Founding the Domain of AI Forensics

SAIL Lab

Ibrahim Baggili
Director - Cyber Forensics Research and Education Group
ibaggili@newhaven.edu
&
Vahid Behzadan
Director - Secure and assured Intelligent Learning (SAIL) Lab
vbehzadan@newhaven.edu

University of New Haven

Inevitable Failures

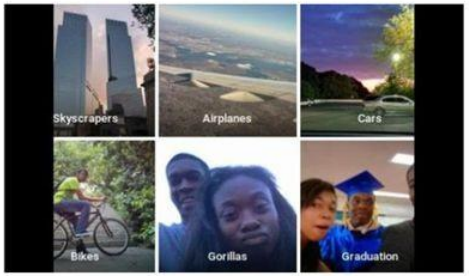
NEWS

Home | Video | World | US & Canada | UK | Business | Tech | Science | Magazine

Technology

Google apologises for Photos app's racist blunder

1 July 2015 | Technology



Andrew J. Hawkins
@andyjayhawk
Follow

In 2016, a Tesla driver using Autopilot crashed into the side of a truck and was killed. It happened again three months ago, but this time with a completely new version of Autopilot. What's the heck is going on?? theverge.com/2019/5/17/1862 ...

1:14 PM - 17 May 2019

GET HYPED

Doctors Are Losing Faith in IBM Watson's AI Doctor

Robust Physical-World Attacks on Machine Learning Models

Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, Dawn Song

(Submitted on 27 Jul 2017 (v1), last revised 30 Jul 2017 (this version, v2))

The need for autopsy

- Problem: What sequence of events led to the failure?
- Technical Analysis
  - Prevention / mitigating
  - debugging
- Legal
  - Liability
  - Responsibility (e.g., criminal negligence, breach of contract)
  - Criminal (e.g., intentional faults, malicious compromise)

Digital Forensics

- ***Definition:*** "scientifically derived and proven methods toward the *preservation, collection, validation, identification, analysis, interpretation, documentation and presentation* of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events … shown to be disruptive to planned operations" (Palmer 2001)

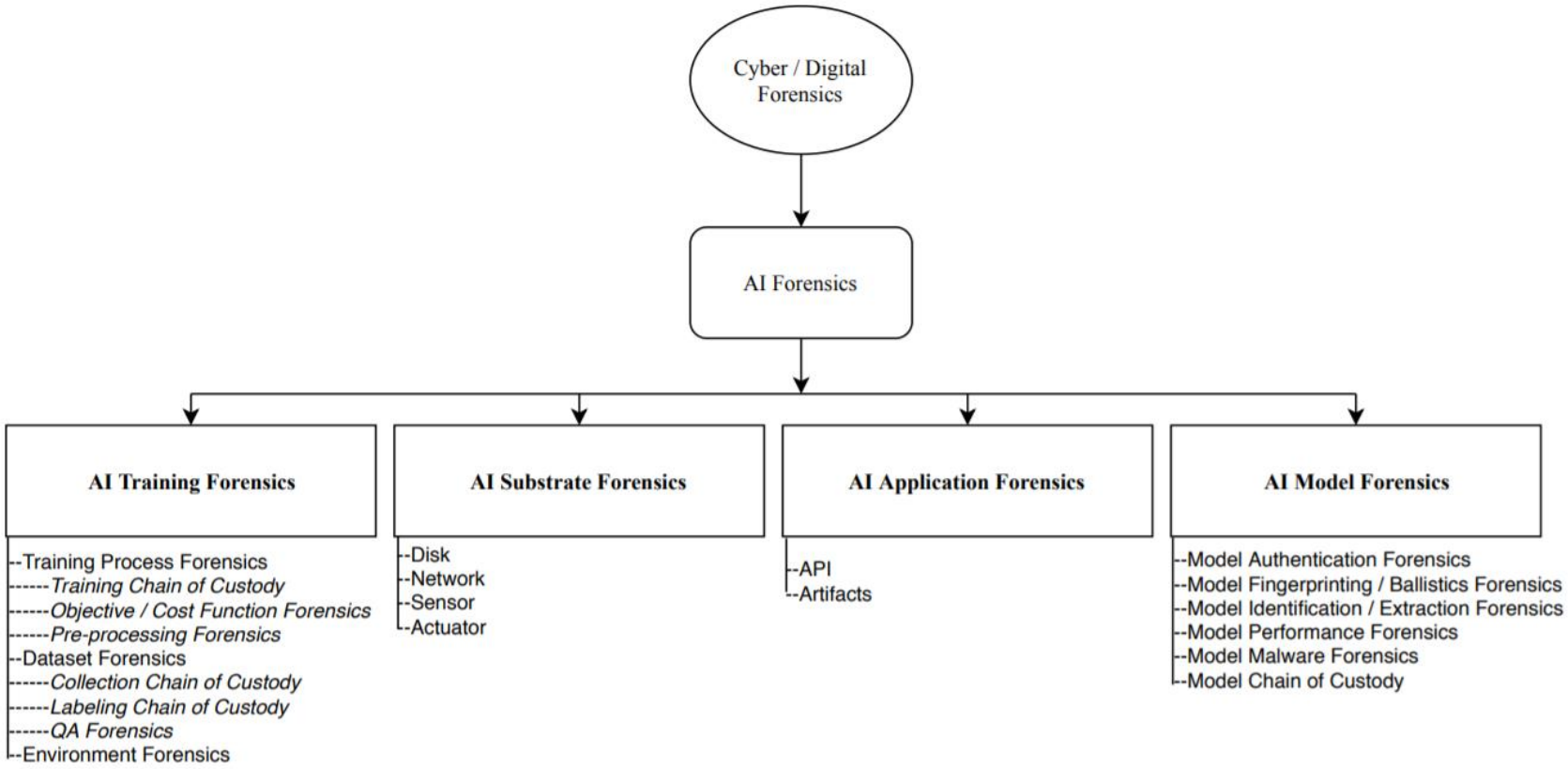
- ***Daubert process*** (Farrell 1993): *Testing, Accuracy, Published, Accepted*

AI Forensics

Definition: *Scientific and legal tools, techniques, and protocols for the extraction, collection, analysis, and reporting of digital evidence pertaining to failures in AI-enabled systems.*

- Scope
  - 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 (at what stage or component) did the failures take place?

The Landscape of AI Forensics



Training Forensics

- Forensic Analysis of faults introduced in training
  - *Training Process Forensics*- Optimization algorithm and hyperparameters
    - Objective/reward, Exploration strategy, regularization, etc.
  - *Dataset Forensics*
    - Inconsistent or unrepresentative data
    - Intentional manipulation – e.g., poisoning, Trojan, etc.
  - *Environment Forensics*
    - Inaccurate representation/model
    - Intentional manipulation

Substrate and Application Forensics

- ***Substrate Forensics:*** Hardware and software platforms hosting the AI agent.
  - Areas overlap with cyber forensics:
    - Network components
    - Disk and memory – e.g., random bit flips due to cosmic rays
    - Actuators and Sensors – e.g., manipulated servos and sensors
- ***Application Forensics:*** AI as a component of application
  - API calls
  - Authentication and access control
  - Data sanitization
  - File system / resource allocation

Model Forensics

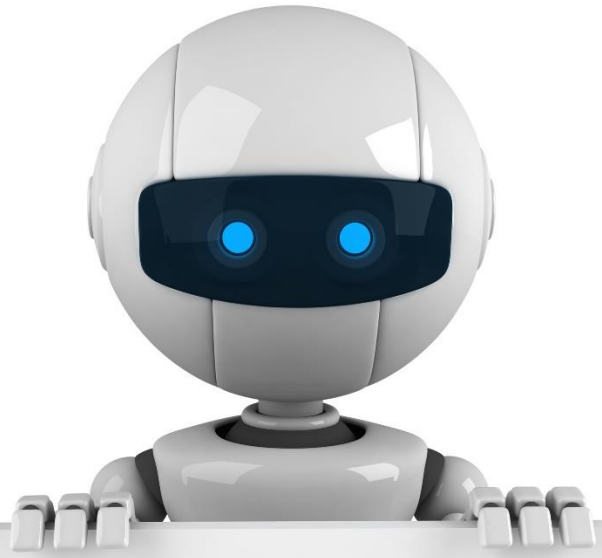
- Forensic analysis of deployed model (post-training)
  - *Model Authentication:* Is this model tampered with or modified from the original?
  - *Model Identification:* What does this model do?
  - *Model Ballistics:* Who created the model, which platform was used?
  - *Model Internals:* Is there anything unusual under the hood?
    - e.g., activation clustering for backdoor detection (Chen et al. 2018)
  - *Malware Forensics:* Is this model infected? If so, what type of infection? (e.g., backdoor, policy trigger)

Challenges

- Unexplainability of AI
  - Forensic need for explainability and interpretability
  - However, advanced AI and complex models may be difficult or impossible to interpret (Yampolskiy 2019)
  - Potential solution: Higher level abstractions
    - e.g., psychopathological modeling of AI Safety (Behzadan et al., 2018)
- AI Anti-Forensics
  - Malicious actors evolve, too
    - Decoys, false evidence, forensic cleaning
  - Need for techniques for proactive identification and mitigation techniques

Conclusion

- AI safety research is generally focused on prevention.
- AI failures are inevitable.
- We need a new set of techniques for establishing the root cause
- Abundance of open problems and challenges
- We are already late!



Thank You

University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

SAIL Lab

UNHcFREG