.g., SVM, RF, etc.) are mentioned in few instances. It is important to emphasize that the models discussed here are far from exhaustive; they represent only a fraction, and the choice of selection is based on their potential suitability for DFAI. Table 1 and Table 2 presents an overview of both model-agnostic and model-specific post-hoc explainability methods and their potential suitability for DFAI tasks.

6.1.1. Explanation by model simplification

The broadest of the model-agnostic post-hoc explanations appears to be model simplification. While they are predominantly focused on rule extraction techniques, Bastani et al. (2018) presented a different extraction approach based on approximating a transparent model to a complex one. Methods, such as G-REX (Johansson et al., 2004a,b; Konig et al., 2008) and CNF (Conjunctive Normal Form) or DNF (Disjunctive Normal Form) (Su et al., 2016) based on this approach seeks to simplify interpretability by extracting information in form of rules.

6.1.2. Explanation by feature importance

By quantifying and analyzing the influence, relevance, and significance of each training variable on the model's prediction, this approach elucidates the operationality of a closed-box model. The SHAP (SHapley Additive exPlanation) SHAP (Lundberg and Lee, 2017) framework, and an interesting approach for explainable image analysis based on saliency detection method proposed in (Dabowski and Gal, 2017), offers a significant contribution to feature importance. Additionally, the Automatic STRuctured IDentification (ASTRID) (Henelius and Ukkonen, 2017; Henelius et al., 2014) is a useful tool for determining feature importance in a predictive model. However, several alternative approaches have been proposed that go beyond the influence measure. The approaches highlighted here provides highly valuable techniques for

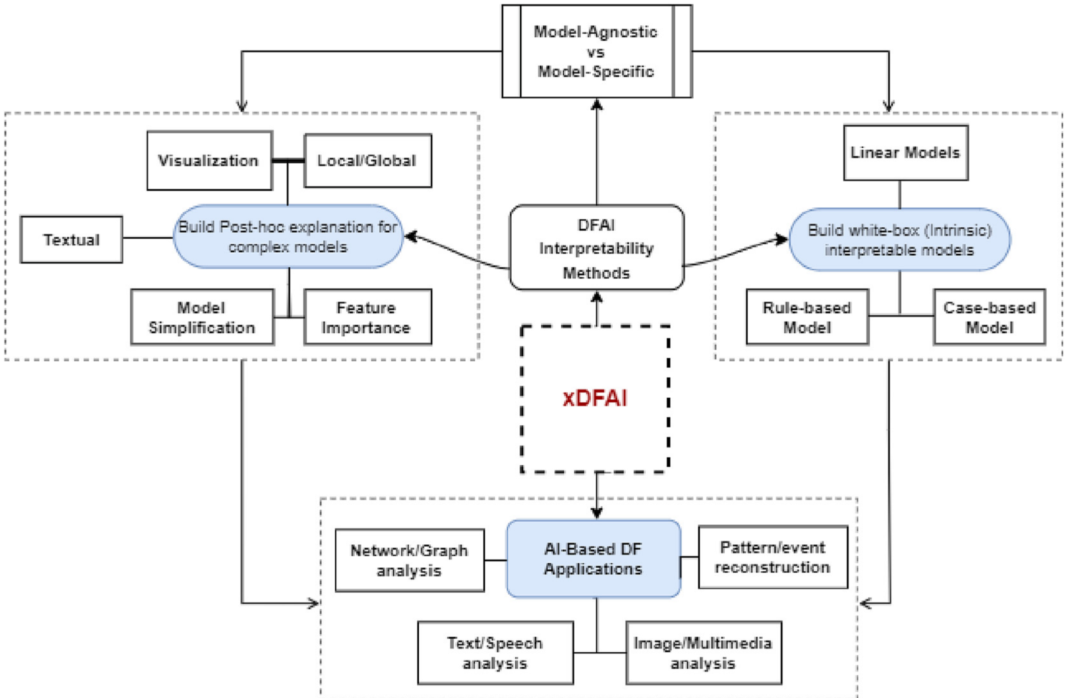

Fig. 1. Mind map representing an illustration of the explainable digital forensic AI (xDFAI) Model.

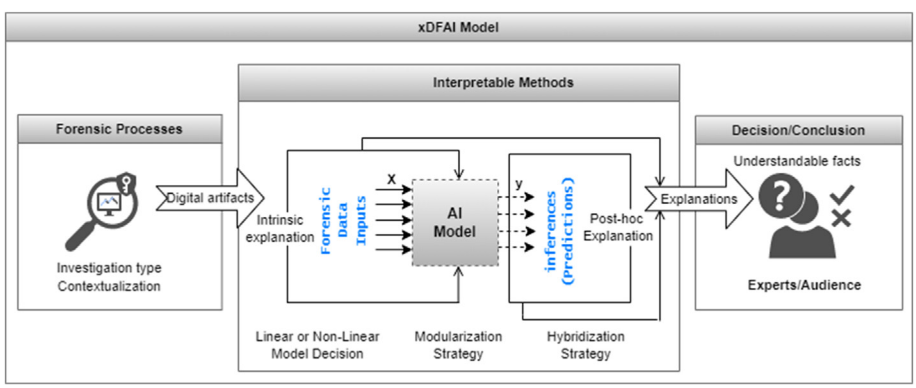

Fig. 2. A typical structure of an interpretable DFAI model.

Table 1
An overview of some model-agnostic explainability methods, proposed tools, and their potential applications to digital forensic.

Explainability Techniques	Post-hoc Explanation	Tools	Potential Applicability to DF
Model-Agnostic	Model simplification	G-REX, CNF or DNF	Pattern recognition, digital file forensic analysis, text analysis etc.
	feature importance	SHAP, ASTRID, Influence function, Saliency detection (Koh and Liang, 2017; Dabowski and Gal, 2017)	Image forensics, object classification, predictive analysis, etc.
	Visualization	SA & Global SA, ICE	Pattern recognition, object identification/classification, document classification, etc.
	Local	LIME, Fairness (Dwork et al., 2012), L2X (Chen et al., 2018), AIX360 (Dhurandhar et al., 2018)	Object classification, predictive analysis, multimedia forensics, etc.
	Text	TextAttack (Gao et al., 2018), HotFlip (Ebrahimi et al., 2018)	Spam message detection, e-mail forensics, attribution, malware detection, etc.

xDFAI, which can be explored further in future research.

6.1.3. Explanation by visualization

Visual explanation is also a strategy for achieving model-agnostic explanations, however it is highly effective, and mostly common in model-specific approaches; especially with DNNs. In a typical model-agnostic settings, developing visualizations based just on the inputs and outputs of an opaque model may be a

Table 2
An overview of some model-specific explainability techniques based on DNNs, proposed/developed tools, and their potential application to digital forensics.

Explainability Techniques		Post-hoc Explanation	Tools	Potential Applicability to DF
Model-Specific	MLNN	Model simplification	DeepRED	Forensic image classification, object identification/detection, pattern recognition, CSEM analysis, etc.
		feature importance	Deep Taylor, DeepLift, Deconvnet	
		Visualization	TreeView	
	CNN	Visualization	LRP, DGN, Grad-CAM, CNN + CRF + bi-LSTM (Ma and Hovy, 2016)	Forensic image/video reconstruction, forensic data visualization, object identification, source identification, deep fakes, image recognition, etc.
		Text	CNN + RNN (Xu et al., 2015)	
	RNN	feature importance	RETAIN	Speech recognition, authorship attribution, determination of intent, forensics linguistics, timeline/event reconstruction, malware detection, email forensics, e-Discovery, IoT forensics, Network intrusion detection, etc.
		Visualization	Finite n-gram horizon + RNN	
		Local	RNN + Hidden Markov Model (HMM)	

difficult task (Arrieta et al., 2020). A frequently utilized technique in this approach is to provide explanations through the use of feature importance techniques. Notable methods for visualization of shallow ML models (e.g., SVM, RF, etc.) are proposed in (Cortez and Emrechts, 2011, 2013) based on Sensitive Analysis (SA), and Individual Conditional Expectation (ICE) (Goldstein et al., 2013) for estimating any supervised learning techniques. While feature importance is beneficial for xDFAI, visualization approaches provide an innovative way to physically observe the interaction of influential variables during the process. Although the approach is quite complex, it offers a promising research direction for xDFAI.

6.1.4. *Local explanation*

Considering that DL models have a high degree of dimensionality and curvature, the concept of local explanation stems from the fact that insight-generating interpretable methods can be applied to a tiny region with detectable changes in individual or grouped features. Using the network's feature space to represent each case (data point) or its neighbors, local explanation provides a semantic explanation for specific cases (Leslie, 2019). However, a *global explanation* entails capturing the internal logic and function of each prediction or classification made by an opaque model as a whole (rather than a tiny region) (Leslie, 2019). The technique, known as LIME (Local Interpretable Model-Agnostic Explanations) (Ribiero et al., 2016) is an example of a model-agnostic approach designed to simplify explanations, which explains model predictions by learning interpretable models locally and modeling them as a submodular optimization problem.

6.1.5. *Text explanation*

Adding explanations in plain natural language to closed-box models is an approach that has not been well discussed in the literature. Each decision-making component of a model can be described using text. In some cases, text explanations are incorporated in a rule-based (or if … then) style, in which all decision-making components are described semantically explained. This approach, when combined with other approaches (e.g., feature importance and visualization), can be quite beneficial for xDFAI.

6.2. *Explainability methods to explain deep learning models*

This section briefly discuss the explainability of DNNs. Three distinct neural network architectures are considered: multi-layered networks (MLNNs), convolutional neural networks (CNNs) (Ośhea and Nash, 2015; Albawi et al., 2017), and recurrent neural networks (RNNs) (Mikolov et al., 2010). The selection is based on their utility/applicability to DFAI. However, in terms of depth and breadth, the descriptions offered here are largely limited, readers are urged to check (Linardatos et al., 2021; Arrieta et al., 2020) for a full overview of explainable approaches.

MLNNs are a sort of closed-box, yet robust AI model that excels at inferring intricate relationships between data variables, and in most cases, are unable to justify their underlying assumptions. Three fundamental explainable methodologies are utilized to explain multi-layer neural networks: model simplification through rule extraction from hidden layer of a neural network (DeepRED) (Zilke et al., 2016; Sato and Tsukimoto, 2001); feature importance of contributing elements with models such as Deep Taylor (Montavon et al., 2017) and DeepLift (Shrikumar et al., 2017); and visualization for which TreeView (Thiagarajan et al., 2016) was proposed. Because DeepLift and deep Taylor are exemplified with image classification, they could be an excellent xDFAI options for forensic image analysis as well as pattern recognition-based investigations.

CNNs (Ośhea and Nash, 2015; Albawi et al., 2017) structure reflects DNN's extremely complex internal cores. They lay the groundwork for computer vision's unique underpinnings —from object identification and image classification to instance segmentation (Arrieta et al., 2020). Because CNN's representations are visual, they connect well with the human thought pattern, making them slightly explainable. An approach for explaining CNN functionality is to either map the output back to the input in order to ascertain which input data were discriminative of the output, or to make interpretations based on how the layers see the external world. A common feature importance and local explanation method is Deconvnet (Zeiler et al., 2010, 2011; Zeiler and Fergus, 2013) that repeatedly occludes sensitive region of an image during training to determine which portion produces desired impact. Another approach based on feature importance and localization is the Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017). Layer-wise Relevance Propagation (LRP) (Bach et al., 2015) proposes a method that visualizes relevant elements that contributes to prediction. Other methods (Dong et al., 2017; Xu et al., 2015) combines CNN models and RNN for the purpose of describing visual material via textual explanations. Perhaps an excellent and easily interpretable approach is the deep generator network (DGN) (Nguyen et al., 2016), which not only generates an incredibly realistic synthetic image, but also reveals the features learned by each neuron. Given that certain DF analysis will require object identification, the DGN approach appears to possess both quality and suitable characteristics for the development of xDFAI.

RNNs are one of the most important techniques for DFAI because they are capable of solving prediction problems using sequential data —which is critical for forensic event reconstruction (Solanke et al., 2021). RNNs take pride in their capacity to retain information about data's time-dependent relationships. There have been two approaches to explaining RNN models: 1) through feature importance techniques that seek to understand what the model has learned over time; and 2) by providing insights into (or explanations of) the model's decision-making process through modification of its architecture (local explanations) (Arrieta et al., 2020). Numerous proposals are offered in this respect, which may spark the interest of DFAI professionals. With RNN, some explanation approaches (Donadello et al., 2017; Donadello, 2018; Garcez et al., 2019) have demonstrated the possibility of merging probabilistic and logical reasoning (Manhaeve et al., 2021) (based on background knowledge) in a symbolic/sub-symbolic (Haugeland, 1989; Ilkou and Koutraki, 2020) fashion. Some other approaches include visualization approach based on finite horizon n-gram models (Karpathy et al., 2016) to study predictions, combination of RNN with a simple and transparent hidden Markov Model (HMM) (Krakovna and Doshi-Velez, 2016) to interpret speech recognition representations, and the RETAIN (Reverse Time Attention) model introduced in (Choi et al., 2016) for detecting influential past visit patterns and significant variables within the patterns. This technique could be useful, for example, in performing forensic analysis on users' log history (e.g., internet browsing history) during a CSEM investigation.

In contrast to the preceding methods, which are either model-agnostic or model-specific, a novel technique dubbed Contextual Importance and Utility (CIU) is proposed (Framling, 2020, 2022; Anjomshoae et al., 2019). It is based on Contextual Importance/Influence (CI) and Contextual Utility (CU) theory. CIU appears promising as it is applicable to both linear and non-linear models and may be represented visually or in natural language. Additionally, feature representations can be read and validated directly from input—output graphs. Although the CIU approach is just developing, its features indicate that it has the potential to considerably aid in xDFAI.

7. Interpretability in DFAI: the need

For a system to be trusted, it must go beyond a simple accuracy evaluation. Sometimes, accuracy does not always reflect the real world use case. Therefore, a critical component for determining whether the correctness of a system's outcome is the interpretability of its decisions and comprehensibility of its features. A model's domain-specific constraints may make it difficult to incorporate interpretable components into a closed-box models. Because constrained problems are inherently more difficult to solve, when AI models are applied in DF investigations, interpretability practically translates to a set of application-specific constraints. Hence, domain expertise will be needed to implement interpretable features in the model. In contrast to explainability, which is mostly concerned with providing post-hoc reasoning for predictions, interpretability provides an answer not only to the question of what was predicted (which is only a partial solution to the problem), but also to the question of why such predictions were made (or what caused them). By incorporating interpretable features into DFAI, it is possible to harmonize and update gaps in domain knowledge, as by attempting to answer why a particular decision was made, new dimensions to the problem or solution can be uncovered, and methods for debugging or auditing can be established. A model that can be interpreted can also help determine the fundamental cause of an error and recommend possible solutions. Interpretable models ensure simulatability (the reasoning in the model is verifiable and reproducible), decomposability (the sub-component interpretation is possible), and algorithmic transparency when opposing parties in an inquisitorial tradition request access to the tool used to infer facts. Fig. 2 represents a structural model of an interpretable model. While building interpretable models can be time and resource intensive, it is less expensive than the expense of creating a flawed model (Rudin, 2019) that could lead to the eventual exculpation or incrimination of the wrong entity for high-stakes decisions such as those involving digital evidence. There is evidence to suggest that it would be desirable to dedicate additional efforts and cost on developing a high-quality interpretable model, even as timeliness is still a challenge in DF.

8. Interpretable DFAI model: recommendations for mitigating distrust

The following paragraphs contain a series of recommendations that may be essential for achieving robust interpretability in DFAI. They are adapted in part from the guidelines provided in (Leslie, 2019).

It is critical to contextualize the scenario (e.g. civil or criminal case), potential impact, and accessible AI tools for analysis prior to integrating AI models in DF, while also considering the investigation's interpretability requirements. There appears to be a significant distinction (in terms of techniques and interpretation requirements) between analyzing e-mails for suspicious deletions intended to conceal incriminating activities, and determining responsibility in e-contract agreements between two or more parties concluded via e-mails. This contextual awareness helps to paint a more complete picture of the stakes involve and the scope of the interpretability needs. Another consideration to make before deployment is whether to use pre-existing AI algorithms or to create new ones. In any case, utilizing existing algorithms may require a detailed examination or evaluation of their functionality, expressiveness, complexity, performance, and interpretability. Alternatively, a custom algorithm could be considered that addresses both the aforementioned components and the investigative task.

It is clear that the DF domain and its components are quite sensitive, as they are task-critical and requires transparency. So when DFAI is necessary, less complex, non-opaque evidence mining techniques —generally, intrinsic approach (such as decision tree, linear/logistic regression, case-based reasoning, rule-based list, etc.) can be considered. Simple interpretable models are usually preferred when forensic data is well-structured, sufficient domain knowledge with meaningful representations is present, or if computational resources are constrained. This is also highlighted in (Rudin, 2019). It is reasonable to avoid the circumstance in which “everything becomes a nail when there is a hammer.” The choice of DNN should be influenced by the nature of task, and unless inefficiency with native ML is observed, use of deep learners to improve performance and accuracy may not be less preferable.

Typical linear models may be unable to handle the majority of DF investigations. Cases such as image classification, speech recognition/audio analysis, or object identification in video footage, or anomaly detection in unstructured data typifies the tasks in DF investigation. Given that only non-linear DL models can be viable for these purposes, interpretable models such as those described in section 6 may be considered. Otherwise, a custom model that: fits the specifics of the case; evaluates the impact of decision; and addresses audience needs can be built and deployed. Nonetheless, stakeholders should be satisfied with the semantic explanations provided by supplemental interpretability tools, as well as with how they are implemented in terms of both interpretability and algorithmic approach.

Interpretable methods should be evaluated on their ability to articulate the logical explanation for their decisions and behaviours in a given scenario, as well as their users' ability to account for the generated output in a decent, coherent, and reasonable manner. Prior to selecting a method, a few critical questions should be asked: 1) what is the affected audience's mental capacity for understanding the outcome?; 2) will the method assist decision-makers (e.g., judges, organizations, etc.) in making informed/justifiable evidence-based judgments?; and 3) Will the method generate counterfactual, misleading, or confusing explanations?

The modularization of design is a vital topic to emphasize. Without a doubt, digital investigation comprises the examination of digital artifacts that may be heterogeneous and unstructured. Before data can be imbued into a DL model, it must be pre-processed. Ordinarily, the pre-processing stage does not require AI techniques, and when it does, like with NLP (Manning and Schutze, 1999) or probabilistic language models (Bengio et al., 2003), the procedures are fairly interpretable. Additionally, in a communication-related investigation, it may be necessary to construct a graph of subjects' relationships; this is not AI, and the construction can be easily comprehended. Modularization enables the development of structured applications where AI is responsible for only a portion of the investigative tasks and not the full process (Asatiani et al., 2020). As a result, it can ensure proper control over functions, reduce the investigator's interpretability burden, and enhance the audience's understanding and trust.To leverage on the benefits of cloud computing, Digital Forensics as a Service (DFaaS) (Van Baar et al., 2014; van Beek et al., 2015; Du et al., 2017; van Beek, 2020) is projected to impact the future of forensics. In such situation, DFAI as a Service may involve online learning (OL), which is when a model learns to adapt with changes in the environment and keeps updating its best predictor. OL can be useful for reconstructing events, but it can be hard to keep track of and explain variable interactions in the feature space over time. OL issues may involve the inability to control the working parameters of the model, which could be a problem in high-stakes domains (Asatiani et al., 2020). The same could be said for transfer learning (Zhan et al., 2017) (especially when offered as a service), which entails applying previously learned knowledge to a different but related problem. They could help DF in terms of sample efficiency (Karimapanal and Bouffanais, 2018), less time spent investigating, and less false positives and negatives. However, they provide less information about how the models were trained or how trustworthy the platforms that host them are (Aditya et al., 2017). The number of transfer learning methods that can be explained is still very limited, and their use in DFAI should be done with caution.

Legal experts are commonly familiar with symbolic algorithms (e.g., expert systems, case-based reasoning, etc.) because they are used in legal rule mining and in the modelling of philosophical norms. It will also be easy for laypeople to understand the logical foundation on which they are built. DFAI methods that make use of symbolic algorithms should be able to easily explain their outcomes in this scenario. However, symbolic algorithms suffer from a number of shortcomings (Faye, 2010; Sally and Terence, 1999) that render them inefficient for the majority of forensic investigations. Researchers have proposed a way to hybridize sub-symbolic (like NN models) and symbolic methods (Zeleznikow and Stranieri, 2017; Mao et al., 2018) that takes advantage of the former's robust unsupervised capacity to learn from complex data and the latter's ease of explanation to produce an explainable model. Neurosymbolic AI (Garcez and Lamb, 2020) is one of such methods. While these systems are still in their infancy, hybrid techniques are likely to give the necessary level of interpretation for predictive DF analysis. Furthermore, an equally helpful method is to incorporate a “human-in-the-loop” or a “man-machine” approach (Nguyen and Choo, 2021) with the hybrid technique. That way, automated decisions can be verified by the gatekeeper (Desai and Kroll, 2017) at different levels and appropriate validations performed prior to reaching a final conclusion.

Generative models (e.g., GAN, VAE, etc.) may be able to help solve interpretability problems in some way. In extension, they can be extremely useful for DFAI when it comes to certain tasks given their robustness in terms of performance and accuracy. With the right visualization tool, the latent features (embedding), which are direct low-dimensional representations of the input data, can be examined and tracked during training to identify which features play a role in a prediction. In this case, providing interpretations for such glass-box operations should be straightforward. Therefore, the use of generative models for complex DF analysis (such as pattern/speech recognition, object classification, event reconstruction, etc) is highly recommended.

9. Conclusion

In this paper, the human–machine relationships involved in interpreting machine-generated output were analysed, as well as the interchangeable usage of terms such as explainability, interpretability, and understandability. Brief examination of the relationship between AI and law was presented, with an emphasis on the *'right to explanation'*. By redefining explainability in the context of AI-based digital forensics (DFAI) analysis, this paper explores the goal of explainability and the methods used to achieve it. Additionally, an overview of the most frequently utilized explanation methods was presented, along with their potential applications in DF. A tentative definition of explainable DFAI was presented, while also presenting an argument for interpretable DFAI as against explainable DFAI. The author expressed an utterly (trivial) personal opinion aiming to de-escalate the controversy over AI applications in DF and their inscrutability. Finally, certain recommendations (mainly based on the construction of interpretable models) were offered that may be critical for mitigating distrust in AI-based digital evidence mining techniques. Additionally, an appendix discusses a brief personal opinion.

Future research in this area will seek to expand the xDFAI use case by evaluating the applicability of various explanation approaches on a real-world DF problem.

10. Discussion

According to the reviewed literatures on XAI and interpretable AI, it is apparent that several efforts have been undertaken to deconstruct, demystify, and improve the transparency of closed-box AI models. Thus, it is likely self-evident that AI researchers now have a substantial grasp of the fundamental underpinnings of AI algorithms, which explains why there have been spikes in research output bringing novel approaches or improvement on existing state-of-the-arts. However, the bulk of non-technical users of AI systems or those who are impacted by AI decisions appear to struggle to comprehend the subtleties of AI systems. In a slightly trivial opinion, while algorithmic biases have been reported and confirmed in some AI-generated decisions —which are more related to training data than to data processing technicalities (and, of course, deserve the attention they are receiving) —one can assume that the distrust is “partly (arguably)” influenced and amplified by the discovery of a new research gold mine. While advocacy for transparent and explainability (led primarily by the Social Science discipline) has aided XAI's penetration and understanding across disciplines, it is hoped that, from the socio-economic sides of AI, we will continue to push for a more standardized and responsible approach to designing AI-powered systems, alongside calls for regulations or understandability. One of these standards could be to make proprietary AI-based technologies that affect the public (of which DFAI is one) more programmatically transparent (which, of course, has been vigorously pushed in the EU), or to mandate that no closed-box should be used for certain high-stakes decisions when an interpretable model with the same level of performance exists (Rudin, 2019). This, however, may be difficult, given current legislation safeguarding trade secrets and the recent advancements enabled by AI that were previously deemed virtually unthinkable. Nonetheless, science advances at a frenetic pace, reacting to (internal or external) stimuli along the way. What is potentially alarming is an attempt to over-simplify science for the sake of comprehension. This is why explanations by simplification should be utilized with caution. *"Some things in life are too complicated to explain … Not just to explain to others but to explain to yourself. Force yourself to try to explain it and you create lies.*[7]*"* While there is a substantial difference between grasping and nearly comprehending something, providing an accurate explanation may result in decreased comprehensibility. Conversely, providing a more comprehensible explanation may result in decreased accuracy (Yampolski, 2020). As a result, it may appear unreasonable or counter-intuitive to assume that technical explanations offered post-hoc or modeled using the internals of AI models will be comprehended by the intended audience even after simplification. Perhaps at that point, a comprehensibility evaluation will be required. Consequently, an explanation for an AI-enabled conclusion should justify not just the mathematical foundations, technical underpinnings, and societal context, but also the human impact.

Lastly, it is worth emphasizing, however, that the discussion here is a trivially expressed opinion of the author; based entirely on personal social observations. They are merely offered to lessen the escalation of debate about whether AI (with its perceived opaqueness) should be applied to DF investigation. According to a famous Albert Einstein quotation, which reads as follows:

"It would be possible to describe everything scientifically, but it would make no sense. It would be a description without meaning — as if you described a Beethoven symphony as a variation of wave pressure."

References

Aditya, K., Grzonkowski, S., Lekhan, N., 2017. Enabling trust in deep learning models: a digital forensics case study. In: In Proc. of the 17th IEEE Intl. Conf. on Trust, Security and Privacy in Computing and Communications, pp. 1250–1255.

Albawi, S., Mohammed, T.A., Al-Zawi, S., 2017. Understanding of a convolutional neural network. In: In 2017 Intl. Conf. on Engineering and Technology (ICET), pp. 1–6.

Alqaraawi, A., Schuessler, M., Weib, P., Constanza, E., Berthouze, N., 2020. Evaluating saliency map explanations for convolutional neural networks: a user study, in: in Proc. of the 25th Intl. Conf. on Intelligent User Interfaces 275–285.

Anjomshoae, S., Framling, K., Najjar, A., 2019. Explanation of black-box model predictions by contextual importance and utility. Int. Workshop Explain. Transparent Autonomous Agents Multi-Agent Syst. 95–109.

Arrieta, A.B., Diaz-Rodiguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., benjamins, R., Herrera, F., 2020. Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges towards responsible ai. Inf. Fusion 58, 82–115.

Arun, N., Nathan, G., Praveer, S., Ken, C., Mehak, A., Bryan, C., Kathrina, H., Sharut, G., Jay, P., Mishka, G., Julius, A., Matthew, D.L., Jayashree, K., 2021. Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artif. Intell. 3, e200267.

Asatiani, A., Malo, P., Nagbol, P.R., Penttinen, E., Rinta-Kahila, T., 2020. Challenges of explaining the behaviour of black-box ai systems. MIS Q. Exec. 19. Article 7.

Ashley, K., 1991. Reasoning with cases and hypotheticals in hypo. Intl. J. Man-Machine Studies 34, 753–796.

Atkinson, K., Bench-Capon, T., Bollegala, D., 2020. Explanation in ai and law: past, present and future. Artif. Intell. 267, 103387.

Aziz, S., Dowling, M., 2019. Machine learning and ai for risk management. Disrupting Finance: FinTech and Strategy in the 21st Century 33–50.

Van Baar, R.B., van Beek, H.M., Van Ejik, E.J., 2014. Digital forensics as a service: a game changer. Digit. Invest. 11, 254–262.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Muler, K., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One, e0130140.

Baggili, I., Behzadan, M., 2020. Founding the domain of ai forensics. safeAI@ AAAI.

Bastani, O., Kim, C., Bastani, H., 2018. Interpretability via Model Extraction arXiv preprint arXiv:1706.09773.

van Beek, H.M., van, B.J., van Ejik, E.J., Schramp, R., Ugen, M., 2020. Digital forensics as a service: stepping up the game. Forensic Sci. Int.: Digit. Invest. 35, 301021.

van Beek, H.M., Van Ejik, E.J., Van Baar, R.B., Ugen, M., Bodde, J.N., Siemelink, A.J., 2015. Digital forensics as a service: game on. Digit. Invest. 15, 20–38.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A neural probabilistic language model. J. Mach. Learn. Res. 1137–1155.

Benjamins, R., Barbado, A., Sierra, D., 2019. Responsible Ai by Design in Practice arXIv preprint arXiv:1909.12838.

Besnard, P., Hunter, A., 2008. Elements of Argumentation. MIT Press.

Bhatt, U., Andrus, M., Weller, A., Xiang, A., 2020. Machine Learning Explainability for External Stakeholders arXiv preprint arXiv:2007.05408.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Carr, N., 2014. The Glass Cage: Automation and us. WW Norton & co.

Chen, J., Song, L., Wainwright, M., Jordan, M., 2018. An information-theoretic perspective on model interpretation, in: in Proc. of the 35th Intl. Conf. Mach. Learning (ICML) 882–891.

Choi, E., Bahadori, M.T., Kulas, J.A., Schuetz, A., Stewart, W.F., Sun, J., 2016. Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. In: 2016 NIPS, pp. 3512–3520.

Clancey, W.J., 1983. The epistemology of a rule-based expert system - a framework for explanation. Artif. Intell. 20, 215–251.

Cortes, C., Vapnik, V.N., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Cortez, P., Emrechts, M.J., 2011. Opening bllack box data mining models using sensitivity analysis. In: 2021 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 341–348.

Cortez, P., Emrechts, M.J., 2013. Using sensitivity analysis and visualization techniques to open black box data mining models. Inf. Sci. 225, 1–17.

Cowan, N., 2010. The magical mystery four: how is working memory capacity limited, and why? Curr. Dir. Psychol. Sci. 19, 51–57.

Coyle, D., Weller, A., 2020. Explaining machine learning reveals policy challenges. Science 368, 1433–1434.

Crave, M.W., 1996. Extracting Comprehensible Models from Trained Neural Networks. Ph.D. Dissertation. The University of Wisconsin-Madison.

Dabowski, P., Gal, Y., 2017. Real time image saliency for black box classifiers. In: In Proc. of the 31st Intl. Conf. on Neural Information Processing Systems, pp. 6970–6979.

Davis, B., Bhatt, U., Bhardwaj, K., Merculescu, R., Moura, J.M., 2020. On network science and mutual information for explainable deep neural networks, in: 2020 IEEE Intl. Conf. on Acoustics. Speech Signal Process. 8399–8403.

Desai, D.R., Kroll, J.A., 2017. Trust but verify: a guide to algorithms and the law. Harv. J. Law Technol. 31.

Dhurandhar, A., Chen, P., Luss, R., Tu, C., Ting, P., Shanmugam, K., Das, P., 2018. Explanations based on the missing: towards contrastive explanations with pertinent negatives. In: In Proc. of the Advances in Neural Information Processing Systems, pp. 592–603.

Donadello, I., 2018. Semantic Image Representation—Integration of Numerical Data and Logical Knowledge for Cognitive Vision. Doctoral Thesis. University of Trento.

Donadello, I., Serafini, L., Garcez, A.D., 2017. Logic tensor networks for semantic image representation. In: 2017 IJCAI, pp. 1596–1602.

Dong, Y., Su, H., Zhu, J., Zhang, B., 2017. Improving interpretability of deep neural networks with semantic information. In: 2017 IEEE Intl. Conf. on Computer Vision and Pattern Recognition, pp. 975–983.

Doran, D., Schulz, S., Besold, T.R., 2017. What does explainable ai really mean? a new conceptualization of perspective. In: 16th Intl. Conf. of the Italian Assoc. for AI.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., Weller, A., Wood, A., 2017. Accountability of Ai under the Law: the Role of Explanation arXiv preprint arXiv:1711.01134.

Dror, i.E., Mnookin, J.L., 2010. The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensic science. Law,. Probability Risk 9, 47–67.

Du, X., Le-Khac, N.A., Scanlon, M., 2017. Evaluation of Digital Forensic Process Models with Respect to Digital Forensics as a Service arXiv preprint arXiv: 1708.01730.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R., 2012. Fairness through awareness. In: In 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226.

Ebrahimi, J., Rao, A., Lowd, D., Dou, D., 2018. Hotflip: white-box example for text classification. In: In Proc. of the Association for Computational Linguistics, pp. 31–36.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Ashkam, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C.O., Raine, R., Hughes, J., Sim, D.A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P.T., Suleyman, M., Cornebise, J., Keane, P.A., Ronneberger, O., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. 24, 1342–1350.

Faye, M., 2010. The use of artificial intelligence in digital forensics: an introduction. Dig. Evid. Electro. Signature Law Rev. 7, 35–41.

Framling, K., 2020. Explainable Ai without Interpretable Model arXiv preprint arXiv: 2009.13996v1.

[7] Quote of Haruki Murakami - https://bukrate.com/quote/544024.

Framling, K., 2022. Contextual Importance and Utility: a Theoretical Foundation arXiv preprint arXiv:2202.07292v1.
Gao, J., Lanchantin, J., Soffa, M., Qi, Y., 2018. Black-box generation of adversarial text sequence to evade deep learning classifiers,. In: In Proc. of IEEE Security and Privacy Workshop (SPW), pp. 50–56.
Garcez, A.D., Lamb, L.C., 2020. Neurosymbolic Ai: the 3rd Wave arXiv preprint arXiv:2012.05876.
Garcez, A., Gori, M., Lamb, L., Serafini, L., Spranger, M., Tran, S.N., 2019. Neural-symbolic Computing: an Effective Methodology for Principled Integration of Machine Learning and Reasoning arXiv preprint arXiv:1905.06088.
Gleicher, M., 2016. A framework for considering comprehensibility in modelling. Big Data 4, 75–88.
Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2013. Peeking inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation arXiv preprint arXiv:1309.6392.
Goodfellow, I., Shlens, J., Szegedy, C., 2014a. Explaining and Harnessing Adversarial Examples. CarXiv preprint arXiv:1412.6572.
Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-farle, D., Ozair, S., Courville, A., Bengio, Y., 2014b. Generative Adversarial Networks arXiv preprint arXiv:1406.2661.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63, 139–144.
Goodison, S., Robert, D., Brian, J., 2015. Digital evidence and the u.s. criminal justice system: identifying technology and other needs to more effectively acquire and utilize digital evidence. URL. https://www.rand.org/pubs/research_reports/RR890.html.
Goodman, B., Flaxman, S., 2016. European union Regulations on Algorithmic Decision-Making and a 'Right to Explanation' arXiv preprint arXiv:1606.088138.
Gross-Brown, R., Ficek, M., Agundez, J., Dressler, P., Laoutaris, N., 2015. Data transparency lab kick off workshop (dtl 2014) report. In: ACM SIGCOMM Computer Comm. Review, pp. 44–48.
Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., 2019. A survey of methods for explaining black box models. ACM Comput. Surv. 51, 42.
Gunning, D., 2019. Dapra's explainable artificial intelligence (xai) program. AI Mag. 40.
Hall, P., 2018. On the Art and Science of Machine Learning Explanations arXiv preprint arXiv:1810.02909.
Hall, S.W., Sakzad, A., Choo, K.R., 2021. Explainable artificial intelligence for digital forensics. Wiley Interdisciplinary Reviews: Forensic Sci., e1434
Haugeland, J., 1989. Artificial Intelligence: the Very Idea. MIT Press.
Henelius, A.P., Ukkonen, A., 2017. Interpreting Classifiers through Attribute Interactions in Datasets arXiv preprint arXiv:1707.07576.
Henelius, A., Puolamaki, K., Bostrom, H., Asker, L., Papapetrou, P., 2014. A peek into the black box: exploring classifiers by randomization. Data Min. Knowl. Discov. 28, 1503–1529.
Ho, T.K., 1995. Random decision forests. In: In Proc. of the 3rd Intl. Conf. on Document Analysis and Recognition, pp. 278–282.
Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844.
Huysman, J., Dejaeger, K., Mues, C., Vanthienen, J., Baesens, B., 2011. An empirical evauation of the comprehensibility of decision table, tree and rule based predictive models. Decis. Support Syst. 5, 141–154.
Ilkou, E., Koutraki, M., 2020. Symbolic vs sub-symbolic ai methods: friends or enemies?. In: CIKM Workshop.
Islam, M., Watters, P., Mahmood, A., Alazab, M., 2019. Toward detection of child exploitation material: a forensic approach. In: Deep Learning Applications for Cyber Security, pp. 221–246.
Johansson, U., Konig, R., Niklasson, L., 2004a. The truth is there —rule extraction from opaque models using genetic programming. In: In Proc. of FLAIRS Conference, pp. 658–663.
Johansson, U., Niklasson, L., Konig, R., 2004b. Accuracy vs. comprehensibility in data mining models. In: In Proc. of the 7th Intl. Conf. on Information Fusion, pp. 295–300.
Karimapanal, T., Bouffanais, R., 2018. Self-organizing maps for storage and transfer of knowledge in reinforcement learning. Adapt. Behav. 27, 111–126.
Karpathy, A., Johnson, J., Fei-Fei, L., 2016. Visualizing and understanding recurrent networks. In: ICLR Workshop Track. URL. http://vision.stanford.edu/pdf/KarpathyICLR2016.pdf.
Kim, B., Wattenberg, G.J., Cai, C., Wexler, J., Viegas, F., Sayres, R., 2017. Interpretability beyond Feature Atttribution: Quantitative Testing with Concept Activation Vectors (Tcav) arXiv preprint arXiv:1711.11279.
Kingma, D., Welling, M., 2013. Auto-encoding variational bayes. In: International Conference on Learning Representation (ICLR).
Koh, P.W., Liang, P., 2017. Understanding black-box predictions via influence functions. In: In Proc. 34th Intl. Conf. on Machine Learning, pp. 1885–1894.
Konig, R., Johansson, U., Niklasson, L., 2008. G-rex: a versatile framework for evolutionary data mining. In: 2008 IEEE Intl. Conf. on Data Mining, pp. 971–974.
Krakovna, V., Doshi-Velez, F., 2016. Increasing the Interpretability of Recurrent Neural Networks using Hidden Markov Models arXiv preprint arXiv:1606.05320.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Leslie, D., 2019. Understanding artificial intelligence and safety: a guide for the responsible design and implementation of aaai systems in the public sector. Alan Turing Institute.
Li, Z., 2002. A saliency map in primary visual cortex. Trends Cognit. Sci. 61, 9–16.
Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2021. Explainable ai: a review of machine learning interpretability methods. Entopy 23, 18–63.
Lipton, Z.C., 2018. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue 16, 31–57.
Lundberg, S.M., Lee, S., 2017. A unified approach to interpreting model predictions. In: In Proc. of the 31st Intl. Conf. on Neural Information Processing Systems, pp. 4768–4777.
Ma, X., Hovy, E., 2016. End-to-end Sequence Labelling via Bi-directional Lstm-Cnns-Crf arXiv preprint arXiv:1603.01354.
Manhaeve, R., Dumnacic, S., Kimming, A., Demeester, T., De Raedt, L., 2021. Neural probabilistic logic programming in deepproblog. Artif. Intell. 298, 103504.
Manning, C., Schutze, H., 1999. Foundation of Statistical Natural Language Processing. MIT Press.
Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B., Wu, J., 2018. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision, pp. 111–126 arXiv preprint arXIv:1904.12584 27.
Marcinowski, M., 2021. Deep learning v. human rights. In: 1st Intl. Workshop on Bias Ethics and Fairness in Artificial Intelligence: Representation and Reasoning (Befair 2021).
Mikolov, T., Karafiat, M., Burget, L., Cernosky, J., Khudanpur, S., 2010. Recurrent neural network based language model. Interspeech 2, 1045–1048.
Miller, G.A., 1956. The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol. Rev. 63, 81–97.
Miller, T., 2019. Explanation inartificial intelligence: insights fro the social sciences. Artif. Intell. 267, 1–38.
Miller, T., Howe, P., Sonenberg, L., 2017. Explainable ai: beware of inmates running asylum. IJCAI 2017 Workshop Explainable AI 36, 36–40.
Molnar, C., 2019. Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/.
Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Muller, K., 2017. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recogn. 65, 211–222.
Montavon, G., Samek, W., Muller, K., 2018. Methods for interpreting and understanding deep neural networks. Digit. Signal Process. 73, 1–15.
Nguyen, T., Choo, R., 2021. Human-in-the-loop xai-enabled vulnerability detection, investigation and mitigation. In: 2021 IEEE Intl. Conf. on Automated Software Engineering (ASE), pp. 1210–1212.
Nguyen, A., Dosovitskiy, A., Yosinki, J., Brox, T., Clune, J., 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Proc. of NIPS, pp. 3395–3403.
Noble, W.S., 2006. What is a support vector machine? Nat. Biotechnol. 24, 1565–1567.
Ośhea, K., Nash, R., 2015. An Introduction to Convolutional Neural Networks arXiv preprint arXiv:1511.08458.
Pasquale, F., 2015. The Black Box Society: the Secret Algorithms that Control Money and Information. Havard Univ. Press, Cambridge.
Pearl, J., 2020. Causality, second ed. Cambridge University Press.
Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Pappalardo, L., Ruggieri, S., Turini, F., 2018. Open the Black Box Data-Driven Explanantion of Black Box Decision Systems arXiv preprint arXiv:1806:09936.
Ribiero, M., Singh, S., Guestrin, C., 2016. Why should i trust you? explaining the predictions of any classifier. In: In Proc. of the 22nd ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, pp. 1135–1144.
Rissland, E.L., Ashley, K., 1987. A case-based system for trade secret law,. In: Proceedings of the 1st Intl. Conference on AI and Law, pp. 60–66.
Roth, A., 2015. Trial by machine. Georgetown Law J. 104, 1245.
Roth, A., 2017. Machine testimony. Yale Law J. 126, 1972–2053.
Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Learning 1, 206–215.
Ruping, S., 2006. Learning Interpretable Models. Ph.D. Dissetation. University of Dortmund URL. https://d-nb.info/997491736/34.
Sally, M.C., Terence, L.L., 1999. Maintenance and limitations issues of case-based reasoning technology in a manufacturing application. In: In Proc.. of AAAI Technical Report.
Samek, W., Weigand, T., Muller, K., 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models arXiv preprint arXiv:1708:08296.
Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., K, M., 2019. Explainable ai: interpreting, explaining and visualizing deep learning. Lect. Notes Comput. Sci. 11700.
Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, Q., Nguyen, C., Ngo, V., Seekins, J., Blankenberg, F., Ng, A.Y., Lungren, M.P., Raipukar, P., 2021. Benchmarking Saliency Methods for Chest X-Ray Interpretation medRxiv preprint.
Sato, M., Tsukimoto, H., 2001. Rule extraction from neural networks via decision tree induction. In: Intl. Conf. on Neural Networks.
Schneider, J., Breitinger, F., 2020. Towards Ai Forensics: Did the Artificial Intelligence System Do it? Why? arXiv preprint arXiv:2005.13635v2.
Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE Intl. Conf. on Computer Vision, pp. 618–626.
Shalaginov, A., 2017. Fuzzy logic model for digital forensics: a trade-off between

accuracy, complexity and interpretability. In: In Proc. of the Intl. Joint Conf. on Artificial Intelligence, pp. 5207–5208.
Shrikumar, A., Greenside, P., Kundaje, A., 2017. Learning important features through propagating activation differences, in: in Proc. of the 34th Intl. Conf. Mach. Learning 3145–3153.
Solanke, A.A., Chen, X., Ramiréz-Cruz, Y., 2021. Pattern recognition and reconstruction: detecting malicious deletions in textual communications, in: 2021 IEEE Intl. Conf. Big Data 2574–2582.
Solanke, A.A., Biasiotti, M.A., 2022. Digital Forensics AI: Evaluating, Standardizing and Optimizing Digital Evidence Mining Techniques. Künstl Intell. https://doi.org/10.1007/s13218-022-00763-9.
Su, G., Wei, D., Kush, R., Malioutov, D.M., 2016. Interpretable Two-Level Boolean Rule Learning for Classification arXiv preprint arXiv:1606.05798.
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2013. Intriguing Properties of Neural Networks arXiv preprint arXiv:1312.6199.
Thakur, A., Jindal, N., 2018. Machine learning based saliency algorithm for image forgery classification and localization. In: 1st Intl. Conf. on Secure Cyber Computing and Communication, pp. 451–456.
Thiagarajan, J., Kailkhura, B., Sattigeri, P., Ramamurthy, K.N., 2016. Treeview: Peeking into Deep Neural Networks via Feature-Space Partitioning arXiv preprint arXiv:1611.07429.
Underwoord, G., Foulsham, T., van Loon, E., Humphreys, L., Bloyce, J., 2006. Eye movements during scene inspection: a test of the saliency map hypothesis. J. Cognit. Psychol. 18, 321–342.
Wachter, S., Mittelstadt, B., Floridi, L., 2017. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int. Data Privacy Law 7, 76–99.
Wachter, S., Mittelstadt, B., Rusesell, C., 2018. Counterfactual explanation without opening the black box: automated decision and the gdpr. Harv. J. Law Technol. 31, 841.
Wolfson, A., 2005. Electronic fingerprints: doing away with the conception of computer-generated records as hearsay. Mich. Law Rev. 104.
Xu, K., Ba, J.L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y., 2015. Show, attend and tell: neural image caption generation with visual attention. In: Proc. of ICML, pp. 2048–2057.
Yampolski, R.V., 2020. Unexplainability and incomprehensibility of ai. J. AI Consciousness 7, 277–291.
Yang, J., Xiao, S., Li, A., Lan, G., Wang, H., 2021. Detecting fake images by identifying potential texture difference. Future Generat. Comput. Syst. 125, 127–135.
Yeh, C., Kim, J.S., Yen, I.E., Ravikumar, P., 2018. Representer point selection for explaining deep neural networks. In: In Advances on Neural Information Processing Systems, pp. 9311–9321.
Zeiler, M.D., Fergus, R., 2013. Visualizing and Understanding Convolutional Networks arXiv preprint arXiv:1311.2901.
Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R., 2010. Deconvolution networks. In: 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, p. 7.
Zeiler, M.D., Taylor, G.W., Fergus, R., 2011. Adaptive deconvolutional networks for mid and high level feature learning. In: 2011 Intl. Conf. on Computer Vision, p. 6.
Zeleznikow, J., Stranieri, A., 2017. The split-up system: integrating neural networks and rule-based reasoning in the legal domain. In: In Proc. of the 5th ICAIL, pp. 185–194.
Zhan, Y., Chen, Y., Zhang, Q., Kang, X., 2017. Image forensics based on transfer learning and convolutional network. In: In Proc. of the 5th ACM Workshop on Information Hiding and Multimedia Security, pp. 165–170.
Zilke, J.R., Menica, E.L., Janssen, F., 2016. Deepred—rule extraction from deep neural networks. In: In Intl. Conf. on Discovery Science, pp. 457–473.