Human-AI Collaboration for Synthetic Media Detection in Training and Operations

Laura Cassani, Michael Davinroy, Tatiana Toumbeva, Peter Bautista, Lauren Fortier, James Cook, Ashley Hart, Svitlana Volkova

Aptima, Inc.

Woburn, MA

lcassani@aptima.com, mdavinroy@aptima.com, ttoumbeva@aptima.com, pbautista@aptima.com, lfortier@aptima.com, jcook@aptima.com, ahart@aptima.com, svolkova@aptima.com,

ABSTRACT

The rapid proliferation of synthetic and manipulated media poses a growing threat to mission resilience, information operations, and public trust. Adversaries exploit advanced technologies to disseminate falsified media, undermining decision-making processes and operational effectiveness. This paper explores emerging methodologies and innovative concepts for enhancing human-AI collaboration in detecting and countering synthetic media, with a focus on applications in defense, homeland security, and broader operational contexts. Building on insights from a comprehensive evaluation of synthetic media detection systems, we highlight novel strategies for integrating human expertise with AI capabilities to advance training and operational effectiveness. The evaluation spanned over 111 tasks involving multi-modal data to assess manipulation detection and localization, generator attribution, and intent characterization. These tasks simulated real-world threats through curated datasets, enabling the identification of system strengths and areas requiring further development. Findings emphasize the complementary roles of human and AI contributors: humans excel in contextual reasoning and nuanced judgment, while AI systems provide unmatched speed and scalability. Case studies piloting these approaches in operationally relevant environments revealed promising pathways for training the workforce of the future, improving decision-making under pressure, and fostering trust in human-machine teaming. We further explore challenges in workflow integration, trust calibration, and human-centric design, offering actionable recommendations for advancing training scenarios and simulation environments. This work underscores the transformative potential of human-AI collaboration in tackling emerging threats and demonstrates how modeling and simulation can accelerate the development of groundbreaking capabilities to secure the information environment.

ABOUT THE AUTHORS

Laura Cassani serves as the Deputy Director of the Intelligent Performance Analytics Division at Aptima, Inc., where she spearheads initiatives at the forefront of artificial intelligence research to enhance both human and machine performance. Ms. Cassani has served as Principal Investigator on numerous high-impact research and development efforts for the Department of Defense (DoD), including projects with DARPA, the Office of Naval Research, Marine Corps Systems Command, Air Force Research Laboratory, and the Office of the Secretary of Defense. Her recent work focuses on the application of generative AI for defense and national security missions. She led as PI the test and evaluation technical area for DARPA's Semantic Forensics (SemaFor) program, developing methodologies to assess AI algorithms that detect, attribute, and characterize semantic manipulation in synthetic multimedia. She continues to oversee the transition of SemaFor capabilities through the ULRI Digital Safety Research Institute (DSRI), building national research capacity in synthetic media detection. Ms. Cassani holds an MA in security studies from Georgetown University and a BA in international relations from Boston University.

Michael Davinroy is an AI Engineer in the Intelligence Performance Analytics Division at Aptima, Inc. He uses his experience in machine learning, low-level systems programming, and algorithmic game theory to design, implement, and evaluate new highly efficient intelligent agents that act in uncertain environments. Mr. Davinroy received an MS in computer science from Northeastern University and a BA with honors in computer science from Swarthmore College.

Tatiana Toumbeva is a Senior Scientist and Team Lead in the Training, Learning, and Readiness Division at Aptima, Inc. with expertise the areas of learning and training, leadership, selection and assessment, and organizational change. Dr. Toumbeva applies learning principles and systems design to develop adaptive training capabilities in complex instructional and operational settings. She holds a PhD in Industrial/ Organizational Psychology from Bowling Green State University, a Global Professional in Human Resources (GPHR) certification from HRCI, a change management practitioner certification from Prosci, an MA in mental health from Boston University, and a BS in biology from University of South Florida.

Peter Bautista is an Innovation Lead in the Intelligence Performance Analytics Division at Aptima, Inc. He has a diverse skill set in the realm of data science, machine learning, and natural language processing, with a background in both academic and industry settings. His expertise spans across developing innovative solutions for large language model (LLM) applications and evaluation, visualization, and management, including the implementation of agentic workflows and reinforcement learning frameworks. Mr. Bautista received an MS in computer science from California State University, Fullerton, and a BS in physics from University of California, Riverside.

Lauren Fortier is a Research Engineer in the Intelligent Performance Analytics Division at Aptima, Inc., specializing in the design, evaluation, and implementation of human-centered automation and decision-support tools, as well as in experimental design and statistical analysis. Her experience in machine learning and applied statistics is honed across her work in data analysis, visualization, graph networks, and reinforcement learning. Ms. Fortier received an MS in statistical practice from Boston University, and a BS in psychology from Western New England University.

James Cook is a Research Engineer in the Intelligent Performance Analytics Division at Aptima, Inc. with expertise in the design, evaluation, and implementation of machine learning tools, experiment design, and statistical analysis, and specializing in generative agents and advanced machine learning applications. M. Cook has an MS in data science from Northeastern University and a BS in business analytics and information management from Western New England University.

Ashley Hart is a Computer Science Ph.D. student at the University of Florida and an AI Engineer intern in the Intelligence Performance Analytics Division at Aptima, Inc. Her research interests include artificial intelligence, game development, simulations, and computer science education. Ms. Hart is passionate about leveraging technology to create immersive, innovative, and impactful experiences.

Svitlana Volkova is Chief of AI in the Office of Science and Technology at Aptima, Inc. She spearheads the company's initiatives to develop trustworthy and human-centric compound AI systems that tackle complex real-world problems for the DoD and other government agencies. Dr. Volkova's research focuses on advancing natural language processing and machine learning techniques, with an emphasis on graph neural networks, causal inference, and multimodal frontier models. Her work has pioneered methods for AI-powered analytics to explain complex social systems and behaviors. She received her PhD in computer science from Johns Hopkins University, and MS degrees in computer science from Petro Mohyla Black Sea State University, Ukraine, and Kansas State University.

Human-AI Collaboration for Synthetic Media Detection in Training and Operations

Laura Cassani, Michael Davinroy, Tatiana Toumbeva, Peter Bautista, Lauren Fortier, James Cook, Ashley Hart, Svitlana Volkova

Aptima, Inc.

Woburn, MA

lcassani@aptima.com, mdavinroy@aptima.com, ttoumbeva@aptima.com, pbautista@aptima.com, lfortier@aptima.com, jcook@aptima.com, ahart@aptima.com, svolkova@aptima.com

INTRODUCTION

The proliferation of synthetic and manipulated media poses a growing threat to national security, operational resilience, and public trust (Volkova & Jang, 2018). Adversaries are increasingly leveraging generative artificial intelligence (AI) to inject deceptive content into information environments, undermining decision-making processes, institutional credibility, and mission readiness. These threats cut across multiple sectors—defense, homeland security, intelligence, law enforcement, and commercial domains—highlighting a systemic vulnerability to synthetic information warfare. Current countermeasures have focused primarily on advancing AI-based detection tools, with notable progress in image, audio, and video analysis (Kaur et al., 2024). Yet despite recent advances, both AI and human analysts face significant limitations when operating independently. Deep learning offers speed, scalability, and sensitivity to subtle visual or acoustic artifacts but struggles with generalization, robustness, and interpretability (Azizpour et al., 2025). Human analysts contribute contextual reasoning and ethical judgment but are constrained by cognitive overload and throughput. Research shows that unaided human detection of synthetic media performs at or near chance levels, especially when content involves unfamiliar languages or human likenesses (Cooke et al., 2024). Effective defense against synthetic media requires a paradigm shift from siloed analysis to hybrid human-AI teams that combine complementary strengths. Drawing on experiments and real-world evaluations, we show how integrated systems can detect, attribute, and contextualize manipulated media more effectively than either humans or machines alone. These insights support a new class of detection architecture—compound AI—where individual AI agents can perform specific tasks such as triage, analytic orchestration, and interpretability to guide users through complex workflows. We present findings from user pilots in operational environments that inform new design, training, and workflow strategies to enhance detection while supporting usability, adaptability, and integration.

Accordingly, with this paper, we make the following contributions and outline future directions:

- Present a multi-modal evaluation framework that benchmarks human and AI performance across detection, attribution, and characterization tasks involving synthetic image, video, audio, and text content
- Introduce and synthesize findings from over 100 task-based evaluations as well as pilot studies conducted with operational users across defense, homeland security, law enforcement, and journalism domains
- Identify key strengths and limitations of standalone analytic tools and human analysts, highlighting when and how collaborative approaches yield superior outcomes
- Propose system design principles for building collaborative, interpretable, and mission-aligned detection platforms that support human decision-making under pressure
- Introduce compound AI as a novel human-AI teaming architecture, specifying technical components such as lightweight triage classifiers, workflow orchestration agents, and modality-specific interpretability modules
- Provide tangible defense training use cases (e.g., ISR feed injects, adversarial mission rehearsal scenarios) to demonstrate how compound AI can be embedded into training environments
- Offer practical and technical recommendations for future implementation, including containerized microservice deployment, reinforcement-learning–based orchestration, active learning pipelines, and human-AI performance metrics that capture trust calibration and time-to-insight

To our knowledge, this is the first work to connect synthetic media detection architectures directly to defense training workflows, highlighting both system design and human teaming requirements.

BACKGROUND AND RELATED WORK

Synthetic media (AI-generated or manipulated content) has rapidly shifted from a research concept to a real-world operational concern across multiple domains. Seeing is no longer believing; free tools now generate photorealistic images from text in minutes (see Figure 1), with more advanced capabilities emerging (e.g., multi-minute videos with physical interactions and detailed scenarios). Adversarial uses of generative AI include impersonation, espionage, fraud, and exploitation, and their real-world impacts are increasingly evident (Kalpokas, 2021; Sayler & Harris, 2023). These capabilities are transforming influence operations, making synthetic media detection a mission-critical need for defense and security agencies. Novel detection technologies can identify deepfakes and manipulated artifacts, but most focus on specific modalities or cues that can be bypassed with laundering techniques (Gu et al., 2025). Meanwhile, evolving generator models (e.g., diffusion, transformers for video synthesis) are outpacing detectors' ability to generalize or remain resilient under real-world conditions (e.g., Zhang et al., 2022).

Figure 1. Image Generated Using Flux Dev 1.1 Pro Using Prompt from GPT4

Research comparing human and machine capabilities reveals tradeoffs in both approaches. AI systems trained on large datasets can outperform humans in specific detection tasks (Groh et al., 2021) but often lack transparency and contextual awareness (Hassija et al., 2024). Humans, while adept at interpreting high-level meaning and intent, perform inconsistently and are easily deceived by realistic media, especially under time pressure or cognitive load (Cooke et al., 2024; Uchendu et al., 2023). Recent findings by Naber et al. (2024) further confirm that both human and machine assessments fall short when analyzing subtle propagandistic cues, reinforcing the need for integrated approaches. Despite this limitation, current synthetic media detection systems remain largely siloed, either automated or human-in-the-loop, without truly collaborative or adaptive interaction. The literature on human-AI teaming has noted advancements in other mission-critical domains (e.g., tactical and combat operations, autonomous systems; Feldman et al., 2024; Hagos et al., 2024) but is still nascent in the context of synthetic media analysis.

A growing body of work emphasizes that trust between human and AI partners is essential for successful collaboration, particularly in high-stakes environments. Diedrich et al. (2023) propose a co-learning and trustworthiness framework rooted in operational contexts like mission command. Rather than treating trust as a subjective belief, their model operationalizes it as a function of behaviorally observable microexperiences: individual interactions that build or erode confidence over time. This framework highlights that trust in AI must be earned through repeated, value-aligned interactions that demonstrate competence, consistency, and shared intent. Importantly, AI systems must not only be explainable but also adapt based on human input and feedback. Similarly, human users must develop calibrated expectations and competencies to use AI outputs effectively. Arrieta et al. (2020) argue that explainable AI (XAI) and mutual feedback loops are foundational to building trust but alone are insufficient without co-adaptive system design.

Despite rising awareness of synthetic media risks, few efforts have examined how to train, evaluate, and design human-AI systems specifically for this domain. Existing research focuses heavily on algorithmic benchmarking (Mirsky & Lee, 2020), often overlooking the cognitive, organizational, and cultural aspects required for operational integration. This paper addresses that gap by introducing findings from operational pilots and over 100 task-based evaluations that assess analytic performance, human-AI trust dynamics, and the need for integrated, workflow-aware systems. These findings lay the foundation for compound AI: an architectural approach that links performance to usability, training, and orchestration to support real-world, mission-aligned media detection.

EVALUATION FRAMEWORK

To enable mission-ready, trustworthy synthetic media detection systems, a robust evaluation framework is essential, not only for benchmarking performance but also for revealing where automation excels, where it fails, and how humans and AI can most effectively team in operational contexts. Over the past several years, the Defense Advanced Research Projects Agency (DARPA) Semantic Forensics (SemaFor) program has contributed to the development of large-scale evaluation methodologies for synthetic media analytics. These evaluations span diverse media types (e.g., image, audio, video, text), manipulation techniques, and adversarial tactics and have helped lay the groundwork for integrating AI-based detection technologies into operational workflows to counter rapidly evolving synthetic threats. Our evaluation framework was designed to rigorously test analytic systems across modality, manipulation type, and interpretive complexity. By combining algorithmic benchmarks, human baselines, and operational pilot feedback, we identified performance thresholds, failure points, and design patterns to guide the future of collaborative detection systems. We structured evaluation tasks around three interrelated analytic functions—Detection, Attribution, and Characterization—each reflecting increasing complexity and interpretative demand:

- **Detection**: Has the media been manipulated? A foundational task for filtering large volumes of digital content
- **Attribution**: Who or what generated the manipulation? Identifies specific generative models or actors
- **Characterization**: What is the manipulation's intent or potential impact? Infers meaning, context, and significance beyond technical indicators

Evaluation Methodology and Metric Innovation

Performance was primarily assessed using the following metrics: (1) Probability of Detection (p(D)) – Proportion of correctly identified manipulated instances, and (2) Balanced Accuracy (bACC) – Used to address label imbalance and reflect overall performance across both positive and negative cases. Evaluation tasks spanned over 100 scenarios across image, video, audio, and text modalities, and included advanced manipulation techniques such as diffusion-generated content, recontextualized propaganda, and cross-modal inconsistencies. Human baselines were collected from over 300 participants (Naber et al., 2024), providing essential reference points for interpreting algorithmic performance and understanding teaming opportunities.

Evaluation also extended into real-world settings. We conducted a series of pilot studies and experimental sessions with operational users across defense, homeland security, commercial, and law enforcement contexts. These engagements placed early system prototypes into realistic workflows to better understand how users interacted with AI-enabled detection tools under operational conditions. Feedback from these sessions centered on three critical dimensions of system design and human-AI collaboration:

- **Trust and usability**: Focused on how users evaluated analytic outputs, balancing explanation clarity with system performance.
- **Decision support**: Explored how systems assist—not replace—human judgment by surfacing uncertainty and offering structured cues and documentation to support decisions under varying time pressures and stakes.
- **Workflow integration**: Effective tools must move beyond isolated detectors to become embedded in broader decision-making processes—not just flagging manipulations but also accelerating insight.

Analytic Performance Across Modalities

The evaluation revealed strengths in current analytic systems, particularly in high-precision tasks where signal-level anomalies were well defined (see Table 1). Analytics achieved impressive results in identifying AI-generated images. However, these same systems showed notable performance degradation on specific manipulation types, especially paste splicing, where detection dropped to a p(D) of 0.47. This variation highlights an important limitation: analytic performance is highly sensitive to the type and subtlety of manipulation, requiring robust generalization strategies that

go beyond narrow training data. Analytic performance for synthetic audio detection was also consistently high, demonstrating strong reliability across varied conditions. In contrast, synthetic video detection proved more challenging. Performance declined when evaluating content generated by newer models, with general deepfake detection averaging a p(D) of 0.71. These results underscore the need for more robust, generalizable video analytics capable of adapting to evolving manipulation techniques. Characterization tasks (e.g., identifying propaganda tactics or inferring the intended target audience), presented an interesting avenue of research. Some analytics achieved high characterization scores when assessing binary labels like audience category. More nuanced tasks involving semantic recontextualization or intent interpretation remain difficult for automated systems. Multi-modal detection, which integrates cues across media types (e.g., video and audio, or image and text), posed the greatest challenge. Persistent inconsistencies across cross-modal signals were difficult for systems to identify reliably, signaling a need for next-generation models capable of learning richer inter-modal relationships. Consistent underperformance on certain thematic clusters suggests underlying bias or blind spots in current model architectures or training datasets.

Table 1. Analytic Performance Evaluation Highlights

Modality / Task	Performance Summary	Key Challenges
Synthetic Image Detection	High precision in detecting well-defined anomalies	Significant drop in detecting subtle manipulations like paste splicing
Synthetic Audio Detection	Consistently high and reliable across varied conditions	Lower performance in laundered audio and certain generators
Synthetic Video Detection	Moderate performance; degraded with newer content and models	Difficulty adapting to evolving manipulation or fully synthetic generation techniques
Characterization Tasks	Strong performance on simple/binary labels (e.g., audience category)	Poor performance on nuanced tasks (e.g., semantic recontextualization, intent interpretation)
Multi-modal Detection	Most challenging; inconsistencies in cross-modal signal integration	Inability to learn rich inter-modal relationships; underperformance on specific thematic clusters

Human Performance and Implications for Teaming

To complement the algorithmic evaluations, a human baseline study was conducted with over 300 participants. Human performance in synthetic media detection varied by task and modality, generally falling short of analytic systems but revealing areas of relative strength. For image manipulations, humans were accurate about two-thirds of the time. In detecting AI-generated text, the human baseline performed just below chance (48%). Tasks requiring contextual judgment, such as propaganda tactic characterization, showed mixed results, with analytics generally outperforming humans. Performance also diverged in detecting semantic inconsistencies across news headlines and images. Together, these results show that while analytics outperform humans in most tasks, human insight remains valuable, particularly for subjective or context-dependent judgments. These complementary strengths underscore the importance of designing systems that support human-AI collaboration, rather than relying exclusively on either human or machine capabilities. Instead of trying to replace human judgment, detection systems should be designed to support it such as by surfacing anomalies, flagging uncertainty, and providing interpretable evidence that enables informed decision-making. To examine how these findings translate into real-world performance, we turn to operational pilot studies conducted across defense, law enforcement, and media contexts.

OPERATIONAL CASE STUDIES

To explore real-world human-AI collaboration in synthetic media detection, we conducted pilot studies and field evaluations across diverse operational domains including the US DoD, federal and international law enforcement agencies, and commercial media outlets. These efforts went beyond controlled environments to examine performance under operational stress: time pressure, high ambiguity, variable user expertise, and mission-critical stakes. In defense and homeland security, tools supported information operations and mission assurance. Law enforcement applications prioritized forensic review, particularly identifying AI-generated child sexual abuse material (CSAM). Commercial pilots explored newsroom integration, assessing how detection capabilities could support journalists in verifying viral or manipulated content. Across all settings, explainability emerged as the top requirement, surpassing accuracy or speed. Users consistently wanted to know what a tool detected, how it worked, and under what conditions it could be trusted. Visual cues like manipulation heatmaps or region highlights were helpful only when paired with clear, natural language explanations. Users sought analytic "profiles"—succinct descriptions of a tool’s purpose, strengths,

limitations, and reliability boundaries. Another insight was the limitation of simple analytic fusion. Early designs combined outputs into a single confidence score, which confused users, especially when analytics disagreed. Users preferred disaggregated results, even when conflicting, if context and rationale were provided. Exposing disagreement generated trust and allowed users to apply judgment, rather than obscuring complexity behind a consensus score.

A persistent challenge was analytic selection. Users frequently lacked guidance on which tools to apply to which media types, leading to ineffective "run-all" approaches. This underscored the need for intelligent orchestration: an AI-enabled triage layer that evaluates inputs, identifies media features, and recommends appropriate analytics aligned with content and mission needs. For example, such a system could skip facial landmark models for images without faces or prioritize characterization in CSAM workflows. Triage was not merely a technical need—it was essential for operational usability and cognitive load management. User needs also varied by context. Journalists favored fast, low-friction tools for real-time decisions. Law enforcement and intelligence users required auditability and traceability of evidence, with the ability to annotate findings and review decision paths. Despite these differences, a unifying theme emerged: trust in AI is dynamic, shaped by experience and explanation. Systems that supported transparent, explainable interactions fostered calibrated reliance; those that obfuscated logic risked disengagement or blind trust.

These insights have direct implications for training and simulation. Teaching tool operation is necessary but insufficient. Users must learn how to interpret outputs, navigate uncertainty, resolve conflicts between analytics, and challenge system recommendations when warranted. Simulations should expose users to ambiguous, adversarial, or conflicting media to build interpretive resilience and foster co-adaptation between user and system. Ultimately, these operational studies demonstrate that successful human-AI teaming in synthetic media detection depends not only on advanced analytics but also workflows and interfaces designed for collaboration, transparency, and trust. Findings from the operational case studies (summarized in Table 2) motivate a new system architecture we term compound AI. As described in the next section, compound AI emphasizes triage, analytic sequencing, and explanation to support human-AI teaming. This emerging model underscores the need to design training, workflows, and simulation environments that explicitly reflect analytic affordances and limitations.

Table 2. Summary of Operational Case Study Insights

Domain	Role of AI	Human-AI Challenges	Outcome/Key Learning
DoD	Support for information operations and mission assurance	Decision-making under time pressure; lack of guidance on analytic selection and fail points; need for trust calibration	Explainability through reporting is essential; users need interpretable outputs and dynamic workflows tailored to mission objectives
Law Enforcement	Forensic analysis, including detection of AI-generated CSAM	Auditability and traceability of analytical findings, matching analytic relevance and performance strengths to media type	Intelligent triage prevents analytic misuse (see Figure 2); users prefer disaggregated outputs with contextual explanations to support judgments
Journalism / Media	Real-time content verification and manipulation detection	Demand for rapid, low-friction tools; difficulty interpreting conflicting tool outputs; variable user expertise	Speed and clarity matter; users value succinct analytic profiles and visual/textual explanations over fused confidence scores

COMPOUND AI: A SYSTEM ARCHITECTURE FOR HUMAN-AI TEAMING

Operational evaluations and field deployments revealed a fundamental challenge in current approaches to synthetic media detection: users were not overwhelmed by the lack of analytic capabilities, but by uncertainty about how to apply them effectively. Faced with an array of options, users struggled to determine which analytics were appropriate, in what order they should be applied, and how to interpret conflicting outputs. This confusion hindered timely decisions and, in some cases, led to either over-reliance on tools or a complete disregard of valid system insights. These limitations point to the need for a new architectural model—compound AI—which moves beyond bundling tools and toward building collaborative systems that actively guide users through complex detection workflows. Previous work (Zaharia et al., 2024) defines compound AI as a system composed of multiple components (e.g.,

models, retrievers, and tools) working together to solve complex tasks. Although useful, this definition does not fully capture the challenges of collaborative, high-stakes environments. We extend and reframe this concept as a human-AI teaming architecture, where the analytic platform functions not as a static toolbox, but as an adaptive teammate that supports decision-making through three core capabilities—intelligent triage, workflow-aware orchestration, and layered interpretability. At the technical level, these correspond to lightweight modality classifiers and pre-screening filters that can rapidly identify content features; orchestration agents that learn to sequence tools dynamically; and explanation modules that pair detector outputs with modality-specific interpretability techniques. Together, these elements enable the system to guide users through the complexities of synthetic media detection, reducing cognitive load and improving trust, efficiency, and decision quality.

At the front end, *intelligent triage* analyzes the input media, identifying features such as modality, presence of faces, manipulation signatures, and other distinguishing characteristics. Based on this assessment, the system recommends the most relevant analytic tools, ensuring that users apply capabilities aligned with both the content and their objectives. For instance, rather than applying a facial landmark model to an image without any faces, the system might recommend general-purpose manipulation detection or semantic characterization tools more suited to the task at hand. Beyond selection, users also need help sequencing tools appropriately. Compound AI supports *workflow-aware orchestration*—an adaptive capability that recommends not just which tools to use, but in what order. Detection tasks vary in complexity and intent: some require fast triage and filtering, while others demand deeper forensic review or semantic understanding. Compound AI adapts tool sequences to reflect these goals, providing dynamic workflows that evolve based on user input, intermediate results, and task context. This ensures that users are not left to manually construct workflows or rely on rigid pipelines that ignore operational variability. Another critical insight from the pilots was that *interpretability* matters more than consolidation. Early attempts to fuse outputs from multiple tools into a single confidence score often obscured valuable nuance, particularly when analytic results conflicted. Users expressed a clear preference for disaggregated outputs, each accompanied by visual and textual explanations. These explanations helped users compare results, assess disagreements, and apply judgment in high-stakes decisions.

Just as importantly, the system incorporates user feedback. User annotations, confirmations, and corrections can be logged into an active learning pipeline that continuously refines triage and orchestration models, allowing the system to adapt to new manipulation types and reduce false alarms over time. Containerized, microservice-based deployment architectures make it possible to update these components modularly and embed them within operational simulators or training environments without disrupting existing workflows. Corrections, confirmations, and annotations become inputs that the system can learn from over time. This co-adaptive feedback loop helps calibrate confidence levels, refine analytic recommendations, and improve system behavior based on operational realities. It also fosters a sense of agency for users, encouraging active engagement with the system rather than passive consumption of its outputs. The underlying design of this system is grounded in several key human-centered principles. Transparency must be prioritized, with systems offering clear explanations of what was predicted, why, and with what level of confidence. Interpretability should be layered, offering high-level summaries for non-expert users and detailed rationale for advanced analysts. Interfaces must align with operational tempo, offering lightweight, rapid insights for fast-paced triage scenarios, and more granular, traceable outputs for forensic investigations. Training must go beyond tool operation to include the interpretive and collaborative dimensions of working with AI under uncertainty.

Taken together, this architecture (Figure 2) offers a transformative approach to synthetic media detection—shifting cognitive burden from user to system, supporting real-time decision-making, and building trust through transparency and adaptability. It integrates three core capabilities—intelligent triage, workflow-aware orchestration, and layered interpretability—into a co-adaptive feedback loop with human analysts. The triage layer employs lightweight classifiers (e.g., shallow CNNs, embedding-based metadata parsers) to rapidly identify modality and manipulation signatures. The orchestration layer sequences detectors dynamically, using reinforcement learning or rule-based policies to optimize analytic workflows for mission context. The interpretability layer provides modality-specific explanations, with outputs presented in both visual and natural language forms. Analyst inputs (annotations, overrides, confirmations) feed an active learning pipeline that refines thresholds, retrains policies, and improves detector generalization over time. This design shifts cognitive burden from the user to the system, embedding transparent and adaptive support directly into operational workflows such as ISR review, information operations training, and real-time content verification. As synthetic media threats grow in scale and sophistication, such architectures will enable mission-aligned, resilient human-AI teams that can operate effectively in complex, high-stakes environments.

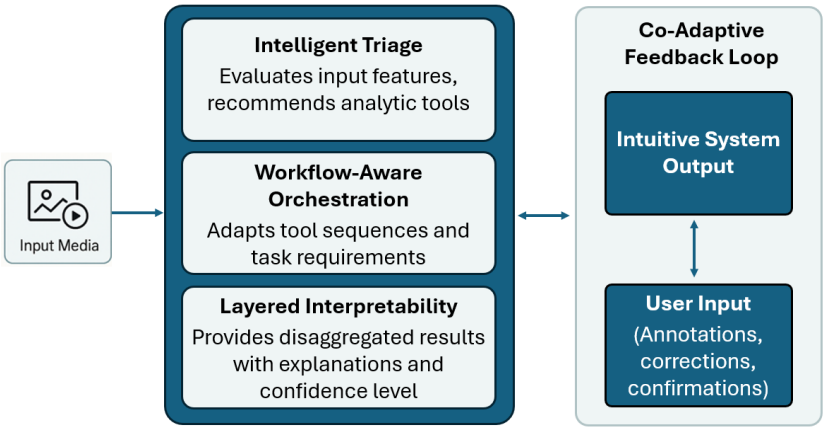

Figure 2. Compound AI System Architecture for Human-AI Teaming

FUTURE RECOMMENDATIONS

As synthetic media threats continue to evolve in scale, realism, and impact, the need for more intelligent, transparent, and collaborative detection systems is clear. Findings from our evaluations and operational pilots underscore a critical shift in system architecture and user interface design—from static toolkits to adaptive, decision-support teammates. The path forward lies in the development and deployment of Compound AI systems that are not only technically capable but also operationally attuned to the complexities of human-AI teaming in high-stakes environments. Future detection platforms must treat triage, orchestration, and interpretability as core design functions—not auxiliary features. They should dynamically assess input media and route it to the most appropriate analytic tools, intelligently sequence those tools based on user goals and task structure, and deliver layered explanations that clarify what was detected, how it was determined, and why it matters. Technically, this requires an architecture that blends fast, low-cost feature extractors for triage (e.g., modality classifiers trained on metadata and shallow embeddings), orchestration policies that use reinforcement learning or rule-based sequencing to manage analytic pipelines, and interpretability modules that generate both visual overlays and natural language rationales tailored to the specific modality. Just as importantly, they must learn from user interactions, adapting recommendations and outputs over time through transparent, meaningful feedback loops. This can be accomplished by integrating analyst feedback into retraining cycles through online learning, using annotation logs to recalibrate model confidence thresholds, and aligning system behavior with mission priorities by updating orchestration policies as adversarial manipulation techniques evolve.

To support this next generation of collaborative systems, we recommend several human-centric design priorities:

1. **Calibrate Trust, Do Not Assume It.** Trust in AI systems should be actively cultivated, not passively expected. Users must be able to see how a system reached its conclusions, understand its limitations, and know when and where it is likely to fail. Trust should emerge from transparency, predictability, and repeated, explainable interactions, and not from blind confidence in algorithmic authority.
2. **Design for Role and Operational Tempo.** Interfaces must reflect the tempo and cognitive demands of the operational environment. In fast-paced triage scenarios, users benefit from lightweight overlays, clear confidence indicators, and rapid insights. In investigative or forensic contexts, the same users may require traceable outputs, detailed rationales, and tools to annotate and archive findings. System design should be flexible to accommodate a range of needs, aligning form and function to user roles and mission phases.
3. **Make Interpretability Layered and Actionable.** Effective interpretation goes beyond visualizations. It requires explanation. Users need access to tiered insights: from quick, high-level summaries (e.g., “possible manipulation detected in face region”) to deeper technical detail (e.g., model performance boundaries, false positive rates, training domain metadata). These explanations must be matched to the user’s expertise, decision context, and time constraints.
4. **Build Feedback Loops for Mutual Learning and Co-Adaptation.** True collaboration requires systems to adapt to their users and vice versa. Interfaces should make it easy to provide feedback, approve or correct outputs, and annotate decisions. These user interactions should be used to retrain prioritization models,

calibrate output confidence, and iteratively improve recommendations. In turn, users should be able to see how the system evolves and responds to their input, reinforcing a co-adaptive learning relationship.

5. **Prioritize Workflow Integration Over Feature Insertion.** Introducing advanced AI capabilities into mission environments should never disrupt existing workflows. Systems must be built in partnership with end-users, reflecting real-world decision points, escalation protocols, and collaboration practices. The most successful tools will be those that feel intuitive, purposeful, and embedded such that they augment, rather than replace, the human element in critical workflows.

These recommendations suggest a dual trajectory for future research: advancing human-centered design principles while also pursuing concrete technical implementations such as containerized compound AI services, active-learning pipelines, and orchestration engines that can be directly embedded into defense training environments. In sum, the future of synthetic media detection lies not in building better individual models, but in designing smarter systems that orchestrate those models to support, guide, and adapt to human users. Compound AI represents this evolution—an integrated framework that aligns machine intelligence with human judgment to meet the demands of increasingly complex information environments. By investing in these collaborative capabilities now, we can ensure that future teams (both human and AI) are better equipped to detect, interpret, and respond to synthetic media threats at scale.

DESIGN AND TRAINING IMPLICATIONS

The development of these systems carries direct implications for training and simulation design. Preparing the future workforce to operate alongside intelligent analytic systems requires far more than technical proficiency with tools. It demands cognitive fluency in collaborating with AI under ambiguity, pressure, and incomplete information. Equally, systems themselves must be trained alongside the human, leveraging synthetic scenario generation to expose detectors to new manipulation styles, incorporating online retraining from user annotations, and measuring progress using joint performance metrics such as time-to-insight or error correction rates. Effective human-AI teaming must be learned, practiced, and evaluated, just like any other mission-critical capability. To that end, simulation and training environments should move beyond static tool demonstrations and incorporate the dynamic, interpretive demands of real-world workflows. The focus should be on developing operators' and analysts' ability to reason with and through AI—leveraging system outputs while retaining human judgment, especially when faced with edge cases, adversarial noise, or analytic disagreement. We recommend the following design strategies to support this evolution:

- **Scenario-Based Training with Ambiguity and Conflict.** Training should include high-fidelity synthetic media with realistic manipulation techniques and adversarial content. These scenarios must extend beyond simple detection to include conflicting outputs, incomplete data, and intentional ambiguity. For example, in a USSOCOM mission rehearsal, trainees might receive adversarially manipulated video injects into an ISR feed. Compound AI would triage the input, route it to appropriate detectors, and present disaggregated outputs with visual heatmaps and textual explanations. The trainee's task is not only to spot manipulation but also to calibrate reliance on the AI outputs under pressure, building resilience that can be assessed in after-action reviews. Embedding compound AI modules into environments such as the Information Operations Network (ION) or Joint Synthetic Environment (JSE) would enable adversarial injects, feedback capture, and co-adaptive system retraining in real time. In this way, training becomes a living testbed where humans and AI systems learn side by side, continuously improving both analytic workflows and operator judgment.
- **Performance Metrics for Human-AI Teams.** Assessment frameworks should move past evaluating individual tool accuracy or user compliance. Instead, metrics should reflect *joint performance*, including time to insight, the ability to detect and correct system errors, confidence calibration under time constraints, and the strategic use of triage to prioritize attention. These team-level metrics better capture the collaborative dynamics that define successful human-AI integration. Additionally, evaluations should consider how AI behaviors and human responses align with organizational values over time, recognizing that trust emerges from repeated, value-consistent interactions, not just point-in-time correctness.
- **Modular, Role-Based Learning Tracks.** Training must be tailored to the operational context and user role. Fast-response personnel need efficient triage, confidence interpretation, and rapid decision-making. For analysts and investigators, deeper engagement is required, focusing on layered interpretation, cross-analytic reasoning, and ability to construct a coherent narrative from potentially conflicting evidence. These distinct roles require *modular training paths* that align with cognitive demands and mission objectives.
- **Embedded Explanation Literacy.** Users must be explicitly trained in how to interpret and evaluate system outputs—whether visual (e.g., heatmaps, bounding boxes), statistical (e.g., confidence scores, thresholds), or

descriptive (e.g., analytic metadata or natural language explanations). This includes teaching users when to defer to the system, when to question it, and how to escalate or override outputs appropriately. Simulation environments should include compound AI behaviors that test users' interpretive reasoning, not just their procedural knowledge. Explanation literacy should also include a focus on values-informed judgment that helps users recognize when system outputs align or misalign with core organizational principles. Simulation environments should support the development of interpretive and ethical reasoning through compound AI behaviors and formative feedback on value-based decision consequences.

Ultimately, preparing for the future of synthetic media defense means designing training systems that not only teach humans to interpret AI outputs but also provide the technical infrastructure for AI to learn from humans. This co-adaptive approach ensures that operational simulators become living testbeds, continuously updating analytic workflows, retraining orchestration engines, and reinforcing explanation strategies that align with user expertise and mission tempo. As synthetic media threats grow more sophisticated, the ability to navigate analytic uncertainty, understanding where the system excels, where it fails, and how to operate in that space between, will be essential. As the I/ITSEC community advances the next generation of simulation platforms, we have a critical opportunity to lead in shaping the cognitive and operational infrastructure needed for trustworthy human-AI teaming. By embedding interpretability, role-awareness, and co-adaptive feedback into our training ecosystems, we can prepare the workforce to meet the evolving challenges of synthetic information with agility, insight, and confidence.

CONCLUSION

The rise of synthetic media demands a new generation of detection systems designed not just for accuracy but also for collaboration. Our findings highlight that effective human-AI teaming requires more than high-performing models; it requires systems that are transparent, adaptive, and aligned with the cognitive and operational demands of users. Compound AI offers a path forward: an architecture that integrates triage, orchestration, and layered explanation to support trust, accelerate insight, and enable resilient decision-making. By embedding these principles into both system design and training environments, we can equip the next generation of operators and analysts to navigate complexity with confidence and build truly mission-ready human-AI teams. At the same time, this work has several limitations that suggest directions for future research. While our evaluation framework and pilot studies provide valuable insights, they represent early-stage efforts with limited samples and contexts. Generalization to new adversarial techniques and rapidly evolving generative models remains an open challenge. The orchestration strategies we outline, such as reinforcement learning for tool sequencing, still require empirical validation at scale, and our interpretability recommendations must be adapted to different levels of operator expertise and operational tempo. Future work should pursue the following: (1) large-scale empirical validation of compound AI architectures under red-teaming conditions and adversarial laundering, (2) staged deployment of modular compound AI services in training environments to evaluate training impact, and (3) development of joint human-AI performance metrics that go beyond accuracy to capture trust calibration, resilience under analytic disagreement, and time-to-insight. With these advancements, the community can move from promising concepts toward operationally proven, trustworthy human-AI systems.

ACKNOWLEDGEMENTS

We thank Defense Advanced Research Projects Agency (DARPA) for their invaluable feedback and guidance throughout the program and our many partners and collaborators on this effort without whom this work would not have been possible.

REFERENCES

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Azizpour, A., Nguyen, T. D., & Stamm, M. C. (2025). Autonomous and self-adapting system for synthetic media detection and attribution. *arXiv Preprint*, arXiv:2504.03615. https://arxiv.org/abs/2504.03615

Cooke, D., Edwards, A., Barkoff, S., & Kelly, K. (2024). *As good as a coin toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli* (arXiv:2403.16760). arXiv. https://doi.org/10.48550/arXiv.2403.16760

Diedrich, F. J., Riccio, G. E., Toumbeva, T. H., & Flanagan, S. M. (2023, November). Toward a theory of human-AI co-learning and trustworthiness. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, Orlando, FL.

Feldman, P., Dant, A., & Dreany, H. (2024). *War elephants: Rethinking combat AI and human oversight*. arXiv. https://doi.org/10.48550/arXiv.2404.19573

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2021). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences, 118*(1), e2110013119. https://doi.org/10.1073/pnas.2110013119

Gu, T., Wang, H., & Zhang, X. (2025). A generalizable deepfake detection framework integrating spatial and frequency features. *International Journal of Intelligent Systems*. https://doi.org/10.1155/int/7084582

Hagos, D. H., El Alami, H., & Rawat, D. B. (2024). AI-driven human-autonomy teaming in tactical operations: Proposed framework, challenges, and future directions. *arXiv*. https://doi.org/10.48550/arXiv.2411.09788

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., ... & Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, *16*(1), 45-74.

Kalpokas, I. (2021). Problematising reality: The promises and perils of synthetic media. *SN Social Sciences, 1*(1). https://doi.org/10.1007/s43545-020-00010-8

Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection: Challenges and opportunities. *Artificial Intelligence Review, 57*(6), 159. https://doi.org/10.1007/s10462-024-10810-6

Mirsky, Y., & Lee, W. (2020). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54*(1), 1–41. https://doi.org/10.48550/arXiv.2004.11138

Naber, A. M., Cassani, L., Cook, J., Bautista, P., & Fortier, L. (2024). Comparing Human to Analytic Performance on Detecting, Attributing, and Characterizing Manipulated Media. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *68*(1), 359-362. https://doi.org/10.1177/10711813241262035

Sayler, K., & Harris, L. (2023). *Deepfakes and national security* (CRS Report No. IF11333). Congressional Research Service. https://crsreports.congress.gov/product/pdf/IF/IF11333

Uchendu, A., Lee, J., Shen, H., Le, T., Huang, T.-H. K., & Lee, D. (2023). Does human collaboration enhance the accuracy of identifying LLM-generated deepfake texts? *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 11*(1), 163–174. https://doi.org/10.1609/hcomp.v11i1.27557

Volkova, S., & Jang, J. Y. (2018). Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the Web Conference 2018 (WWW '18)* (pp. 575–583). https://doi.org/10.1145/3184558.3188728

Zaharia, M., Khattab, O., Chen, L., Davis, J. Q., Miller, H., Potts, C., ... & Ghodsi, A. (2024). *The shift from models to compound AI systems.* Berkeley Artificial Intelligence Research Lab. Retrieved on May 28, 2025 from https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/

Zhang, D., Lin, F., Hua, Y., Wang, P., Zeng, D., & Ge, S. (2022). *Deepfake video detection with spatiotemporal dropout transformer*. arXiv. https://arxiv.org/abs/2207.06612

DISCLAIMER

The research described herein is sponsored by the Defense Advanced Research Projects Agency (DARPA, Contract No. 47QFLA22F0137). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the US Government.