Founding The Domain of AI Forensics

Ibrahim Baggili and Vahid Behzadan
University of New Haven
ibaggili@newhaven.edu, vbehzadan@newhaven.edu

Abstract

With the widespread integration of AI in everyday and critical technologies, it seems inevitable to witness increasing instances of failure in AI systems. In such cases, there arises a need for technical investigations that produce legally acceptable and scientifically indisputable findings and conclusions on the causes of such failures. Inspired by the domain of cyber forensics, this paper introduces the need for the establishment of *AI Forensics* as a new discipline under AI safety. Furthermore, we propose a taxonomy of the subfields under this discipline, and present a discussion on the foundational challenges that lay ahead of this new research area.

Introduction

Recent advances in Artificial Intelligence (AI) have given rise to the rapidly growing adoption of such techniques by a vast array of industries and technologies. The penetration of AI in our day-to-day lives is easily observed in everyday technologies such as advertisement and road navigation (e.g., Google Maps), as well as critical sectors such as cybersecurity (Li 2018), healthcare (Jiang et al. 2017), and smart cities (McKee et al. 2018). However, the growing complexity of AI techniques renders the assurance and verification of safety and reliability of such systems difficult (Yampolskiy 2018). Therefore, it is not surprising to observe the growing frequency of reported failures in AI-enabled systems (e.g., (Yampolskiy and Spellchecker 2016)).

In response, the evolving field of *AI safety* (Amodei et al. 2016) aims to tackle the problem of reliability and safety in AI-enabled systems. The resulting body of work to date is largely focused on the prevention of unsafe behavior in current and future AI technologies. However, the rapid penetration of AI into critical technologies has greatly outpaced the research efforts of the AI safety community. Hence, an increase in the frequency of failures in deployed AI seems inevitable. In the event of such failures in critical systems, it becomes necessary to investigate the causes and sequence of events leading to the failure. Besides the analysis of underlying technical deficiencies, such investigations will need to determine a variety of other aspects, including: whether the failure has been the result of malicious actions, which party is liable for the damages caused by the failure, and whether the failure could have been prevented. Furthermore, interested parties such as law enforcement and insurance providers may require this investigation to result in legally acceptable and indisputable findings and conclusions.

Similar needs in the domain of computer safety and security have given rise to the field of *Cyber Forensics*. Digital Forensics, also known as Cyber Forensics, revolves around the scientific and legal extraction of digital evidence. This field is multidisciplinary and involves computing, law, criminology, psychology and other disciplines. At the core of the domain, however, is the Acquisition, Authentication and Analysis (AAA) of digital evidence.

Inspired by this analogue, we argue for the need to establish the formalism of *AI Forensics* as a new discipline under AI safety. This formalism will aim to develop the tools, techniques, and protocols for the forensic analysis of AI failures. Accordingly, this paper makes the following contributions:

- We offer the first working definition of AI Forensics.
- We conceptualize the first formal attempt of the AI Forensics domain, and propose a taxonomy for the corresponding types and sources of evidence.
- We enumerate a number of notable challenges in the domain of AI Forensics.

Related Work

In recent years, there has been a growing interest in failure detection and analysis techniques for algorithmic decision making(Goodman and Flaxman 2016). This is partly due to the European Union's General Data Protection Regulations (GDPR), which requires the explainability of consequential decisions made by algorithms. Similarly, the research on Explainable AI (Samek, Wiegand, and Müller 2017) aims to create tools and techniques that enable the explainability of black-box models such as deep neural networks. However, current state of the art in the analysis of failures in AI, and in particular machine learning models, is largely focused on the technical diagnosis and troubleshooting of design and training issues (e.g., (Nushi, Kamar, and Horvitz 2018)). Hence, there remains a gap with regards to tools and techniques that enable the forensic analysis of failures in AI-enabled systems.

Digital Forensics and Digital Evidence

Digital forensics is defined as *"The use of scientifically derived and proven methods toward the preservation, col-*

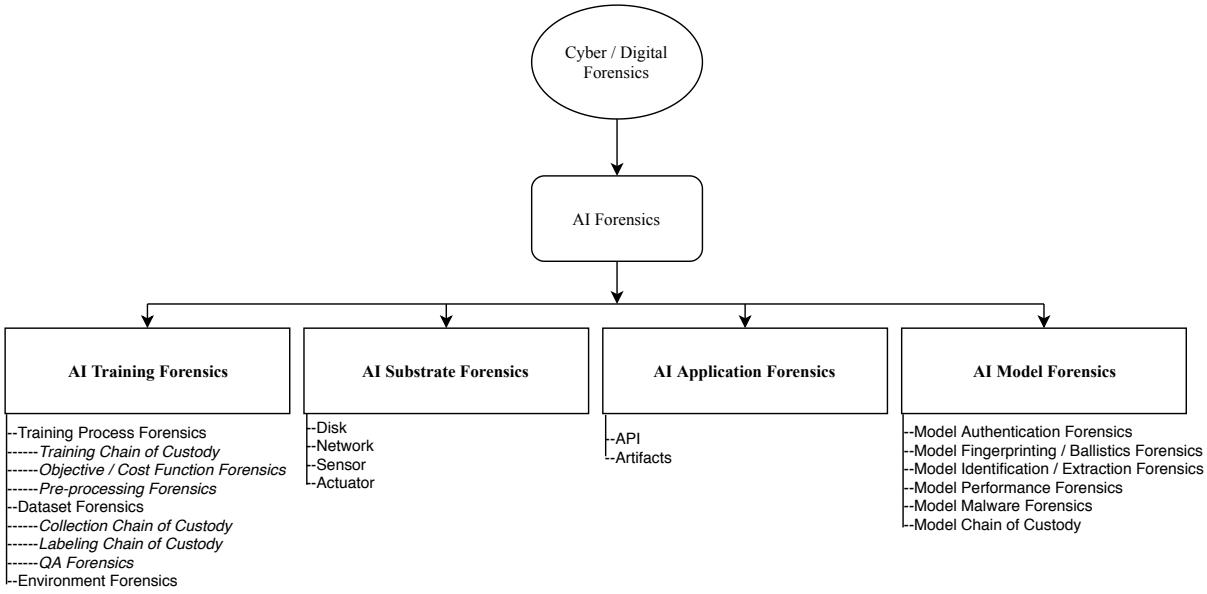

Figure 1: Diagram of the AI Forensics Research Domain and Sub-Domains

lection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating 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). An important component of digital evidence is its admissibility to the court of law.

The admissibility of evidence was mostly dominated by what is known as the Frye test that articulated expert scientific evidence is admissible only if the scientific community generally accepts it. Since 1993, courts in the United States have adopted Rule 702 of the federal rules of evidence, resulting in what many refer to as the *Daubert process*. This process comprises of four major guidelines (Farrell 1993):

1. Testing: Can and has the procedure been tested?
2. Error Rate: Is there a known error rate of the procedure?
3. Publication: Has the procedure been published and subject to peer review?
4. Acceptance: Is the procedure generally accepted in the relevant scientific community?

Digital forensics has made strides by exploring areas such as disk forensics (Garfinkel 2009), memory forensics (Ligh et al. 2014), network forensics (Karpisek, Baggili, and Breitinger 2015), cloud forensics (Ruan et al. 2013), artifact forensics (Grajeda et al. 2018), blockchain storage and cryptocurrency forensics (Ricci, Baggili, and Breitinger 2019), social media forensics (Al-khateeb et al. 2016), authorship attribution (Mohan, Baggili, and Rogers 2010), and mobile forensics (Baggili et al. 2015).

The aforementioned forensic areas also benefit from AI techniques, as process automation in digital forensics is of importance given the volume, variety and velocity production of data.While the opportunity for AI Forensics has been recently noted by experts (Luciano et al. 2018), at the time of writing this work, the domain was ill-defined.

The Landscape of AI Forensics

We view the scope of AI Forensics as a subfield of digital forensics, defined as the *scientific and legal tools, techniques, and protocols for the extraction, collection, analysis, and reporting of digital evidence pertaining to failures in AI-enabled systems.* In compliance with the Daubert process, AI Forensics provides a framework to enable the systematic and scientific resolution of such questions as:

- What were the sequence of events and conditions that led to the failure?
- Did the failure result from malicious actions?
- Which party or parties is responsible for the failure?
- Would it have been possible to prevent the failure?
- Where did the failures take place?

The core of any forensic investigation is the collection and extraction of evidence. To this end, we introduce a classification of the various types evidence that can be of relevance to the investigation. Furthermore, we identify possible sources of evidence for each of the enumerated types.

AI Training Forensics

In forensic investigations of machine learning systems, a necessary step is to identify potential intentional or unintentional faults introduced during the training of the system. Such faults may stem from any of the components involved in training, as detailed below:

Training Process Forensics: The training process comprises of the optimization algorithm and its corresponding hyperparameters. A major cause of AI failures is the misspecification of the objective or cost function, which may result in behaviors that are misaligned with the goals of the designer (Arnold, Kasenberg, and Scheutz 2017). Furthermore, design choices such as exploration techniques in reinforcement learning agents (Behzadan and Munir 2018) and regularization techniques may result in brittle behaviors that fail to adapt to distributional shifts in their settings of deployment (Papernot et al. 2018). Therefore, such parameters and choices constitute valuable forensic evidence.

Dataset Forensics: The dataset used in the training of a machine learning model may be of inconsistent or unrepresentative samples, thus resulting in a model that is not compatible with the conditions of its deployment settings. Furthermore, the training dataset may be subject to intentional manipulations (e.g., data poisoning attacks (Papernot et al. 2018) and backdoor injections (Chen et al. 2017)). Therefore, forensic investigations can benefit from access to the dataset and knowledge of its compilation methodology. Also, access to the modification history of the dataset can help with identifying intentional manipulations, as well as the responsible parties.

Environment Forensics: The analogue of training data for reinforcement learning agents is the training environment. Similar to the case of training data, unrepresentative or manipulated environments may result in faulty behavior and failures (e.g., (Behzadan, Yampolskiy, and Munir 2018) and (Behzadan and Hsu 2019)). Hence, access to the training environment and its modification history can prove useful in forensic investigations

AI Substrate Forensics

Substrate refers to the hardware and software platform that hosts the AI system. Forensic investigations of the AI substrate is essentially the domain of Digital Forensics. However, there are certain aspects of AI substrates that may give rise to unique circumstances. For instance, random bit-flips in the processor or memory due to cosmic rays has been shown to result in potentionally significant failures (Santoso and Jeon 2019). Furthermore, in cyber-physical AI systems, impaired or manipulated actuation of mechanical components (e.g., robotic locomotion) may result in misrepresentations of states and consequences of actions (Behzadan and Munir 2018). Main sources of forensic evidence in the AI substrate include the disk and memory, the network component, as well as the conditions of sensors and actuators in the cyber-physical settings.

AI Application Forensics

AI is often deployed as a component of an application system. For instance, cloud-based cognitive services provide Application Programming Interfaces (APIs) to enable the integration of an AI service in software products. Forensic evidence collected from the usage logs of such APIs may indicate manipulation attempts, as well as failures in correct data cleaning and processing of queries. Furthermore, analysis of artefacts such as system resource utilization logs, authentication logs, and file system logs may also provide useful information on anomalous behavior and its causes.

AI Model Forensics

Deployed machine learning models may result in failures that are independent of the aforementioned sources. For instance, original models may have been manipulated or replaced at some point in the machine learning supply chain. This type of malicious behavior may manifest in the form of backdoored models (e.g., (Chen et al. 2017)), corrupted models (e.g., poisoning of malware classification (Chen et al. 2018b)), or intentionally malicious models. While the domain of explainable AI offers an array of tools that may prove helpful in the analysis of the decision-making process in such models, a comprehensive forensic investigation requires further evidence which may be obtained from alternative sources, as discussed below:

Model Authentication Forensics: The aim of such evidence is to enable the verification of the authenticity of the model under investigation. Recent advances in watermarking techniques for machine learning models (Behzadan and Hsu 2019; Zhang et al. 2018) provide the means for direct authentication of models. However, such approaches are yet to be commonly adopted. Furthermore, watermarks may also be prone to tampering and forging. Hence, alternative evidence such as software-level hashing techniques can provide a more reliable alternative.

Model Identification / Extraction Forensics: If the model under investigation is a blackbox (i.e., model parameters and architecture are unknown), techniques such as model inversion and extraction (Tramèr et al. 2016) may provide the means for replicating its behavior for further testing and analysis. However, due to the approximate nature of such replicas, there remains the need for techniques that enable the quantification of uncertainty to maintain the legal soundness of the resulting forensic analyses.

Model Ballistics Forensics: In the general domain of Forensics, ballistics refers to the analysis and identification of the type and owner of a weapon used in a shooting incident. Similarly, in the forensic investigation of suspicious or malicious models, it is of importance to determine the type and creators (tools and individuals) of the model.

Model Performance Forensics Recording the values of internal metrics and variables in the model may provide a detailed insight into the inner workings of the model. For instance, (Chen et al. 2018a) demonstrate that the analysis of activation values in deep learning models facilitates the detection of hidden backdoors. Also, higher-level measurements of model internals, such as state-action value estimates of reinforcement learning agents, and the multi-class probability distribution of classifiers, may provide useful evidence on the origins of failures.

Model Malware Forensics As mentioned before, machine learning models can be infected with backdoors and trigger-activated policies, or have been intentionally trained

to act maliciously. An AI forensic investigation needs to detect and establish the existence of such malware, and provide the means to determine their malicious intent. Inspired the sandboxing techniques of computer malware analysis (Greamo and Ghosh 2011), a preliminary source of forensic evidence in such cases is to replicate the conditions in simulation or a controlled environment to observe and analyze the behavioral dynamics of the model under investigation.

Challenges

Unexplainability of AI

Sound and indisputable root-cause analysis of failures in AI may require transparent and accurate interpretations of the decision-making process which resulted in undesired behavior. However, the research on the explainability of complex AI systems is still at its early stages, and the state of the art is far from solving the problem of explainability. Furthermore, some recent literature (e.g., (Yampolskiy 2019)) argue that as the AI technology and capabilities advance over time, it may become more difficult, or even impossible for AI systems to be explainable. In such circumstances, simpler abstractions of the decision-making process may enhance the forensic analysis of such failures. For instance, (Behzadan, Munir, and Yampolskiy 2018) propose a psychopathological abstraction for complex AI safety problems. Similar abstractions may be required to enable accurate forensic analysis of advanced AI.

AI Anti-Forensics

In the domain of digital forensics, criminals constantly adapt to the state of the technology, and utilize techniques such as decoys, false evidence, or forensic cleaning to impede the forensic investigation. It is likely that such anti-forensics techniques may also be invented and adopted by criminals to manipulate AI forensic investigations. Proactive identification of such techniques and development of mitigating solutions will thus become an increasingly important area of research in this domain. A recent anti-forensics general taxonomy was devised by researchers (Conlan, Baggili, and Breitinger 2016), yet, AI anti-forensics was not included in the taxonomy. While this is a challenge, it also presents a ripe opportunity for researchers.

Disconnect Between the Cyber Forensics and AI Communities

One of the biggest challenges we face is the disconnect between the AI Safety and Cyber Forensics communities. Scientists from those two domains are not working together, thus, the domain of AI Forensics has not been conceived and is ripe for future work. This disconnect was apparent in a recent survey study, where the majority of digital forensic practitioners (67%) (disagreed, agreed or were neutral) on their competency in Data Science (Sanchez et al. 2019).

Conclusion

We argued that the widespread integration of AI in everyday and critical technologies is bound to result in increased instances of failure, which will require technical investigations that produce legally acceptable and scientifically indisputable findings and conclusions. Inspired by the domain of cyber forensics, we thus introduced the need for the establishment of AI forensics as a new discipline under AI safety. Furthermore, we proposed a taxonomy of the subfields under this discipline, and presented a discussion on the foundational challenges that lay ahead of this new research area.

References

Al-khateeb, S.; Conlan, K. J.; Agarwal, N.; Baggili, I.; and Breitinger, F. 2016. Exploring deviant hacker networks (dhm) on social media platforms. *Journal of Digital Forensics, Security and Law* 11(2):1.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Arnold, T.; Kasenberg, D.; and Scheutz, M. 2017. Value alignment or misalignment–what will keep systems accountable? In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Baggili, I.; Oduro, J.; Anthony, K.; Breitinger, F.; and McGee, G. 2015. Watch what you wear: preliminary forensic analysis of smart watches. In *2015 10th International Conference on Availability, Reliability and Security*, 303–311. IEEE.

Behzadan, V., and Hsu, W. 2019. Sequential triggers for watermarking of deep reinforcement learning policies. *arXiv preprint arXiv:1906.01126*.

Behzadan, V., and Munir, A. 2018. The faults in our pi stars: Security issues and open challenges in deep reinforcement learning. *arXiv preprint arXiv:1810.10369*.

Behzadan, V.; Munir, A.; and Yampolskiy, R. V. 2018. A psychopathological approach to safety engineering in AI and AGI. In *Computer Safety, Reliability, and Security - SAFECOMP 2018 Workshops, Västerås, Sweden, September 18, 2018, Proceedings*, 513–520.

Behzadan, V.; Yampolskiy, R. V.; and Munir, A. 2018. Emergence of addictive behaviors in reinforcement learning agents. *arXiv preprint arXiv:1811.05590*.

Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.

Chen, B.; Carvalho, W.; Baracaldo, N.; Ludwig, H.; Edwards, B.; Lee, T.; Molloy, I.; and Srivastava, B. 2018a. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*.

Chen, S.; Xue, M.; Fan, L.; Hao, S.; Xu, L.; Zhu, H.; and Li, B. 2018b. Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach. *computers & security* 73:326–344.

Conlan, K.; Baggili, I.; and Breitinger, F. 2016. Anti-forensics: Furthering digital forensic science through a new extended, granular taxonomy. *Digital investigation* 18:S66–S75.

Farrell, M. G. 1993. Daubert v. merrell dow pharmaceuticals, inc.: epistemilogy and legal process. *Cardozo L. Rev.* 15:2183.

Garfinkel, S. L. 2009. Automating disk forensic processing with sleuthkit, xml and python. In *2009 Fourth International IEEE Workshop on Systematic Approaches to Digital Forensic Engineering*, 73–84. IEEE.

Goodman, B., and Flaxman, S. 2016. Eu regulations on algorithmic decision-making and a "right to explanation". In *ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY. http://arxiv.org/abs/1606.08813 v1*.

Grajeda, C.; Sanchez, L.; Baggili, I.; Clark, D.; and Breitinger, F. 2018. Experience constructing the artifact genome project (agp): Managing the domain's knowledge one artifact at a time. *Digital Investigation* 26:S47–S58.

Greamo, C., and Ghosh, A. 2011. Sandboxing and virtualization: Modern tools for combating malware. *IEEE Security & Privacy* 9(2):79–82.

Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; and Wang, Y. 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology* 2(4):230–243.

Karpisek, F.; Baggili, I.; and Breitinger, F. 2015. Whatsapp network forensics: Decrypting and understanding the whatsapp call signaling messages. *Digital Investigation* 15:110–118.

Li, J.-h. 2018. Cyber security meets artificial intelligence: a survey. *Frontiers of Information Technology & Electronic Engineering* 19(12):1462–1474.

Ligh, M. H.; Case, A.; Levy, J.; and Walters, A. 2014. *The art of memory forensics: detecting malware and threats in windows, linux, and Mac memory*. John Wiley & Sons.

Luciano, L.; Baggili, I.; Topor, M.; Casey, P.; and Breitinger, F. 2018. Digital forensics in the next five years. In *Proceedings of the 13th International Conference on Availability, Reliability and Security*, 46. ACM.

McKee, D. W.; Clement, S. J.; Almutairi, J.; and Xu, J. 2018. Survey of advances and challenges in intelligent autonomy for distributed cyber-physical systems. *CAAI Transactions on Intelligence Technology* 3(2):75–82.

Mohan, A.; Baggili, I. M.; and Rogers, M. K. 2010. Authorship attribution of sms messages using an n-grams approach. *Technical Report, CERIAS 2010-11 College of Technology*.

Nushi, B.; Kamar, E.; and Horvitz, E. 2018. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

Palmer, G. 2001. A road map for digital forensics research-report from the first digital forensics research workshop (dfrws). *Utica, New York*.

Papernot, N.; McDaniel, P.; Sinha, A.; and Wellman, M. P. 2018. Sok: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, 399–414. IEEE.

Ricci, J.; Baggili, I.; and Breitinger, F. 2019. Blockchain-based distributed cloud storage digital forensics: Where's the beef? *IEEE Security & Privacy* 17(1):34–42.

Ruan, K.; Carthy, J.; Kechadi, T.; and Baggili, I. 2013. Cloud forensics definitions and critical criteria for cloud forensic capability: An overview of survey results. *Digital Investigation* 10(1):34–43.

Samek, W.; Wiegand, T.; and Müller, K.-R. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Sanchez, L.; Grajeda, C.; Baggili, I.; and Hall, C. 2019. A practitioner survey exploring the value of forensic tools, ai, filtering, & safer presentation for investigating child sexual abuse material (csam). *Digital Investigation* 29:S124–S142.

Santoso, D., and Jeon, H. 2019. Understanding of gpu architectural vulnerability for deep learning workloads. In *2019 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, 1–6. IEEE.

Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M. K.; and Ristenpart, T. 2016. Stealing machine learning models via prediction apis. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, 601–618.

Yampolskiy, R. V., and Spellchecker, M. 2016. Artificial intelligence safety and cybersecurity: A timeline of ai failures. *arXiv preprint arXiv:1610.07997*.

Yampolskiy, R. V. 2018. *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC.

Yampolskiy, R. V. 2019. Unexplainability and incomprehensibility of artificial intelligence.

Zhang, J.; Gu, Z.; Jang, J.; Wu, H.; Stoecklin, M. P.; Huang, H.; and Molloy, I. 2018. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, 159–172. ACM.