



Practitioner Perspectives on Making AI Defensible in Digital Forensics Casework

A report of the collective working session involving all DFRWS attendees

From **2026 Digital Forensics Conference Americas**

Arlington, VA (July 25th - 30th)

DFRWS is dedicated to the sharing of knowledge and ideas about digital forensics research. Ever since it organized the first open workshop devoted to digital forensics in 2001, DFRWS continues to bring academics and practitioners together in an informal environment.

As a non-profit, volunteer organization, DFRWS sponsors technical working groups, annual conferences and challenges to help drive the direction of research and development.

<http://dfrws.org>

Practitioner Perspectives on Making AI Defensible in Digital Forensics Casework

Sep 30, 2026

Authors:

- Eoghan Casey — Cyber Knowledge Engineering.
- Jim Jones — George Mason University.
- Jessica Hyde — Hexordia.
- Cameron Kelley — Mobasi.
- Chad Steel — George Mason University.

This document captures numerous discussions, including planning meetings, hallway conversations, email exchanges, and the working session. The authors carefully curated content from presentations slides, speaker notes, and audience responses. Although it was necessary to condense the content, every effort was made to incorporate participants' perspectives accurately. Uses of AI in experiments, examples, and editing were explicitly noted.

1) Background

Use of artificial intelligence (AI) in digital forensics brings opportunities, particularly efficiency in tool development and data analysis, but also has associated risks that must be addressed. To address the questions we actually face when AI enters the digital forensics workflow, a practitioner-led interactive working session was held during the Digital Forensics Conference Americas 2026, organized by DFRWS.org at George Mason University. This session concentrated on forensic practitioners using AI and treating AI generated evidence. AI defensibility in digital forensics focused on human oversight, performance evaluation, responsible use, error mitigation, system engineering, and expertise grounding. Five practitioners shared their experiences and led group discussions, guided by live audience polls projected on a large screen to help focus on participants' perspectives.

2) Key Concerns

The working session opened with an audience poll "What are your concerns about AI in digital forensics?" Answers spanned a wide range of concerns, including errors, eventual costs, difficulty validating or reproducing results, and human factors such as investigators trusting it blindly and a judge/jury misunderstanding capability. The potential for errors took several forms: LLM creativity, hallucinations, misleading information, convincingly wrong answers, incorrect or incomplete responses, and wrong analysis in reports. Bad prompting, insufficient guardrails, and non-deterministic outputs were also listed as concerns.

An emerging challenge is the lack of knowledge and digital forensics capabilities where AI is used as part of the crime. Forensic analysis of AI usage artifacts, including LLM audit trails, requires further research and development.

On the human side, primary concerns revolved around deskilling and knowledge gaps, specifically decline in digital forensics expertise and investigators blindly trusting AI. As practitioners rely more on automation and AI (e.g., an “easy button” for finding digital evidence), their technical and critical thinking skills may atrophy, and they may accept outputs as facts without assessing their veracity and uncertainty. Vendors are marketing AI enhanced tools that deliver rapid results for investigators, without consideration for the aforementioned potential errors. General principles that flow from this are:

Principles

Statistics from LLMs are not definitive proof

AI is not a substitute for digital forensics analysis.

Judges or juries misunderstanding AI capabilities was also raised, which could cause blind trust in some situations and lack of trust in others.

Although vibe coding was mentioned, later discussion addressed this in terms of traditional responsibility to tool testing and quality assurance.

3) Understanding Uncertainty (Led by Jim Jones, GMU)

An illustrative example of evaluating performance of digital forensics tools was presented to set the stage for further discussion of statistics and uncertainty in AI-enabled capabilities.

Recovering and analyzing data from raw media and file systems is one of the most fundamental practices in digital forensics. After a file is deleted and partially overwritten, how confident can you be that a particular file was previously on the device? The answer depends on various factors, which brings uncertainty and requires statistical analysis.

Patterns of persistence of non-allocated files on storage media are nondeterministic. It is hard to predict what part will persist over time (beginning, middle, end, etc.). The predictive value of recoverable file content depends on its relative uniqueness, with low entropy content having low value (e.g., white space) and high entropy content having higher value (e.g., compressed or encrypted data). Sector hashing statistics can help establish relative uniqueness to further suggest a recovered fragment’s inferential value. More information in and around the recovered data may provide context and contribute to the evaluation of the question about prior full file presence:

- Filename / metadata
- Exif metadata
- Relative offset within cluster
- Plausibility (e.g., 1st sector of a fully utilized cluster intact while rest of cluster is overwritten is a red flag)

Trying to classify/identify file fragments using existing AI models provides insights into performance. The following illustrative experiment was performed:

- Gave 10 single sector (512 bytes) samples from multiple files to Claude Sonnet
- No fragments were attributed to specific source files
- Some fragments were correctly classified as executable code (e.g., PE Headers)
- Some fragments were correctly recognized as indicating PyInstaller use

Then the middle part of cmd.exe and HxD.exe (publicly available code with stripped starting and ending sections) were provided to Claude Sonnet, which correctly identified the content as part of the original files.

To reinforce the effects of model thresholds and tradeoffs, the audience was given a recap of classifier performance with a confusion matrix (FP FN TP TN) to determine precision, recall, accuracy, and F1 score. A survey was conducted for different ROC Curves to demonstrate the point Want higher Area Under Curve = better Performance of the model (max 1).

4) Using AI Results in Casework (Led by Chad Steel, GMU)

The discussion opened with opportunities for using AI in digital forensics. An illustrative example of using GenAI conversation logs to show that someone was looking for ways to hide a body. This is analogous to the forensic examination of browser search history for “how to hide body?” Now people use AI to elaborate and clarify, with AI chat history stored and automatically replicated. For example, someone might type the initial prompt “How to extort?” and then elaborate on “How to delete data from a SQLite database?” Because of privacy concerns related to GenAI storage and use of prompts, that query is instead stored within the account of the individual entering the query, and is replicated between devices they use and are logged into so they have a seamless interaction when switching between laptop and mobile clients. This provides forensic opportunities - an individual who deletes their account info may still have their chat history replicated and recoverable on other devices.

Further discussion on this topic highlighted that there is an increasing amount of AI generated CSAM. This raises questions, including:

- What is the goal of the investigation, and what is the role of victim identification?
- How do we detect AI image (which is a trick question, even normal images are AI. A typical iPhone image has stability management, lighting corrections, histogram equalization, sharpening, and other processing that is now transparent and AI-driven)?
- What do we mean when we say AI generated? In-painted images have different detection techniques than fully prompt-created AI images.
- How good are detection tools (classifiers) that use AI (use ROC curves / F1 score)? Many are tested on images from a single model (and may detect either the model or the export library), and don't necessarily reflect in-the-wild scenarios. Additionally, even with high precision/recall numbers, they may not be effective in many use cases where even a 99% performance level results in too many false positives/negatives (e.g., processing a billion images a day on Instagram).

Offenders using apps that have AI built-in may grant access to their Photos library. In doing so, these tools may access photos, which leaves significant artifacts including local, lower resolution versions of those images that persist even when the originals are deleted (lower resolution versions are often more effective/efficient for object detection). However, the presence of a thumbs.db with artifacts of photos, or similar data cache in an app, does not necessarily mean the app was actually utilized by the user to view those photos. The app might have automatically indexed photos in the background without the user realizing it was happening.

4.1) Appropriate AI Usage

The audience was polled with the question: *What are some appropriate ways you use AI in your work?* Responses were anonymous and unseeded, and the top five results are ranked in order of popularity in the following table.

Theme	Rank
Writing and text improvement	1
Coding assistance and tool development	2
Data classification and pattern detection	3
Search, research, and brainstorming	4
Bulk data/document summarisation and triage	5

Table 1: Categorized and ranked audience responses to poll question “*What are some appropriate ways you use AI in your work?*” (AI summarized, human verified)

The group discussion delved into these topics. Although AI is commonly used to assist with writing, for integrity and transparency, it is important to disclose when and how a document or paper has been influenced by AI. It is also essential for authors to take full responsibility for written content they produce during their work. Any errors introduced by AI must be caught and corrected by the author.

For digital forensics work, using advanced processing is helpful for lead purposes, but any finding should be corroborated from other sources. For example, an image enhancement tool might provide a probabilistic plate number from a blurred license plate. This should be treated as a starting point or hypothesis for testing, not taken as ground truth.

AI-based image and video processing in general is a critical part of most digital forensics pipelines. At ingest, automated transcription and translation are performed for audio and video prior to indexing, including OCR of text content in still images. Increasingly, semantic image classification is used as well, providing concept-based searches (e.g., identifying all images in a hotel room). These can then be leveraged for further classification and clustering, making triage

and review more rapid. In child sexual abuse material (CSAM) cases, automated identification can reduce the amount of content that examiners need to review, with perceptual hashing automatically flagging relevant content, though the offending content still needs manual confirmation for court use.

Frontier models are well suited to assisting with computer programming, and digital forensics has existing approaches to validating tool results. Specific use cases include parsing unknown system logfiles, analyzing AI conversation logs, and parsing logs of installing forensic tools when troubleshooting. As with written work products, the author of a parser or tool is ultimately responsible for the work and must run and own the code. It is important to retain any code as an artifact for future review and verification and for replication purposes.

Using AI to prioritize and triage large amounts of data, such as bulk documents or emails can improve efficiency and help find relevant information. Risks include missing relevant information and errors or hallucinations, so important findings must be carefully reviewed by a person. Knowledge of domain-specific content is also an issue - a highly technical corpus might require a custom-trained (or LoRA boosted) model. Cost is also a concern - running an LLM against millions of records can end up with costs in the tens of thousands of dollars using naive approaches at current token costs, so run test prompts on data samples to manage expense. AI can also be used to apply labels or taint markings to email and documents, classified against a *defined code book* rather than ad-hoc categories. By creating a documented code book in advance, the digital forensics practitioner has a grounded basis for why individual artifacts were classified into a particular category.

Although it might be appropriate to use AI for lead generation, there are risks associated with even running AI processing when presenting results in expert reports and court testimony. For example, in a narcotics investigation involving a violent attack with a weapon, if you find a photo of the defendant with piles of money, drugs, and firearms (“money shot”) do you want to have a tool report that it is 60% AI generated? What are your discovery obligations for that analysis?

4.2) Inappropriate AI Usage

The participants were then polled with the question: *What are some inappropriate uses of AI in your work?* Responses were anonymous and unseeded, and the top results are ranked in order of popularity in the following table. The general concern that AI use causes distraction from work and time waste was prominent, but is broader than just digital forensics and not included in the ranked results.

Theme	Rank
Illegal and overtly illegitimate use	1
Unverified output and absent human oversight	2
Processing sensitive or confidential data	3

Manufacture evidence and disinformation	4
Deterministic and computational tasks	5

Table 2: Categorized and ranked audience responses to poll question “*What are some inappropriate ways you use AI in your work?*” (AI summarized, human verified)

During the group discussion, there was strong consensus around it being unacceptable to abdicate investigations to AI, including to find evidence if not subsequently verified, failing to verify the output, trusting it and not checking its results. A lack of person-in-the-loop, outsourcing AI to do the investigation (a “find evidence” button), doing the work of a digital forensics practitioner for them were further identified as inappropriate, with the general theme that AI is not a substitute for thought.

CSAM presents specific challenges related to AI:

- Minor detection. AI has the same problem as the Tanner scale in that it can identify developmental stages, which may not fully correlate to victim age. While it can be useful for triage (presenting the most likely images for review first), it should not be used as a definitive answer related to age,
- Current algorithms do not generally take into account background. The same individual in a bedroom with features consistent with a middle school student has a different prior probability than one in a neutral environment (e.g., a living room).
- We throw away information by not using all available information for AI classification. For example, if there are a series of photographs of the same individual that were taken at the same time, current models treat them as discrete images rather than trying to classify them as a series.
- AI models may have inaccurate ground truth datasets that they were trained on. Many general models ingest images from Instagram or Flickr that are self-labeled. There are unknown error rates on how many of those are mislabeled.

A particular topic of concern was the increasing trend of people generating and sharing inaccurate AI slop “digital forensics posters” on social media. Mistakes and misinformation is being disseminated on social media in “digital forensics posters” produced by inexperienced practitioners using AI models that are not grounded in digital forensics expertise. Knowing what tools and techniques have been accurately vetted and sorting those out from the quantity of poor quality, high volume noise is becoming more difficult.

There was some question about whether it might be okay to “Ask one AI to check another.” This can look like verification, produces a second opinion, and generates an audit trail, but does not actually provide independent evidence. Such cross validation may not catch errors because models sharing training data, architecture or prompt framing fail together, and a confident second model laundering a first model's mistake is worse than no check at all, because it manufactures confidence.

This is becoming worse as models are being trained on content created *by other models*. This leads to circular reporting and corroborated “reference sources” may actually have been generated by another AI tool and suffer from the same issues.

Principles

AI is appropriate when used responsibly as a tool, with people in control, and the output is verified for accuracy and completeness. AI is not a substitute for expert knowledge and critical thinking. Critically evaluate AI output, don't rubber-stamp or treat as factual. Use it for lead purposes, and make it source and justify its output.

5) AI in Court (Led by Jessica Hyde, Hexordia)

Courts are grappling with the use of AI. In one case, a witness was forced to produce their prompts in preparation for court. In LA Superior Court AI is being used to help judges understand expert witness testimony. According to a [post by an examiner](#), a judge used AI to assess answers from and ask questions of a digital forensics expert. The participants' concerns about AI in digital forensics included insufficient accuracy, explainability, and reproducibility

The Daubert standard has requirements for testability, error rate, peer review, standards/controls, and general acceptance. According to the [Digital Investigation Techniques: A NIST Scientific Foundation Review](#) key takeaway 4.4 “Digital processes tend to have systematic errors rather than random errors. Therefore, an error mitigation analysis provides more information and is the correct way to manage uncertainty. An error rate is only useful where there are random errors.” However, the use of AI in digital forensics processing introduces the possibility of random errors. As such, an error rate needs to be considered when using AI-enabled technologies in the realm of a Daubert challenge.

Errors that have been observed by members of the audience:

- Telling me a generated image did not have content that could be visually seen in the image
- Estimating the age of a real child as an adult - CSAM classifier use case
- Disobeyed orders and broke out of its sandbox
- Regex that would've missed half the expected results
- Said I was right when I knew I was wrong
- Said I was wrong when I knew I was right
- Confirmation bias (just like human)
- Wrong face reconstruction
- Wiping a production database
- Told me data in a LevelDB was a SQLite db
- Wrote SQL joining a billion row table with itself
- Using non-existent libraries
- Providing wrong sources
- Hallucinated 50 findings for a Red Team report

- Make up a SWGDE publication
- Over fitting data

This encourages the need for responsible use of AI in ways that the digital forensics practitioner can clearly articulate testing in order to meet the threshold of a Daubert challenge. The practitioner testifying to the events must be able to articulate the error rate, perform testing, and utilize AI in accordance with standards and controls.

The use of AI in courts for digital evidence is being looked at the highest levels including via the Proposed U.S. Federal Rule of Evidence (FRE) 707. The proposed rule has gone back to committee after an open comment period in early 2026 that resulted in over 300 pages of response. While we do not know what the next proposed FRE 707 will bring, it is clear that this will have an impact on the role of AI in the courtroom.

One way that digital forensics practitioners may deal with these changes is to focus on the reliability of the data that is surfaced by a tool as opposed to the reliability of the model. The digital forensics practitioner is still responsible for providing the proper interpretation of the data, noting that the existence of data in and of itself does not provide meaning. The results of an AI model may be treated as a lead or a starting point for traditional digital forensic analysis.

Another potential challenge is that of attribution. As users of digital systems incorporate AI, and particularly agentic AI, into their processes, digital forensics practitioners will need to determine if a trace was left by the user or an AI acting as the user. As Ovie Carroll recently wrote we may soon be facing the [Some AI Did It \(SAIDI\) Defense](#).

6) Engineering Around AI (Cameron Kelley — Mobasi, Edited with AI assistance)

The session advanced to GenAI guardrails. A guardrail is how a system is designed and engineered to prevent incorrect or improper use. No single guardrail is sufficient for AI agents - a combination of techniques is necessary to prevent problems, and a digital forensics practitioner must stay in control.

The audience was asked what guardrails they would like to see around GenAI, and the top results are ranked in order of popularity in the following table.

Guardrail category	Rank
Containment and access control	1
Provenance and grounding of output	2
Transparency and reproducibility	3
Clear inferences and multiple hypotheses	4

Design principles for the guardrails themselves	5
---	---

Table 3: Categorized and ranked audience responses to poll question “*What guardrails would you like to see?*” (AI summarized, human verified, original responses in Appendix A)

The guardrail that digital forensics practitioners asked for the most was *containment*. The responses included:

- *the ability to record, inspect, and optionally approve/reject any external (off-system) communications*
- *controlled access to external resources while in a sandbox*
- *a whitelist for tools/functionality it can use*
- *strict scoping of data access*
- *specified limitations on dates and locations in storage.*

These requirements describe a system whose egress is mediated and auditable by default, not just trusting a model to behave as expected. The approve/reject requirement is the strongest form of observability and constraint: a digital forensics practitioner sees each outbound call before it happens and can block if needed.

The ability to explain and reproduce AI results was another requirement expressed by participants alongside always linking directly to data sources and supporting knowledge resources, and preventing it from inventing citations. This relates to the poll response “*Ontology based LLMs, integrated with Agents, defense in depth codified at two different layers*” to constrain AI generation so that outputs are expressed in defined classes and relations rather than free text, with enforcement at more than one layer so a single failure is not a total failure. Approaches to accomplish these requirements were discussed in the final topic of the session dealing with AI grounding using structured validated knowledge bases and neurosymbolic AI.

Distinguishing between facts and inferences restates the earlier poll’s emphasis on statistical inferences not being deductions, elevating this to a design requirement rather than just a warning.

The session proposed a set of guardrails based on findings from early development of LLM-based tools by AI engineers in tandem with practitioners. The proposed guardrails take into account the current state and capabilities of large language models, as well as the manner in which they are applied during a digital forensics examination (e.g. the user interface, standard operating procedures, human-in-the-loop requirements). Principles for safe, reliable GenAI use were presented along with the digital forensics examiner’s responsibility to critically evaluate results.

Three categories of mistakes were discussed. Each category shows where design and engineering is necessary to improve outcomes:

1. How AI queries data
2. How AI presents findings

3. How humans use AI output

Problems with how AI queries data. These mistakes occur before the digital forensics practitioner sees an answer. They are therefore the most difficult to see.

- **Truncated data:** A model can only read a limited quantity of data at one time. When a query returns more than that limit, the model only sees a part of the result, and may not provide a statement that it stopped. A model also tends to stop when it finds something interesting, not when it has examined everything. The result is that the digital forensics practitioner gets an answer that looks complete but is based on a fraction of the evidence.
- **Wrong artifact:** A model may search artifacts that it is most familiar with, rather than ones that are most relevant. It goes to the well-documented artifacts that are common in its training data, such as browser history and message databases. It can miss the less common artifact that answers the question. The result is a confident answer from the wrong source.
- **Misread filters:** When a question is ambiguous, a model may make assumptions without asking the examiner. A common example is the time zone. A model that assumes UTC for a question about “the evening of March 3” can silently examine the wrong window of time. The same problem occurs with date ranges, accounts, and devices.
- **Reinventing the wheel:** Models are good at writing code, so they frequently write a new parser for an artifact when a validated forensic tool already exists. The result is that an examination loses the trusted testing and validation of existing forensic tools, and may introduce silent parsing or interpretation errors. The severity of this issue increases for data types with complex parsing or interpretation requirements.

Problems with how AI presents findings. These mistakes occur when the model converts data into a description to the examiner.

- **Misunderstanding the meaning of artifacts:** A model can reinterpret a “created at” timestamp as “this event occurred at” that time. A digital forensics practitioner knows that the timestamp can record a file copy, a data synchronization, a restore from backup, or a database migration. The label of a field is not its forensic meaning.
- **Citing absence as evidence:** A model may interpret “not found” as “did not happen.” The absence of a record in the data that was searched does not show that an event did not occur. The record could be deleted, stored in a different location, not parsed by the tool, or outside the search. A failed query is also not a negative finding.
- **Omitting scope and coverage:** The scope of what was searched, such as the exact period of time or set of artifacts, may be left out of findings. Without this information, the reader cannot know whether “no relevant messages were found” refers to the full extraction or to one keyword search of one database.

- **Treating a device as a person:** A model may implicitly link observed findings on a device to the custodian of that device. Digital evidence records the actions of devices and accounts. To attribute those actions to a person is a separate inference, which often requires other information and always belongs to the examiner.
- **Drawing conclusions instead of presenting findings:** A model is trained to give an answer, and may tend to prioritize its interpretations and conclusions rather than present raw findings to the digital forensics practitioner for interpretation

Problems with how humans use the output. A system can do everything above correctly and still fail at the last step.

- **Not validating findings in forensic tools:** A citation makes verification possible, but it is not verification. Output with many citations looks rigorous, and that appearance makes it less probable that the reader will check it.
- **Jumping straight to the conclusion:** The digital forensics practitioner may skip over findings and read only the conclusion. This is a natural response to a fluent, confident summary, especially under time pressure. It is also the deskilling concern from the opening poll, seen at the level of a single report.

Eight guardrails were proposed to combat these challenges. Illustrative good and bad examples of each guardrail were presented and discussed.

Guardrail 01: *AI must log every action taken.* Forensic practice requires that a second digital forensics practitioner can follow the same steps and get the same result. A summary that a model gives of its own work is not a record. The system must record each action taken with its parameters, its input, and its output. The record must also identify the specific model that was used. With this record, a digital forensics practitioner can reconstruct how each result was obtained.

Good	Bad
Which tool ran, with which parameters, on which input, producing which output.	"The LLM looked at X and told me Y."

Guardrail 02: *Let AI drive known forensic tools.* The discipline already has a method to establish trust in tools: testing, validation, and community review. When a LLM operates a validated tool and queries its output, the result has the known error profile of that tool. When the AI writes its own parser during an investigation, the result depends on code that nobody has tested. Where possible, AI should decide which tool to run and what to ask of its output, not replace the tool.

Good	Bad

iLEAPP tool parsed the biome, and LLM queries the iLEAPP parsed table.	The model wrote its own parser for the biome, on the fly, unvalidated.
--	--

Guardrail 03: *AI must separate findings from interpretation.* When findings and interpretation are mixed into one response, the reader cannot tell which sentences are records and which are inferences. When they are in separate sections, the digital forensics practitioner can verify the findings one at a time and then decide whether to accept, change, or reject the interpretation.

Good	Bad
Two distinct sections. One for evidence citations, the other for interpretation.	One paragraph of findings & interpretations nobody can pull apart later.

Guardrail 04: *AI must link back to the source artifact.* A finding that cannot be traced to its source cannot be verified. A specific reference gives the file, the path on the original device, and the table and row or the line. A specific reference is also the most practical defense against fabricated or hallucinated output. An invented record fails immediately when the digital forensics practitioner opens the source location.

Good	Bad
We know the file, the artifact, and the line where it was located. Specific.	Unspecific.

Guardrail 05: *AI must explain how to validate results using other tools.* Verification that is difficult does not get done. When the output includes the guidance to confirm a finding in an independent tool, the correct behavior becomes the easy behavior. An example is “open this database, run this query, view this row.” This guardrail is directed at the human’s use of AI outputs more than the AI itself. It also keeps the examiner’s skills in use, which answers part of the deskilling concern.

Good	Bad
Guidance on how to verify results, e.g., Open this SQL database, run this query.	Nothing.

Guardrail 06: *AI must qualify conclusions.* A model writes with the same confidence when it is right and when it is wrong, and a reader can mistake that confidence for accuracy. Qualified language keeps the statement inside what the evidence supports. It also keeps alternative explanations visible. The statement “Jill was at the scene” is a conclusion about a person. The evidence can only show that a device that is associated with Jill recorded a location.

Good	Bad
------	-----

The observations are much stronger given the claim that Jill's device was at the crime scene rather than some other location.	"Jill was at the scene of the crime."
---	---------------------------------------

Guardrail 07: *AI must be honest about coverage.* Each answer must state what was examined, what failed, and what was skipped. Without this statement, the digital forensics practitioner cannot tell the difference between "nothing was there" and "the AI did not look fully." If a keyword search returns 500 records and the model reads 50, the system must say so. This guardrail is the direct defense against truncated data and against absence that is cited as evidence.

Good	Bad
What was examined, what failed, what was skipped.	"I read all the chats..." when actually it ran three keyword searches.

Guardrail 08: *AI must run in a real sandbox.* Instructions in a prompt are requests, and a model can ignore a request. One of the errors that the participants reported was an AI that "disobeyed orders and broke out of its sandbox." A real sandbox enforces its limits at the system level. The evidence is read-only, the AI can write only to a separate work area, and no network route goes out. The sandbox holds whether the model follows its instructions or not. This ensures that AI can be used with the necessary data security, privacy, and compliance controls during a forensic examination.

Good	Bad
Evidence only. Read only. Write blocked. No route out.	An agent running web searches mid-investigation.

Additional discussion went beyond the eight guardrails into common components of engineered AI systems. A guardrail controls one behavior, but the safety and accuracy of an AI system also depend on the other engineered components around the model, and on how that system is developed and tested.

Sandboxing: A sandbox applies system-level controls to an AI process to create secure boundaries on behavior. The file system part controls what the AI can read and change. It ensures that AI can interact only with evidence that is in scope for the case, it cannot modify that evidence, and its own output goes to a separate location. The network part controls what can enter and leave. Without network control, case data can leave the environment improperly, and material from outside can enter the analysis. For example, a model that does a web search in the middle of an investigation brings in content that is not evidence. Models tend to treat web results as authoritative, and the text that results is fluent, so the bias is difficult to see. Outbound communication must be blocked by default, and each exception must be recorded.

Evaluations: An evaluation is a repeatable test of the full AI system on tasks that have known correct answers. It is the AI equivalent of tool validation, and it is the practical method to produce the performance data that a Daubert challenge will ask for. Evaluations answer questions that intuition cannot. How frequently does the system follow a specified guardrail? What changes when the model, the prompt, or a tool is changed? How does AI-assisted work compare with human-only work? The tasks must be verifiable, with ground truth that is documented, and they must represent the conditions of real casework. It is possible to add a second “checker” prompt that reviews output against the rules. As noted in section 4.2, one AI that checks another is not independent verification. The output of a checker is one more signal to measure, not a guarantee.

Temperature: A common suggestion is to reduce the temperature of a model, which reduces the randomness of its output. This does not necessarily make a model safe for forensic use. A lower temperature makes output more consistent, but it does not make output more correct. A wrong answer that is repeated each time is still wrong, and the consistency can give a false impression of reliability. In practice, hosted models at low temperature still do not always produce the same output. The goal is not identical text on each run. The goal is a result that is reproducible in the forensic sense: a different examiner, with different tools, gets the same finding from the same source data.

Switching models: Guardrails that are implemented as instructions do not transfer automatically from one model to another. Models differ in how closely they follow instructions and in how they prioritize evidence. They can reach different conclusions from the same data. Smaller models usually adhere less to instructions. Each change of model, including a version update from the same vendor, is a change to the tool. The evaluations must be run again, and the audit log must record the model that was used.

Open-weight models: Models that run locally are attractive for forensics, because the evidence stays in the laboratory and the system can operate in an air-gapped environment. The open-weight frontier improves quickly, but the models that fit on hardware that a typical laboratory can buy are much smaller than the hosted frontier models. At the time of the session, these smaller models can do bounded tasks, such as writing an SQL query. They were less capable at complex reasoning and investigative work, which includes long tasks with many steps, the correct use of tools, and adherence to structured processes and guardrails during a long session. A laboratory that deploys a local model must evaluate it on its own tasks and not assume that results from a larger model apply.

Context is not forgotten: Anything that is put into the context of a model can influence its output, even after an instruction to disregard it. A model cannot forget. Examples are a case theory, the name of a suspect, a label from a prior report, or data from a different case. The engineering response is isolation, not instruction. Use a new environment for each case and each task, reset the environment between them, and control what enters the context. This is the AI form of a control that examiners already use to limit contextual bias.

Principles

Do not rely on a model to police itself.

Make verification the easiest path for the examiner.

Enforce limits in the system, record every action, and test the full system each time a component changes.

7) Grounding AI using Codified Digital Forensics Expertise (Led by Eoghan Casey, Cyber Knowledge Engineering)

The presentation and discussion of knowledge management centered around codifying digital forensics expertise in forms that support both AI reliability and human competency.

Essentially, a general-purpose language model produces fluent text by predicting plausible word sequences from patterns in its training data. It has no native concept of which forensic techniques are valid for a given device, which weaknesses occur in the real world, or which associated mitigations have been validated by qualified examiners. When asked a forensic question, an unconstrained model can produce answers that are subtly wrong, partially correct, or entirely fabricated, which is treacherous in a courtroom-facing discipline.

For reliability and verifiability purposes, digital forensics needs a blending of generative AI with explicit, structured knowledge and logic, known as neurosymbolic AI. This approach grounds AI in validated knowledge specifically designed to improve AI accuracy and help digital forensics practitioners learn from shared codified expertise. The foundation for the symbolic side of neurosymbolic AI is a common body of codified expertise, stored in machine-readable forms, institutionalized and available. Examples presented during the conference included [NIST ArtCat](#) and [SOLVE-IT](#), and the resulting improvements to accuracy and reliability were demonstrated.

These knowledge bases are designed to support learning and reasoning, and also to support decision-making and problem-solving by connecting ideas and indicating relationships between them. This structured, codified expertise includes [ontology compliant](#) knowledge graphs of precisely defined relationships between forensic techniques, failure modes, risk factors, and remediation steps, curated and validated by experienced digital forensics practitioners. Results from grounded AI can reference the specific knowledge base entries it used, improving traceability, rather than relying on a general-purpose language model's unconstrained generative ability. Participants expressed that the presented examples helped them see the benefits of ontologies for digital forensics, which they had not appreciated before.

While the common body of knowledge is stored in machine-readable formats, it is accessible to both humans and AI-enabled applications. AI applications can consume it natively through

integrations or common protocols. That same knowledge can be presented to digital forensics practitioners through applications dedicated to converting knowledge graphs and similar structures into human-readable forms. Qualified and authorized personnel can manage the common body of knowledge directly through applications designed for managing knowledge graphs and ontologies.

The practical effect of grounding in expert knowledge is that AI behaves less like a generative author and more like a disciplined laboratory assistant working from a curated case file. It is less likely to invent techniques, weaknesses, or mitigations. If the relevant details are not in the digital forensics knowledge base, it can be instructed to inform the user rather than produce a potentially incorrect answer.

Summarizing the benefits of using Neurosymbolic AI:

- Codifying validated knowledge improves accuracy and reliability of AI results
- Structuring knowledge with cognitive modeling enhances AI reasoning
- Referencing supporting sources increases AI transparency and auditability
- Versioning knowledge bases enables reproducibility of AI on specific versions
- Flagging out-of-scope queries not in knowledge base for human handling

Neurosymbolic approaches are useful for gap analysis of standard operating procedures (SOPs) to improve digital forensics practices, and for adversarial analysis of digital forensics outputs and human-generated deliverables (e.g., examination notes, expert reports) to mitigate potential weaknesses and misinterpretations before going to court.

8) Investigator/Attorney Self-Service AI Analysis of Digital Evidence

To wrap up the working session, the audience was asked about the risks of investigators/attorneys analyzing digital evidence on their own utilizing AI.

The primary concern was that investigators/attorneys do not know enough about digital forensics and AI to detect errors or explain how they reached a conclusion. They do not understand the underlying data or how a given digital artifact was created, and they do not understand the limitations of the tools being used. An untrained non-specialist using AI can have the illusion of expertise/knowledge, can believe the output is actually ground truth, not take uncertainties into consideration, and jump to a conclusion.

Where the earlier audience polls in this session treated lack of verification as carelessness, this discussion elevated the concern of investigators/attorneys accepting AI results as reliable and not realizing that verification is needed.

Statistical inference is not a deduction, but non-specialists may not see the difference. When AI reports a likely pattern alongside an established fact, the investigator/attorney is given no indication to distinguish between them. The risk is presenting an interpretation with unevaluated uncertainty as incontrovertible fact. This risk can be mitigated by guardrails discussed in Section 6 above.

A broader concern is that AI takes away opportunities for people entering digital forensics to learn and grow their knowledge, skills, and abilities to an expert level. Expertise is developed by working through problems, and a tool that supplies the answer removes opportunities for the development of expertise. The poll response “No SMEs anymore, hostages of subscriptions” is a forward looking ramification of irresponsible AI use: if a discipline stops producing SMEs, the capability does not merely atrophy, it shifts to software/AI vendors. Responsible use of AI reinforces competence of digital forensics practitioners by providing them with required knowledge and guidance to validate and evaluate results.

It is important to note that this does not negate the need for or the openness to investigators and attorneys having access to tools that conduct AI analysis of digital evidence. To the contrary, it allows for attorneys and investigators to self-discover potential items relevant to the case and then note those things to the digital forensics practitioner who can provide the context of what the existence of the data means from a digital forensics perspective. The use of AI tools by non-examiners has the opportunity for the people who have the context of the investigation to identify items that the digital forensics practitioner may not have seen as relevant to the case. The important part is that the digital forensics practitioner be consulted to interpret that meaning of an artifact, assess possible reasons that a trace could exist, and prevent potential mistakes or misunderstandings.

Principle

Appropriate use of AI is a combination of the task and the operator.

9) Key Outcomes

The following is a summary of key outcomes from this session:

- Learn to analyze traces of AI usage for digital forensics purposes.
- Treat AI tools in a similar manner to conventional digital forensics tooling including an audit trail of what was done and the ability to trace a result back to its source data (provenance, chain of evidence, transparency, auditability, explainability).
- Guardrails are essential and there needs to be a combination of techniques to prevent problems, ultimately with humans in control
- AI that is grounded using digital forensics expertise is more accurate and reliable than ungrounded AI
- Digital forensics practitioners are responsible for critically evaluating and verifying the veracity of outputs from AI-assisted and AI-generated tools. Anyone cannot push a button and blindly accept the output or pass it off to a lawyer or judge to determine whether or not it is reliable.
- Using advanced processing is helpful for lead purposes, but any forensic finding should be corroborated with information from other sources

- Evaluate the performance of AI using ground truth datasets that represent real world digital forensics conditions.
- Don't leave it to an investigator, lawyer, or judge to determine whether or not AI output is accurate and reliable for digital forensics purposes.

The following high risks were discussed during this session:

- AI generated content is becoming increasingly challenging for digital forensics. Updated training and tools are needed to help.
- AI output has uncertainty that must be treated accordingly, but can be presented confidently as factual. This increases the risk of someone taking statistics from LLMs as proof without considering inherent uncertainty.
- Inaccurate AI slop "forensics posters" on social media spreading misunderstanding and misinformation.
- Using AI to check another AI may be incorrectly perceived as verification, producing a second opinion, but does not actually provide independent evidence. Such cross validation may not catch errors because models sharing training data, architecture or prompt framing fail together, and a confident second model laundering a first model's mistake is worse than no check at all, because it manufactures confidence.
- Investigators and attorneys using AI as a substitute for digital forensics SMEs increases risk of mistakes and misinterpretations. Then commercial vendors rather than digital forensics SMEs control the quality and reliability of digital evidence used for high-stakes decisions (investigative, court, military, corporate). It is imperative that a digital forensics practitioner be a critical part of the analysis for the interpretation and validation of digital evidence.
- Guidance addressed only to forensic examiners will miss investigators and attorneys, where the risks are rising in relation to irresponsible use of AI in digital forensics. The investigators and attorneys now able to analyse evidence unaided are outside the discipline's professional structures, its training, its competency standards, its peer review, and are the least likely to encounter a policy written for practitioners.
- Cost is a concern - running an LLM against millions of records can end up with costs in the tens of thousands of dollars using naive approaches at current token costs, so run test prompts on data samples to manage expenses.

Not discussed during this session were risks associated with tampering that conceals, alters, or destroys digital evidence, and prompt injection attacks planted in evidential sources designed to undermine AI-assisted digital forensics. This is an area that would benefit from further research and safeguards.

Appendix A - Categorized audience responses to poll question "*What guardrails would you like to see?*"

Containment and access control

- Ability to record, inspect, and optionally approve/reject any external (off system) comms

- Controlled access to external resources while in a sandbox
- Strict scoping of data access
- Avoid being tainted by previous data / analysis
- Whitelist for tools/functionality it can use
- Simple to express the bounds needed
- Provide specified limitations on dates and locations in storage

Provenance and grounding of output

- Providing offsets for output evidence locations
- Always link directly to data source
- Detailed logging of activity
- Banning ambiguity from LLM development
- Prevent it from inventing citations?

Transparency and reproducibility

- Ontology based LLMs, integrated with Agents, defense in depth codified at two different layers
- Proof of exact generated methods used to process datasets and validation of no difference in generated methods between datasets
- Clarity on methodology
- Transparency in the reasoning process

Clear inferences and multiple hypotheses

- Having AI evaluate multiple hypotheses
- Decision tree for AI inference
- Distinguish between facts and inferences

Design principles for the guardrails themselves

- Don't use same set of guardrails for every task/project
- Guardrails are very necessary but don't make them overly strict or it will backfire - from experience at Project VIC
- Affirming messages, it should not change its answer to Yes if I just say "you're wrong"

Uncategorized

- Detection of adversarial techniques targeting AI-assisted analysis
- Low temperature for forensics uses for lower randomness
- Ones that are tested to make sure humans will use them appropriately
- Runtime script plus, the Rules of 3, plus LIVE.