A Taxonomy of Bullshit: Classifying AI Truth-Indifference in Policing Contexts

Brandon May[1], Jordan Nunan[2], Marshall Jones[1], Marek Palace[3]

[1]Florida Institute of Technology, United States

[2]University of Northampton, UK

[3]Liverpool John Moores University, UK

*Corresponding author. Email: bmay@fit.edu

Abstract

Since 2022, Generative AI has seen rapid development across sectors, and more recently deployments in policing, including report writing, administrative automation, intelligence analysis, investigations and crime scene investigations. However, there remains a lack of understanding toward (i) how these systems systemically fail, and (ii) the threats this can create within evidentiary workflows, and criminal justice as a whole. Current discourse typically associates these failure modes under the umbrella term hallucination; however, such anthropomorphism is inappropriate and disregards the nuance of how these systems might fail on a human-AI interactionalist, diagnostic level. Using a structured approach to conceptual analyses, this article develops a more precise framing grounded in Frankfurt's concept of bullshit and synthesises existing technical, philosophical, and policing-relevant literatures into three categories of error: factuality errors, faithfulness errors, and task-specific errors. The central contribution was to map each category to the minimum review resource required to detect it: external verification, source-output comparison, and domain-expert evaluation respectively. This approach to detectability carries direct operational consequences for how policing should allocate review effort and decide which use cases are appropriate for current LLM systems. In addition, this article argues that no single approach can mitigate these issues. Specifically, factuality errors can be detectable via verification, whereas faithfulness and task-specific errors require a motivated, domain-specific expertise. In many cases, the latter two categories presuppose that policing has the sufficient resources, time and expertise required, which operational settings rarely afford. This taxonomy provides a working framework for practitioners, but is subject to refinement in future research.

Keywords: artificial intelligence, hallucination, bullshit, policing, large language models, taxonomy, confabulation, truth-indifference

Introduction

The implementation of Generative Artificial Intelligence (GenAI) in policing has accelerated (Coleman, 2025; Halford, 2025; Home Office, 2026) and seen significant investment (Home Office, 2026) and deployment (e.g., Halford, 2026) across jurisdictions (Tao et al., 2026). Police agencies now employ or are evaluating Large Language Models (LLMs) for report writing (Brems, 2026; Ferguson, 2025; Halford, 2026), threat and risk assessment (Halford & Webster, 2024; King & May, 2026), intelligence analysis (May et al., 2026), witness statement processing (Ministry of Justice, 2026), investigative interviews (Järvilehto et al., 2026; May, 2025; Sun, 2025), and administrative functions (Halford, 2026). The UK Home Office has reported that AI-assisted document redaction saved over 600,000 staff hours (6,578 minutes per employee, per year; Home Office, 2025), and commercial products, such as Axon's Draft One, Code Four, and Truleo's systems, are already generating police incident reports from body-worn camera audio in operational settings (e.g., see, Alderden et al., 2026; Watts et al., 2024; White et al., 2025).

These developments are occurring against increasing regulatory frameworks, including the European Union's AI Act (European Commission, 2021), the California AI Transparency Act (Senate Bill 942), and the UK National Police Chiefs' Council's Covenant for Using Artificial Intelligence in Policing (National Police Chiefs' Council, 2023). Under the European Union's AI Act, most law-enforcement uses of AI are expressly classified as high-risk (European Commission, 2021, Annex III). Notably, the UK National Police Chiefs' Council's Covenant (2023) and the California AI Transparency Act (SB 942) do not use the high-risk designation but impose comparable transparency, accountability, and oversight obligations on AI use within their respective jurisdictions (see, for instance, Independent Review of the Criminal Courts, Part 1, Chapter 4, Sir Brian Leveson).

The scale of this raises distinctive concerns for policing as a discipline. Policing has long been characterised by wide officer discretion exercised under conditions of time pressure, incomplete information, and asymmetric power (Bittner, 1970; Lipsky, 2010; Manning, 2010), and by an evidentiary standard in which documentation produced at the point of contact becomes the

record on which subsequent decisions depend (Ericson & Haggerty, 1997; Ferguson, 2025). The introduction of LLM-assisted drafting into this environment changes the conditions in ways that current policing research has anticipated but seldom evidenced. Where officer discretion has historically been constrained by the need to commit a written account to the record (e.g., plea bargains; Ferguson, 2025), AI-assisted drafting reduces the cognitive cost of producing that account. In so doing, this may reduce the deliberative friction that has been an implicit safeguard against careless or motivated reporting (see also Frydl & Skogan, 2004; Walker, 2005, on accountability mechanisms in policing). The accountability literature has consistently identified the audit trail (i.e., the link between observation, account, and decision) as the mechanism through which policing decisions are reviewable after the fact; AI-mediated drafting alters the shape of that audit trail without yet being matched by corresponding revisions to oversight design (e.g., Green, 2022).

Indeed, the conceptual and operational frameworks available to policing for understanding when, where, and how AI, specifically LLM, systems fail remain underdeveloped. Existing technical typologies of LLM error (e.g., Huang et al., 2023; Ji et al., 2023; Rawte et al., 2023) describe the failure modes of these systems in general terms, but they do not address what policing in particular needs to know. For instance, which errors can be caught by which review resource; which categories of error are most likely to pass undetected in operational contexts; and, which use cases place the heaviest detection burden on the practitioners least equipped to discharge it. Nor does the existing policing research on discretion and accountability yet extend to AI-mediated documentation, which leaves a conceptual gap on both sides of the literature.

As such, this article provides four key contributions. First, it evaluates the terminology used to describe AI-generated errors (e.g., Ariso & Bannister, 2025), arguing that the dominant hallucination metaphor is anthropomorphic, clinically inappropriate, and operationally ambiguous (see, Maleki et al., 2024) when applied to policing contexts. More specifically, it makes the case that Frankfurt's (2005) concept of bullshit, understood as output produced with constitutive indifference to truth-value, provides a more precise and practically useful characterisation of LLM behaviour. This is important, as once the terminology of perception and belief is removed, what arguably remains is a system whose outputs can be classified by what is required to evaluate them rather than by what the system was trying to do. Second, building on

this theoretical principle, the article develops a three-category taxonomy (henceforth referred to as the *Taxonomy of Bullshit*) of AI truth-indifference operationalised for policing: *factuality errors, faithfulness errors, and task-specific errors*. Third, the article maps each category to the minimum review required to detect error, including external verification for factuality errors, source-output comparison for faithfulness errors, and domain-expert evaluation for task-specific errors. Detectability of errors has direct operational consequences for how policing organisations should allocate review effort, design oversight, and decide which use cases are appropriate for current LLM systems. Fourth, the article argues that these error categories rarely occur in isolation, and that their joint behaviour under conditions of cognitive bias produces a compounding problem in which the probability of detection decreases as (i) the number of co-occurring error types increases, and (ii) as the errors propagate across the criminal justice system (i.e., cascading and snowballing biases; Dror, 2025).

Toward A Taxonomy of Bullshit: Classifying AI Truth-Indifference for Policing

The Taxonomy of Bullshit posits a theoretical contribution and methodological framework, which can be best characterised as a structured conceptual analysis. This approach is appropriate where the prior question (i.e., what kind of phenomenon is an AI-generated error?) remains underdeveloped, and where premature deployment of LLMs in policing risks measuring the wrong thing. The paper is therefore constructed in four stages, where each is anchored in extant literature and each with its own object of analysis. More specifically, the framework seeks to posit an analytic rather than empirical claim whereby it synthesises and orders existing technical, philosophical, and policing relevant work under conditions of inclusion. This includes the requirement that (i) each error class describe a distinct output level failure, (ii) corresponds to an identifiable detection mechanism, and (iii) be observable in policing processes. Note, these conditions are not protocol driven (e.g., a systematic review) and instead they present the conditions within which a source was treated as relevant evidence for the present article, keeping in mind that (i) there is relatively limited evidence that substantiates the use of LLMs in a criminal justice context explicitly, and (ii) the claims made henceforth are presented on the basis for later revision as the literature grows.

Stage One. The first stage of the framework was concerned with conceptual critique, and in particular with the terminology that has come to dominate discussion of AI-generated error

(Hicks et al., 2024). It is therefore important to highlight that the object of analysis at this stage was not the error itself. The focus remained on the language used to describe AI errors, on the basis that the terminology the field adopts shapes how its phenomena are understood, weighted, and acted upon (Cheng et al., 2024; Ibrahim & Cheng, 2025). The procedure, as noted above, was conceptual in the analytic philosophical tradition (Frankfurt, 2005; Hicks et al., 2024), whereby each term (hallucination, confabulation, bullshit) was examined against the constitutive features attributed to it in its originating literature, and those features are then examined against how LLMs actually generate output (Vaswani et al., 2017; Fredrikzon, 2025). A term is treated as inappropriate where its definition presupposed properties that the referent does not possess, such as percepts (Telles-Correia et al., 2015; American Psychiatric Association, 2022), beliefs (Hicks et al., 2024), or processes of source monitoring (Schnider, 2008; Smith et al., 2023). The anchors at this stage are therefore twofold: (i) the clinical and philosophical sources in which each term is originally defined (Frankfurt, 2005; Hatem et al., 2023; Schnider, 2008; Telles-Correia et al., 2015); and, (ii) the technical literature describing transformer-based next-token prediction (Ji et al., 2023; Vaswani et al., 2017; Wanner et al., 2025). It is worth highlighting at this stage, that the analysis does not report on the prevalence of any given error; however, it does attempt to make a claim about what kind of phenomenon is being named, and whether the name currently in use is the one that should be retained.

Stage Two. The second stage of the framework sought to develop the taxonomy itself, and was concerned with how the various forms of AI-generated error could be ordered into categories that were meaningful for a policing context. The object of analysis at this stage relied on the existing body of work that sought to describe LLM error at the time of analyses (up to March 2026). For transparency, sources were located through iterative keyword searches of Scopus, Web of Science, EBSCO Discovery, and preprint repositories (e.g., ArXIV) conducted between July 2025 and March 2026, using combinations of terms relating to LLM error (e.g., hallucination, confabulation, sycophancy) and to policing and criminal justice applications. In addition, backward and forward citations were identified and included where they were deemed relevant to the study aims. Note, two distinct literatures were relied on: (i) technical typologies of LLM failure (e.g., Alhejaili et al., 2025; Huang et al., 2023; Ji et al., 2023; Rawte et al., 2023), and policing relevant or adjacent (e.g., criminal justice) studies that documented how those failures manifested in operational settings (e.g., Halford & Webster, 2024; Magesh et al., 2025; Varotsis,

2025; Watts et al., 2024; White et al., 2025). The synthesis also relied on three conditions of inclusion. First, each category had to describe a distinct output level failure, such that a given error instance could be classified into one category, or, in the compounding case, into more than one, without category overlap collapsing the distinction between them (e.g., Huang et al., 2023; Ji et al., 2023). Second, each category had to correspond to an identifiable detection mechanism and had to differ qualitatively across categories (Farquhar et al., 2024; Magesh et al., 2025). Third, each category had to be observable in policing processes (e.g., Halford & Webster, 2024; Watts et al., 2024; White et al., 2025). Categories from the technical literature that did not meet the third condition (e.g., purely linguistic measures such as token-level factors) were considered as part of the broader factuality and faithfulness classes rather than retained as separate categories, on the basis that they did not map cleanly onto a review step that a police officer or supervisor would perform in practice.

Stage Three. The third stage of the framework mapped each category in the taxonomy to the minimum review resource required to detect it and was concerned with translating the conceptual distinctions established in stage two into an ordering with practical consequences for how police agencies, supervisors, and oversight bodies might allocate review effort. The object of analysis at this stage was the relationship between errors and detection. What was later established was three interconnected families of errors: (i) Factuality errors were paired with external verification, on the basis of benchmarking evidence that such errors were surfaced, where they were surfaced at all, by comparison against ground-truth sources (Farquhar et al., 2024; Magesh et al., 2025); (ii) Faithfulness errors were paired with source-output comparison, on the basis that these errors were invisible to external verification, in that the fabricated content might be plausible, and were only exposed when the original input was held alongside the generated output (Ji et al., 2023; Watts et al., 2024); and (iii) Task-specific errors were paired with domain-expert evaluation, on the basis that the empirical work documenting them had required practitioner judgement to identify the failure in the first place (Halford & Webster, 2024; Varotsis, 2025).

Stage Four. The fourth stage of the framework was not part of the original conceptual plan but emerged as a necessary extension once the taxonomy itself had been developed. Specifically, during stages two and three it became clear that the three error categories did not behave

independently in practice, and that the conditions under which they were most likely to occur in policing settings were also the conditions under which they were most likely to co-occur (Dror, 2025). This stage was therefore the joint behaviour of the taxonomy under conditions of cognitive bias, and the procedure was one of conceptual integration between the taxonomy developed in the preceding stages and the established literature on cascading error in forensic and criminal justice decision-making (Dror, 2018, 2020, 2025; Kassin et al., 2013; Kukucka & Dror, 2023). The primary argument posited was that the cognitive mechanisms documented in the literature, including anchoring (Tversky & Kahneman, 1974), confirmation bias (Ask & Granhag, 2005), and automation bias (Parasuraman & Manzey, 2010; Strauß, 2021; Vered et al., 2023), operated on AI-generated outputs in much the same way they had been shown to operate on other forms of input, with the additional and consequential feature that AI-generated outputs were produced under conditions of truth-indifference rather than truth-orientation (Frankfurt, 2005; Hicks et al., 2024). The compounding claim that followed was therefore a derived prediction. In other words, if the taxonomy was correct that the three categories required qualitatively different detection mechanisms, and if the cognitive bias literature was correct that practitioners operating under time pressure rarely applied all three review conditions simultaneously, then the joint probability of detection across all three classes is predicted to be lower than the probability of detection for any single class taken alone; this prediction, and the assumptions on which it rests, are stated formally in Bayesian terms in the section on the compounding problem below. To evidence this, extant materials for forensic decision-making on cascade, snowball, and bias-on-bias effects were used (Dror, 2018, 2020, 2025; Kassin et al., 2013; Kukucka & Dror, 2023), and the cognitive bias studies in which the relevant mechanisms had been independently established (Ask & Granhag, 2005; Parasuraman & Manzey, 2010; Tversky & Kahneman, 1974; Vered et al., 2023).

Why Hallucination Is The Wrong Metaphor

The term hallucination is commonly used to describe AI-generated errors across technical, policy, and media (Ariso & Bannister, 2025; Liu & Lin, 2026; Maleki et al., 2024; Smith et al., 2023), yet this framing introduces a lack of conceptual clarity by implying human-like perception or cognition where none exists (May, 2025). Whilst it might be an argument of semantics (e.g., generation of ungrounded content; Berk, 2024), language used to describe AI

errors in policing shapes how those outputs are understood, weighted, and operationalised (e.g., by officers, supervisors, courts, and oversight bodies). Referring to AI errors as hallucinations risks normalising them as rare or unavoidable anomalies, rather than as predictable by-products of probabilistic text generation, thereby obscuring their implications for evidentiary reliability, auditability, disclosure obligations, and procedural fairness.

In clinical psychiatry and psychology, hallucination refers to a perceptual experience that occurs in the absence of corresponding external sensory stimulation (Telles-Correia et al., 2015). In other words, it is a phenomenon of conscious awareness wherein the individual perceives stimuli (e.g., visual, auditory, tactile) that are not present in their environment but are experienced as genuinely real (American Psychiatric Association, 2022). Hallucinations are clinically associated with conditions, such as schizophrenia, bipolar disorder, Parkinson's disease, and substance-induced psychosis (Hatem et al., 2023). As Brender (2023) argues, employing the clinical term hallucination for this process implies acceptance that LLMs are engaged in perceiving (in becoming consciously aware of sensory input) for which there is no evidence whatsoever.

The broader concern with hallucination is its contribution to the anthropomorphisation of AI systems, commonly discussed as the attribution of human cognitive properties to computational processes (see, Forster & Skop, 2025; Ibrahim & Cheng, 2025). For instance, Cheng et al. (2024) developed AnthroScore, a computational measure of anthropomorphism in scientific writing about LLMs, finding that anthropomorphic language had steadily increased in recent years, and more specifically Ibrahim & Cheng (2025) later found that over 40 per cent of LLM research paper abstracts contained anthropomorphic language as of January 2023, rising to 48 per cent by December 2024. The tendency to describe AI systems using terms drawn from human cognition (e.g., hallucinating, understanding, reasoning, knowing) arguably creates implicit mental models that shape how practitioners interact with these tools.

For policing, the anthropomorphism rooted in hallucination is particularly problematic because it may lead officers and analysts to conceptualise AI systems as cognitive agents [or co-pilots] that are trying to get it right but occasionally fail, rather than as statistical text generators that have no concept of right or wrong in any meaningful sense (Harnad et al., 2024; Yuan et al., 2024). Research on automation bias (Strauß, 2021) demonstrates that operators tend to over-rely on

automated systems and under-scrutinise their outputs (Parasuraman & Manzey, 2010), and anthropomorphic framing of these systems is likely to exacerbate this tendency.

Confabulation versus bullshit

Two principal alternatives to hallucination have emerged in the literature: confabulation and bullshit. Confabulation originates in neuropsychology, where it refers to the generation of narrative accounts that, while incorrect, are produced without conscious intent to deceive, typically to fill gaps in memory (Schnider, 2008). Unlike hallucinations, confabulations involve accounts of past events or factual claims that the individual genuinely believes to be accurate. Here, what is presupposed, from a biological context, is that the neural substrates differ (e.g., Corota & Bogousslavsky, 2019). Specifically, hallucinations are associated with aberrant perceptual processing, while confabulations involve failures of memory monitoring and source attribution (Smith et al., 2023). Sui et al. (2024) identify several reasons confabulation has emerged as a preferred alternative for AI contexts, in that it does not pertain to percepts or stimuli, thereby avoiding the implication that LLMs have sensory experiences or consciousness. In other words, it is a more neutral term sensitive to mental health communities, and it is applicable to non-pathological scenarios, as everyday memory reconstruction in healthy individuals routinely involves confabulatory processes.

However, it is worth noting that confabulation retains anthropomorphic limitations. Whilst it is considered preferable over hallucination (Shumaker, 2024), it still implies a cognitive agent attempting to recall information and inadvertently producing errors. This is a characterisation that, while closer to the phenomenology of LLM output than hallucination, still attributes cognitive states (belief, memory, attempted recall) that LLMs do not possess. As Hicks et al. (2024) observed, even the preferred alternative of confabulation risks anthropomorphising the technology. Furthermore, confabulation implies good faith; the confabulator believes their account is accurate. Again, LLMs have no beliefs about the accuracy of their outputs; therefore, the concept of belief does not apply.

Bullshit, in Frankfurt's (2005) philosophical sense, avoids both of these problems. It describes output produced with constitutive indifference to truth-value, without requiring the attribution of either perceptual experience (as hallucination does) or memory and belief states (as

confabulation does). This makes it the most technically precise of the available framings for characterising LLM outputs, particularly in contexts where the consequences of those outputs are severe.

Frankfurt's analysis of bullshit

Harry Frankfurt's (2005) essay On Bullshit provides what has become the philosophical analysis of a category of discourse that is distinct from both truth-telling and lying. Frankfurt's central argument is that the bullshitter differs from the liar in that the liar knows the truth and deliberately subverts it, engaging with the truth-value of their statements even as they seek to invert it. The bullshitter, by contrast, is indifferent to truth altogether. The bullshitter's statements are *'grounded neither in a belief that [they are] true nor, as a lie must be, in a belief that [they are] not true'* (Frankfurt, 2005, p. 4). The bullshitter operates in a mode where truth-value is simply not a governing constraint on output.

Frankfurt (2005, p. 7) famously argued that bullshit is a *'greater enemy of the truth than lies are'*, because while the liar at least acknowledges the authority of truth (by taking care to subvert it), the bullshitter undermines the very concept of truth as something worth caring about. The liar respects truth enough to work against it; yet the bullshitter is indifferent to whether truth is even at stake. This distinction is important, as LLMs that generate fabricated legal citations are not lying (i.e., it has no knowledge of the truth to subvert) nor confabulating (i.e., it has no beliefs to inadvertently distort). Therefore, it is reasonable to conclude that it is producing output in a mode where truth-value has never been a governing constraint (i.e., the paradigmatic case of bullshit).

Frankfurt (2006) refined his definition to identify three core characteristics of bullshit, (i) indifference to the truth-value of one's assertions; (ii) the concealment of that indifference; and (iii) the strategic deployment of truth-indifferent output to achieve some communicative or social effect. Harris (2023) further distinguished between strategic and unintentional indifference to truth, arguing that bullshit is morally culpable only when it constitutes negligence, which represents an important distinction when evaluating the responsibilities of those who deploy LLMs in high-stakes contexts, as opposed to the systems themselves.

Hicks et al. (2024) in their now seminal paper 'ChatGPT is bullshit', introduces a useful distinction between hard and soft bullshit. Hard bullshit requires an active attempt to deceive the audience about the nature of the enterprise. When we consider this, the bullshitter not only produces truth-indifferent output but actively conceals this indifference. Soft bullshit, on the other hand, requires only a lack of concern for truth, without the additional layer of concealment. Hicks et al. (2024, p. 2) argue that, at minimum, LLM outputs constitute soft bullshit (i.e., 'speech or text produced without concern for its truth… produced without any intent to mislead the audience about the utterer's attitude towards truth').

More provocatively, they suggest that LLMs may also produce hard bullshit, if the system is viewed as having design-level intentions. The observation is that LLMs are designed to produce text that is indistinguishable from human-generated text, that is, to be convincing rather than accurate. Indeed, Colbert's (2005) concept of truthiness captures this exact perspective, where ChatGPT functions not to convey truth or falsehood but rather to convince the reader of the truthiness of its statement (Hicks et al., 2024). The system is designed in such a way as to make bullshit efficacious. It produces text that looks truth-apt without any mechanism for ensuring that it is.

Truth-indifference: why LLMs bullshit

Understanding why the bullshit framing is not only metaphorical but architecturally descriptive requires an examination of how LLMs function. These systems are trained on a massive corpora of text data to predict the statistically most probable next token (word or sub-word unit) in a sequence, given the preceding context. The training objective is to minimise prediction error across the training corpus; that is, to produce text that is statistically consistent with the patterns observed in human-generated language (Vaswani et al., 2017). At no point in this process does the model learn to verify claims against external reality, check factual consistency, or distinguish true statements from false ones. What it does is learn patterns of language, not patterns of truthfulness, because it has no concept of truthfulness, nor accuracy (Hicks et al., 2024), treating all claims equally without ranking them (Wanner et al., 2025).

As Fredrikzon (2025) argued, hallucinations are not temporary flaws in an otherwise sound system but the predictable outcome of the generative pre-trained transformer architecture itself.

The system's function of predicting statistically likely continuations means that outputs are unremittingly indifferent to knowledge and truth claims (Fredrikzon, 2025). When the model generates a plausible-sounding legal citation, it does so because citations of that form are statistically common in its training data, not because it has verified that the cited case exists. The generation of fabricated but plausible content is not a malfunction in this instance, and therefore what is often concluded is that the system has operated exactly as designed (i.e., to bullshit).

Techniques such as retrieval-augmented generation (RAG), grounding, fine-tuning, and reinforcement learning from human feedback (RLHF) can reduce the rate of factually incorrect outputs, but they cannot eliminate the underlying truth-indifference (e.g., Huang et al., 2025). RAG provides the model with relevant source documents, but the model's use of those documents remains governed by statistical prediction rather than logical inference. The model may ignore, misinterpret, or selectively attend to the provided information in ways that reflect linguistic patterns rather than factual reasoning (Ji et al., 2023). These mitigations reduce the symptoms of truth-indifference without addressing the underlying architecture that produces it.

A Taxonomy of Bullshit: Classifying AI Truth-Indifference for Policing

The categories developed below are the output of the structured conceptual analysis discussed in the above, and do not represent a systematic or scoping review of the LLM error literature. The sources cited within each category were selected against the three conditions of inclusion stated earlier (distinct output-level failure, identifiable detection, and observability in policing processes), rather than through the protocol-driven approach that would typically govern a systematic synthesis. The taxonomy should therefore be understood as a conceptual framework grounded in extant literature at the time of writing, not as an exhaustive list. As noted earlier, the categories that follow are working categories, intended to be tested, contested, and revised as the field develops.

The Taxonomy of Bullshit synthesises and extends existing technical classifications of LLM hallucination (Ji et al., 2023; Huang et al., 2023; Rawte et al., 2023) by grounding each category in the specific operational, evidential, and procedural requirements of policing. The three categories: factuality errors, faithfulness errors, and task-specific errors, represent increasingly context-dependent failure modes, from errors that can be detected by comparison with external

reality, through errors that require comparison with task instructions, to errors that can only be identified by professionals with domain expertise.

Category one: factuality errors

Factuality errors occur when AI-generated output contradicts verifiable real-world information (see Table 1, for a full breakdown). These represent the most straightforward category of truth-indifference, wherein the system produces claims that are demonstrably false against external sources of ground truth (e.g., verifiable detail; Vernham et al., 2020), including dates, names, relationships, events, geographical locations, legal statutes, and established procedures. More simply put, factuality errors are, in principle, detectable through external verification (e.g., a fabricated citation can be checked against a database). However, detection requires that practitioners know to check, have the resources to do so, and are not subject to the cognitive bias that discourages scrutiny of machine-generated content (e.g., Kücking et al., 2024; Lindo et al., 2024).

This challenge is compounded by evidence that automation bias is not limited to inexperienced users, although they are more commonly affected by it (Goddard et al., 2012). Parasuraman and Manzey (2010) demonstrated that automation bias occurs in both naïve and expert participants, which cannot be prevented by training or instructions, and affects decision-making in individuals as well as teams. Their integrated attentional model shows that automation bias produces both omission errors (failing to notice problems because the automated system did not flag them) and commission errors (following incorrect automated recommendations), with commission errors being especially dangerous in policing contexts where an AI-generated factual claim may go unchallenged. More recently, a systematic review found that automation bias manifests across high-stakes domains including healthcare, law, and public administration, and that even explainable AI (XAI) interventions intended to mitigate over-reliance may paradoxically increase automation bias by fostering misplaced trust (Romeo & Conti, 2025).

Vered et al. (2023) similarly provided experimental evidence that having an intelligent agent explain its recommendations did not reduce automation bias and, in some instances, increased it. What this suggests is that the very conditions under which police officers would interact with AI-generated content (i.e., high workload, time pressure, trust in institutional technology, and

multitasking demands) are precisely those most likely to produce uncritical acceptance of factual errors, and thus increases the potential likelihood for automation bias.

Surveys of LLM hallucination (e.g., Ji et al., 2023; Huang et al., 2023; Rawte et al., 2023; Alhejaili, Jiang, Rajan, and Karami, 2025) have found that factuality errors are an inherent and persistent feature of large language model output, arising from the probabilistic nature of next-token prediction, which prioritises linguistic plausibility over factual accuracy (Shuster et al., 2022; Kadavath et al., 2022). Recent empirical benchmarking (e.g., SimpleQA factuality benchmark), demonstrate error rates in the tens of per cent on straightforward knowledge retrieval tasks, with the measured rate depending heavily on whether the model is permitted to abstain and whether evaluation penalises confident errors (Brinsa, 2026). For instance, Farquhar et al. (2024), developed entropy-based methods for detecting what they termed 'confabulations', showing that LLMs frequently produce confident but false outputs that are internally consistent yet factually baseless. What is important to recognize in this work is that these systems cannot reliably predict when they are producing such errors, undermining any assumption that confidence calibration can serve as an internal safeguard (Kumar et al., 2024).

[Insert Table 1 here]

Category two: faithfulness errors

Faithfulness errors occur when AI output diverges from the instructions, prompts, context, or source materials provided to the system (see Table 2, for a full breakdown). Unlike factuality errors, which can be detected by comparison with external reality, faithfulness errors require comparison with the specific input that the system was given (Huang et al., 2025). The output may be factually accurate in isolation but unfaithful to the task as defined by the user or the system's operating parameters. For instance, a factuality error can, in principle, be identified by anyone with access to the relevant ground truth; however, a faithfulness error can only be identified by someone who has access to both the output and the original input, and who is motivated to compare the two. In policing contexts, this comparison rarely occurs systematically (e.g., transactional myopia and police auditing; see, Fan et al., 2018; Jiao, 2024).

This category is particularly relevant to policing because many policing applications of GenAI involve processing specific documents (e.g., witness statements, interview transcripts, crime reports) where fidelity to the source material is essential (Bull et al., 2025; May et al., 2025). An LLM tasked with summarising a witness interview transcript must reproduce the substantive content of that transcript (Milne et al., 2022); if it introduces details not present in the original, omits material information, or restructures the account in ways that alter its meaning, the resulting summary is unfaithful regardless of whether the introduced content happens to be factually accurate (May & Jones, 2026). The Huang et al. (2023) taxonomy of LLM hallucination distinguishes faithfulness hallucination from factuality hallucination precisely on this basis. For example, faithfulness hallucination captures the divergence of generated content from user instructions or provided context, including instruction inconsistency, context inconsistency, and logical inconsistency within the generated output itself (Malin et al., 2024).

Further, Peters and Chin-Yee (2025) tested ten prominent LLMs across 4,900 summaries of scientific texts published in journals such as Nature, Science, and The Lancet, finding that LLM-generated summaries were nearly five times more likely to contain broad generalisations than human-authored summaries of the same articles. Models often overgeneralized, omitting qualifiers, strengthening hedged conclusions, broadening scope beyond what the source warranted, in 26 to 73 per cent of cases, even when explicitly prompted for accuracy. In addition, accuracy prompts nearly doubled the overgeneralisation rate compared to simple summary requests, suggesting that instructions to be accurate may paradoxically trigger more assertive and therefore less faithful output. If this pattern holds in policing applications, officers requesting that an LLM accurately summarise a witness statement may receive output that is more confidently unfaithful to the source than a simple unprompted summary would have been (e.g., the difference between “the suspect may have been present” and “the suspect was present,” or between “the witness thought it looked like a knife” and “the witness saw a knife”).

Context limitations represent further structural vulnerability. Liu et al. (2024) demonstrated the lost in the middle phenomenon, wherein LLM performance was highest when relevant information occurs at the beginning or end of the input context (i.e., anchoring effects; see, Hämäläinen, 2025), and significantly degraded when models accessed information positioned in the middle of long documents, even for models explicitly designed for extended contexts. Du et

al. (2025) later showed that even when models could perfectly retrieve all relevant evidence (e.g., verified by 100 per cent exact-match token recitation) performance still degraded substantially as input length increases. What this suggests is that an LLM processing a lengthy case file or multi-page witness statement may systematically neglect information from the middle portions of the document, producing summaries or analyses that are structurally biased toward recency and primacy biases (e.g., regardless of where the most important information actually falls).

One of the primary problems of faithfulness error is sycophantic validation (Cheng et al., 2025), whereby models privilege alignment with the user's framing or implied expectations rather than adhering to accurate, source-faithful processing of the underlying material. Sharma et al. (2023) demonstrated across five AI assistants that sycophancy is a consistent and general behaviour. Interestingly, Sharma et al. (2023) found models wrongly admitted mistakes when questioned, gave predictably biased feedback aligned with perceived user preferences, and mimicked errors introduced by users. Interestingly, their analysis of human preference data revealed that sycophantic responses were systematically preferred by human evaluators (e.g., within Reinforcement Learning from Human Feedback; RLHF). Chen et al. (2025) further demonstrated that frontier LLMs complied with illogical requests at rates of up to 100 per cent, prioritising helpfulness over logical consistency even when the model possessed the knowledge to identify the request as illogical. The implication is that an LLM processing, for instance, a witness statement, or police report in the context of an officer's existing case theory may selectively emphasise details consistent with that theory (i.e., confirmation bias; Ha & Kim, 2024) while downplaying contradictory evidence (i.e., through the same sycophantic tendency that makes these systems appear cooperative and useful).

The Axon Draft One frog case (Police1, 2026) is a particularly helpful example of this type of faithfulness error; specifically between source processing (e.g., the system processed pre-recorded movie dialogue as if it were witness or officer speech relevant to the incident) and input discrimination (e.g., the system could not distinguish between officer speech pertaining to the incident and a Disney film playing on a television in the background). What distinguishes this incident as a faithfulness error is that the transcribing audio input failed to distinguish between relevant incident audio and irrelevant background noise, integrating both into a coherent

narrative. Interestingly, research has shown that this is to be expected in forensic-audio contexts, as transcription performance often degrades in conditions characteristic of the above case (Loakes, 2024; also see, Srivastava et al., 2025).

Ferguson (2025) has also identified the broader risk as generative suspicion wherein AI-generated narrative attributions reshape the evidential record with no audit trail separating algorithmic contributions from human observations. While the frog case was caught before submission, the structural conditions that produced it (i.e., automated processing of ambiguous audio input, absence of source discrimination, lack of audit mechanisms, and reliance on officer review as the sole safeguard) remain present in every deployment of this technology. In other words, when an AI mediates the translation of experience to text, this can become a fraudulent account, not in the sense that the officer intended to deceive, but because the document claims a relationship to its source that cannot be guaranteed (Hicks et al., 2024).

[Insert Table 2 here]

Category three: task-specific errors

Task-specific errors represent failure modes that are unique to particular operational applications and that require domain expertise to identify (Dressel & Farid, 2018; Gaube et al., 2021; see Table 3, for a full breakdown). Unlike factuality errors, which can be detected by comparison with external ground truth, or faithfulness errors, which can be detected by comparison with input instructions and source material, task-specific errors can only be identified by someone who possesses the professional knowledge required to evaluate whether the AI has applied the relevant framework, standard, or analytical method correctly. The output may be factually accurate, faithful to its input, and formatted appropriately, and yet fundamentally wrong in ways that only a domain expert would recognise (Duenas & Ruiz, 2024; Ji et al., 2023).

The literature on domain-specific LLM failures is convergent, most commonly in regard to performance (e.g., Song et al., 2026). Here, performance deteriorates where task complexity demands the kind of inferential, contextual, and procedural reasoning that distinguishes expert from novice judgment. For instance, Dahl et al. (2024) found that LLMs hallucinated between 58 percent (GPT-4) and 88 percent (Llama 2) of the time when asked verifiable questions about federal court cases. Further, hallucination rates increased with task complexity; when asked to

assess the precedential relationship between two cases (e.g., a core legal reasoning task) most LLMs performed no better than chance. When asked about a court's core holding, models hallucinated at least 75 percent of the time. What is unique here, is that those errors can be coded not as factual retrieval failures, but instead failures of domain-specific reasoning. The models could locate cases but could not reason about what those cases meant (i.e., it failed to interpret its operational significance). Magesh et al. (2025) subsequently showed that even RAG systems marketed specifically to lawyers (e.g., Westlaw AI and Lexis+ AI) still produced incorrect information in 17 to 33 per cent of cases, which would suggest, in part, that domain-specific errors persist even in purpose-built tools.

This can be illustrated by Kim et al. (2025) who conducted a review of medical hallucinations in models, finding that even LLMs developed explicitly for medical purposes remain vulnerable to domain-specific errors, with these failures arising from reasoning failures rather than knowledge gaps. It was noted that medical hallucinations frequently used domain-specific terminology and presented coherent logic, making them difficult to recognise without expert scrutiny. For instance, an AI system that produces a superficially competent risk assessment using the correct terminology and format, but that applies the underlying framework incorrectly, can exhibit this kind of expert failure.

When we consider this problem in a criminal justice context, Halford and Webster (2024) tested ChatGPT (versions 3.5 and 4) on thirty life-like police scenarios using the THRIVE (Threat, Harm, Risk, Investigation, Vulnerability, Engagement) risk assessment framework, finding only 62.5 per cent accuracy as judged by expert practitioners, with failures attributable to the system's departure from the specified assessment methodology. The system could produce THRIVE-formatted outputs, using the correct categories and terminology, while also applying the framework incorrectly. Similarly, King et al. (2026) found that LLMs could produce optimal convergence to heuristic decision-making tasks; however, verbosity and coherence maximisation often limited the models' applicability to real-world context. Varotsis (2025) extended this line of inquiry using a purpose-built platform (CopBot) to compare LLM risk assessments of missing person scenarios against over 600 human assessments from police practitioners. While high-performing models such as GPT-4o showed mean absolute errors below 0.6 on a four-point scale, they exhibited concerning biases. Interestingly, unlike human assessors who showed no

significant racial bias, all models demonstrated some form of racial bias in their risk assessments (also see Pulvino et al., 2026, for their overview on default bias).

Comparably, there is a similar pattern in predictive machine learning (ML) applications within domestic abuse risk assessment. Structured tools such as the Domestic Abuse, Stalking, and Honour Based Violence (DASH) checklist have long been criticized for their limited capacity to identify cases that subsequently escalate to serious harm. Turner et al. (2022) found that 96.3% of domestic abuse cases that later resulted in serious harm had initially been classified as standard or medium risk, reflecting an exceptionally high false-negative rate. In other words, both structured actuarial tools and the practitioners applying them frequently fail to detect escalation patterns at the point of assessment.

What is important to recognise is that this is not primarily a problem of statistical accuracy or any specific frontier model calibration. As Richards (2025) argues, risk in coercive control contexts is not additive; instead, it is inferential. A better way to understand this, is that serious harm does not emerge from the accumulation of checklist indicators, but rather it can emerge from patterns (e.g., stalking behaviour, coercion, escalation) that unfold relationally over time. This distinction has direct implications for task-specific LLM deployment, given that even domain-specific predictive models and trained practitioners struggle to identify high-risk escalation within structured assessment frameworks. If those systems already exhibit substantial false-negative rates (Turner et al., 2022), introducing statistically distributed language models that treat the same material as a classification, summarisation, or scoring task is unlikely to correct the problem (e.g., Busch et al., 2025; Xu et al., 2024). Instead, such systems risk reproducing the same blind spots while presenting their outputs as consistent, systematic, and analytically robust (Busch et al., 2025).

Beyond inconsistencies, research has also shown that LLMs operate over statistical regularities in linguistic form rather than grounded representations of meaning or world models (Bender et al., 2021), raising concerns about their ability to capture causal or inferential structure. Indeed, Bender et al. (2021) argues that frontier LLM models are often trained exclusively on strong prediction talk, learning probabilistic associations between linguistic forms without access to intent, shared knowledge of models of causality (i.e., it functions as a stochastic parrot). However, even in instances where a benchmark using fresh corpora not present in training data

was provided, research has found that all studied LLMs (including GPT-4) have shown significant performance drops compared to earlier benchmarks where corpora overlap with training data was likely (i.e., LLMs can only provide shallow reasoning; Chi et al., 2024). This suggests that apparent causal reasoning performance on established benchmarks may largely reflect memorisation of common causal patterns rather than genuine causal inference. Interestingly, Lithgow-Serrano et al. (2025) demonstrated that exposure to causal statements during pretraining did not improve causal classification performance, nor did it reduce uncertainty, suggesting that LLM failures in causal tasks stem from representational limitations rather than memorisation of causal patterns. For task-specific errors, this limitation is particularly dangerous as an LLM may identify a statistical pattern on, for example, offending data and present it as a causal relationship with appropriate analytical vocabulary, while the relationship is in fact spurious (e.g., concluding that individuals who miss a probation appointment are more likely to escalate into serious offending behaviours, when the observed association is likely driven by exogenous factors; see Harcourt, 2006; Lum et al., 2016).

What unites these domain-specific failure modes is a structural asymmetry. In other words, the same features that make AI output appear competent (e.g., correct formatting, appropriate terminology, internal coherence) are precisely the features that make task-specific errors hardest to detect (e.g., LLM provides surface structure, absent of meaning; Bender & Koller, 2020; Brown et al., 2020; Marcus & Davies, 2020). A factuality error produces a false claim that can be checked (Ji et al., 2023); a faithfulness error produces output that departs from its source in ways that can be compared (Maynez et al., 2020). However, a task-specific error produces output that looks right to anyone who is not an expert in the specific domain, and that may even look right to experts under conditions of time pressure, high workload, and automation bias (Skitka et al., 1999). Note, task-specific errors, as a consequence, are most likely to pass undetected through human review.

[Insert Table 3 here]

Applying the Taxonomy in Practice

The taxonomy's three categories are distinct in that they answer three independent questions about an output, but overlapping in that the categories are not mutually exclusive. In other words,

a single output rarely fails in one way only, and an error in one category often produces errors in the others. Classification therefore proceeds by assessing every output against all three, in fixed order, assigning each classification that applies; two or more constitutes a compound error. To achieve this, one must evaluate whether any claim in the output conflicts with verifiable external information, such as a case, statute, date, or statistic. Where it does, the output contains a factuality error, and only a check against an authoritative source can establish this, because the error lies between the output and what is verifiably true, rather than anywhere inside the document.

The second is whether the output diverges from its task-defined input, whether by adding what was never there, omitting what was, altering meaning, or importing extraneous material. Where it does, the output contains a faithfulness error, and only a side-by-side comparison of output against input can establish this, because the output may be entirely plausible, and often true, while still distorting or misrepresenting the material it was meant to represent. The third is whether the output misapplies a professional framework governing the task, irrespective of whether it is accurate and faithful. Where it does, the output contains a task-specific error, and only domain-expert evaluation can establish this, because the error perhaps lies the reasoning rather than in the facts, and is invisible to anyone who cannot assess how the framework should have been applied. See Table 4 for details.

[Insert Table 4 here]

With this in mind, it is important to recognise that the three *questions* are logically independent, and overlap between categories does not collapse the distinctions between them. In other words, a compound error is not a fourth category. The Heber City report (Police1, 2025), for instance, represents (i) a factuality error because it asserts something that did not happen, which only the body-worn video established, and (ii) a faithfulness error because it treated dialogue from a film that formed no part of the incident as though it were incident speech, which only a comparison of the narrative against the recording identified.

The Compounding Bullshit Problem

In applied contexts, LLM error categories rarely occur in isolation. Instead, they tend to co-occur, interact, and compound, producing effects that exceed the risks posed by any single error

type (e.g., Dror, 2025). Indeed, Hicks et al. (2024) argue, ontologically, that compounding bullshit is a reflection of repeating bullshit. In other words, if a non-bullshitter repeats something that a bullshitter has said, that itself becomes bullshit, regardless of whether the individual repeating the content believes it to be genuinely true. Beyond the ontological claim, operationally what this means for policing is that the repetition of truth-indifference becomes the compounding problem. This can be operationalized through the sequential decision-making processes in policing, including amplifying errors (e.g., runaway feedback loops; Ensign et al., 2018), inflation confidence, and reducing the likelihood of detection through cognitive bias (e.g., automation bias).

This also reflects findings from the deception and forensic decision-making literature, which demonstrate that cognitive biases seldom operate independently. Confirmation bias, for example, increases anchoring effects (Hämäläinen, 2025), which in turn reinforce belief perseverance, generating cascading distortions in judgment (Chen et al., 2024; Kassin et al., 2013; Kukucka et al., 2017; Lou & Sun, 2026). AI-generated errors exhibit similar cascading effects (e.g., 'when AI systems hallucinate or produce other errors, these mistakes can cascade through human review processes and create systematic biases'; Beck et al., 2025, p. 2), such that an initial failure can trigger a chain of reinforcing inaccuracies that magnify downstream harm (see Figure 1; Hicks et al., 2024; Xu et al., 2025). Once these compounded errors enter the justice system, their joint influence is not contained within the task or stage that produced them, nor are the cognitive biases that influence them contained; rather, they spread across actors, decisions, and institutional processes, snowballing (see, Dror, 2025) through the legal system in ways that research on cognitive bias has only recently begun to formally articulate (Dror, 2018, 2020, 2025; Kukucka & Dror, 2023).

Frankfurt's (2005) core insight was that bullshit is more dangerous than lying because the bullshitter operates without reference to truth as a governing constraint. In other words, the liar must at least know the truth in order to subvert it, while the bullshitter is indifferent to whether claims are true or false (see also Fredrikzon, 2025). The compounding problem extends this, as compound AI bullshit is simultaneously indifferent to factual truth, source fidelity, and domain-specific validity, presenting a sort of triple indifference within a surface of professional competence that actively delays (or completely subverts) detection (Bender et al., 2021; Hicks et

al., 2024). While a factuality error alone can be caught by fact-checking; a faithfulness error by source comparison; a task-specific error by expert review; however, compound errors require that all three detection methods be applied simultaneously. The relationships among the three categories can be represented diagrammatically (see Figure 1), with the caveat that the diagram is offered as a heuristic illustration of the conceptual structure of the taxonomy rather than as evidence for the prevalence, weighting, or empirical separability of the three categories.

[Insert Figure 1 here]

The compounding prediction as sequential Bayesian model

The taxonomy does not only predict that compound errors are harder to detect, it also predicts that the sequential structure of criminal justice system makes their detection progressively less likely at each successive stage. For example, consider an error moving through officer review, supervisory review, prosecutorial review, and trial. The Swiss cheese model (Reason, 1990) treats these stages as independent layers, such that survival requires passing each in turn. This is the baseline against which the compounding prediction is set, namely how well a system would detect errors if each reviewer examined the document with fresh eyes, without knowing that anyone had approved it before them. In that regard, the probability of surviving all stages is the product of the probabilities of surviving each, which is to say the model is multiplicative across review stages only. However, it carries no claim that co-occurring error classes multiply one another's effects. Operationally this is where the assumption of independence fails. Case in point, a prosecutor does not receive a raw document; what they receive is one that an officer has signed and a supervisor has approved, and those approvals are part of what they see. When one considers this, each carries forward the judgement of the one before it.

Indeed, review processes rarely generate false alarms, in that accurate reports are almost never rejected as erroneous; this reflects the response bias that signal detection accounts predict where the prior probability of error is perceived to be low and the cost of rejection is high (e.g., Green & Swets, 1966; Hautus et al., 2021). Consequently, survival at one stage is far more consistent with there being nothing to find than with something having been missed. In other words, every prior approval lowers the perceived posterior probability of error, such that each reviewer inherits a lower expectation than the reviewer before. Where scrutiny tracks expected error, as

the anchoring and automation bias literatures suggest (Parasuraman & Manzey, 2010; Tversky & Kahneman, 1974; Vered et al., 2023), scrutiny itself declines as the document advances.

In Dror (2025), a forensic examiner who learns that a suspect has confessed no longer examines the evidence independently; indeed, that examiner's conclusion then travels to the next actor, who receives it as corroboration rather than as the same original judgement. Here, the snowball grows because each stage's output is passed forward as though it were fresh evidence. Sequential review in the criminal justice system operates in a similar way, where a document passes through *k* stages of review, and each reviewer inherits its prior (i.e., the approvals of everyone preceding them). What this means is that those approvals are not neutral, in that they carry information, or at least appear to. Consequently, the rational response is to treat a document that has already been reviewed and approved as less likely to be flawed than one that has not been reviewed.

In this regard, the prediction is that each additional error category does not only add a new failure mode but further reduces the probability that existing failures will be detected, both because each class exploits a different gap in the review process and because each approval that a mismatched review process issues lowers the prior carried into every subsequent stage. This parallels the double-blind problem in eyewitness identification research (Wells & Olson, 2003), where procedural safeguards addressing one source of contamination can leave others entirely unmitigated if the interaction between sources is not anticipated. Xu et al. (2025) formalised this through learning-theoretic proofs demonstrating that hallucination is an innate limitation of LLMs used as general problem solvers rather than a defect amenable to engineering fixes (i.e., the compounding of error categories is therefore structural rather than incidental).

The systemic implications of this compounding problem become fully apparent when situated within Dror's (2025) analysis of bias cascade and bias snowball effects throughout the criminal justice system. Dror's central argument is that cognitive bias, the class of effects through which an individual's pre-existing beliefs, expectations, motives, and situational context influence the collection, perception, and interpretation of evidence during the course of a criminal case, within individual elements of the justice system has been researched and addressed primarily in silo, neglecting the interactions through which bias propagates across elements in ways that are more than their sum. Another way to understand this problem, is that a siloed approach to bias mirrors the siloed approach to AI error. Existing research on LLM hallucination has focused

predominantly on the accuracy of individual outputs (Ji et al., 2023; Huang et al., 2023) without examining how those errors propagate through the decision-making chains that characterize criminal justice processes (Berthet, 2022; Simon, 2012).

Dror (2025) adapts the Swiss cheese model (Reason, 1990) to illustrate why the criminal justice system should, in theory, be resilient to compounding bias at any single element. The model suggests that even when one element has a fault (i.e., a hole in a layer of cheese), it will not, by itself, result in an error in the final outcome, because subsequent elements will, or should, catch the fault through checks and balances. As shown in Figure 2 (Panel A) an error entering the system may pass through the first two layers (e.g., an AI-generated report and supervisory review) but encounter a layer where no corresponding vulnerability exists, thereby preventing propagation (A-1). Similarly, a different error may penetrate further through the system but is ultimately stopped before reaching the final outcome (B-1). For an innocent person to be wrongfully convicted, errors and biases would have to be aligned across multiple elements (e.g., police investigation, forensic examination, prosecutorial decision, trial proceedings) in a configuration that the model treats as an unlikely event.

However, as Dror (2025) demonstrates, this protection hinges entirely on the assumption that the different elements are relatively independent from each other. The reality is that supposedly independent elements actually work in parallel, influencing and biasing each other (Brook et al., 2021; Murrie et al., 2013; Simon et al., 2020), creating and enabling what Dror describes as bias cascade and bias snowball effects. Panel B demonstrates how AI-generated compound bullshit (i.e., output that is simultaneously factually fabricated, unfaithful to its source, and domain-specifically invalid) acts as the mechanism that drives both effects. In bias cascade (A-2), errors at earlier stages shift the vulnerabilities of later stages into alignment; for example, an AI-generated report containing fabricated details primes the expectations of a forensic examiner, effectively moving the hole in that layer to align with the holes in preceding layers, allowing the error to pass through where it would previously have been caught (A-1). In bias snowball (B-2), compounding cognitive biases (e.g., automation bias reinforcing the initial error, anchoring on the AI-generated output, confirmation bias selecting for consistent details, and belief perseverance rooting the narrative) amplify the error at each successive stage until it becomes powerful enough to penetrate safeguards that would ordinarily contain it. These mechanisms are

represented below, which is offered as a heuristic depiction of the cascade and snowball processes described in the cognitive-bias literature. It is not in itself evidence that AI-assisted policing workflows behave in the specific way the figure depicts; the alignment of vulnerabilities shown in Panel B should be read as a conceptual representation of what Dror's framework predicts, not as a measured pattern observed in any specific dataset. Future research should seek to test whether the alignment of vulnerabilities depicted here corresponds to observable patterns in operational policing settings, whether the cascade and snowball mechanisms operate at the rates and in the directions the framework predicts, and whether the introduction of AI-assisted documentation measurably shifts the alignment of vulnerabilities relative to comparable workflows in which AI is not used.

[Insert Figure 2 here]

One possible reason is the reliance on top-down information processing, in which prior expectations, experiences, ideology, and motivations shape if and how incoming information is processed (Kunda, 1990; Nickerson, 1998; Nisbett & Wilson, 1977; Wilson & Brekke, 1994). These top-down processes impact not only judgments but also what people perceive, including selective attention effects (Bruner et al., 1955; Balcetis & Dunning, 2006; Correll et al., 2002, 2007; de Lange et al., 2018; Desimone & Duncan, 1995; Neisser, 1967; Simmons & Chabris, 1999) and motivated perception (Balcetis & Dunning, 2006, 2010). Indeed, many of these top-down influences occur automatically and without awareness (Bargh et al., 1996; Eberhardt et al., 2004; Kunda, 1990; Nickerson, 1998), and people are relatively blind to their own biases (i.e., the blind spot; Pronin et al., 2002, 2004; Neal & Brodsky, 2016). The belief that expertise and experience reduce or immunize against cognitive bias is a widely held fallacy; on the contrary, experience can increase bias by enhancing the prevalence, power, and impact of top-down processes (Dror, 2011, 2020; Eeden et al., 2018). Similarly, the belief that technology and AI per se protect from bias is equally fallacious (Benjamin, 2019, 2023; Logg et al., 2019; Parasuraman & Riley, 1997; Skitka et al., 1999), as is the illusion of control. The belief that people can eliminate their biases through willpower, which can actually increase bias through ironic and rebound processing (Dror, 2025; Lynch et al., 2022; Macrae et al., 1994; Pronin et al., 2002; Pronin & Kugler, 2007; Wegner, 1994).

AI-generated compound bullshit therefore introduces a qualitatively new and uniquely dangerous vector into these cascade and snowball effects. For instance, at Level 1 (Data), the actual evidence can create empathy, sympathy, and motivation that bias observations and judgments. For example, it is only natural that horrific crimes and cases (e.g., child abuse) provoke emotions (Almada et al., 2021; Ask & Granhag, 2005; Bell & Loftus, 1989; Kensinger & Schacter, 2006; Kensinger et al., 2006). When an AI system generates a police report that fabricates or embellishes victim details (e.g., a faithfulness error that adds invented emotional content not present in the officer's input) it artificially amplifies the Level 1 biasing potential of the case data for every downstream reader. Case in point, when we consider this problem, the forensic pathologist, the prosecutor, the juror all respond emotionally to details that were never observed, but were generated by a system constitutively indifferent to whether those details are true (i.e., it generated plausible bullshit; Hicks et al., 2024).

In addition, Dror (2025) argues that bias drives the cognitive process rather than the actual evidence, creating backward and circular reasoning. The main principle is that the initial information has a great impact in priming and creating expectations and thus can bias if and how subsequent information is processed (Carlson & Russo, 2001; Chen et al., 2024; Darley & Gross, 1983; Lawson, 1967). For instance, early task framing can redirect attention from the evidence to a pre-specified hypothesis (e.g., implicit signals toward a preferred suspect can increase selection rates; Lindsay & Wells, 1985), and research has consistently shown that initial information sets the interpretive baseline for subsequent judgements (Tversky & Kahneman, 1974). More recent studies suggest a similar mechanism, particularly when managing task-agnostic triggers with LLMs (e.g., let's think step by step; Kojima et al., 2022). In practice, what this does, is change how the existing inputs are processed, serving as an anchoring analogous to hypothesis framing (i.e., it establishes a reference structure within which later information is evaluated; see, also Yin et al., 2025).

Policy and Practice Implications

The framework developed in this article carries direct consequences for how policing organisations procure, deploy, and oversee LLM-assisted technologies. For instance, procurement decisions should be conditional on whether the detection of bullshit can actually be discharged. Specifically, use cases dominated by task-specific error (e.g., risk assessment and

safeguarding) should not be deployed without domain-expert review of every output, because the detection such errors require cannot be discharged at the speed at which operational decisions are made. Use cases dominated by factuality error (e.g., report summarisation against verifiable records) are lower risk, provided that the verification step is built into the workflow rather than assumed. Supervisory review should be specified differently for each error class (external verification for factuality, source-output comparison for faithfulness, domain-expert evaluation for task-specific), and policing organisations should make explicit, documented decisions about which review applies to which output class. Where no such decision is made, the default review tends to be the cheapest available; typically a rapid read of the output alone. Because such a read consults no external ground truth, no original input, and no domain framework, it is unlikely to address any of the three error classes adequately. In this regard, a review process that has not been specified against an error class cannot be assumed to discharge the detection that class requires.

Documentation and oversight follow from these procurement and review decisions. Wherever AI-assisted material enters the record, audit trails should preserve both the original input and the generated output, alongside a record of the review actually performed. Without this preservation, each error becomes structurally undetectable in its own way: factuality errors when the ground-truth reference is not consulted, faithfulness errors when the input is discarded, and task-specific errors when domain-expert review is not applied. The compound case is the most severe, because compound detection requires all three review processes to be available simultaneously; the foreclosure of any one of them renders the entire compound undetectable, irrespective of whether the remaining processes are performed. External oversight bodies should also require policing organisations to specify, for each LLM use, which error class is dominant and which review resource is therefore the primary safeguard. Where that primary safeguard is domain-expert review and consistent application cannot be demonstrated, the use case should be treated as operating without an effective safeguard, regardless of the technology's reported accuracy. More broadly, the framework supports a taxonomy-before-technology principle, in that the categorisation of error and the specification of corresponding review should precede, rather than follow, the procurement of the technology that produces the error.

Limitations

Several limitations of the present analysis should be acknowledged. The first concerns the methodological approach of the framework itself. The Taxonomy of Bullshit was developed through a structured conceptual analysis rather than through systematic empirical sampling, and the conditions of inclusion that governed which sources were treated as relevant evidence were stated as conditions and not the kind that would govern a traditional systematic review (Page et al., 2021). The framework therefore makes claims about how AI-generated error should be classified and reasoned about; it does not make claims about how frequently each error class occurs in any given deployment. Estimating prevalence requires a different methodology that is not attempted here.

A second limitation concerns the empirical literature used for this framework. The policing studies that anchored the detectability approach were heavily concentrated in UK and US contexts. In addition, the categories that might be salient in other jurisdictions, in non-anglophone policing systems, or in policing functions that have not yet been the subject of published research may be underweighted in the present taxonomy. In addition, the rapidly evolving nature of the LLM literature meant that the framework relied heavily on preprint materials, which are identified as such in the reference list. While every effort was taken to assess the methodological quality of these sources, it is reasonable to assume that some may not ultimately meet peer review requirements, and the framework should be read with that caveat in mind. The framework is therefore best understood as a working framework that will need to be extended as the empirical literature on AI-assisted policing expands beyond its current geographical and functional concentration and as preprint sources progress through formal review. Finally, the compounding argument warrants particular caution, as it remains untested, and at the time of writing, the parameters that might enable testing have not been estimated in an operational policing setting.

Conclusion

This paper has developed four key contributions. Firstly, it has criticised the terminology used to describe AI-generated errors, and argues that the use of anthropomorphic language such as hallucination and confabulation are both theoretically incorrect, practically detrimental, and potentially malevolent (i.e., institutionally harmful). Because the language we use to describe AI errors shapes how policing interacts with these tools, calling errors hallucinations, albeit less so

confabulations, encourages practitioners to anthropomorphise the models, and implies that it is attempting to be accurate, thereby encouraging officers to treat its outputs as a reliable first draft, that requires only light editing. This, as a consequence, reframes the task from verification to revision, increasing the likelihood that inaccuracies are carried into the evidential records. We posit that Frankfurt's (2005) concept of bullshit is more appropriate, as it provides a more precise and practically useful characterisation of truth-indifference in respect to LLM outputs. For policing, this means a shift from an analytical perspective which would suggest that a system is trying but failing, to a position that considers whether its outputs are produced through processes that are structurally indifferent to evidential standards.

Secondly, the paper has developed a three-category taxonomy of AI truth-indifference operationalised for policing: *factuality errors, faithfulness errors, and task-specific errors*. Developing upon Huang et al. (2025) and Hicks et al. (2024), the taxonomy organises a body of LLM error typologies that has, until now, treated failure modes as descriptive categories rather than as meaningful classes for a policing context. Thirdly, the paper maps each category to the minimum review resource required to detect it, including external verification for factuality errors, source-output comparison for faithfulness errors, and domain-expert evaluation for task-specific errors. Importantly, the notion of detectability provides a diagnostic tool for practitioners, to identify what kind of error has occurred, what would be required to detect it, and where existing safeguards are likely to fail. It also provides a basis on which policing organisations can decide which use cases are appropriate for current LLM systems, which require additional review resource, and which should not be deployed at all without corresponding investment in domain-expert oversight. This is of practical importance, as noted earlier, given the motivation to deploy AI across policing (Home Office, 2026), and the need for policing to rely on auditable documentation (e.g., original reports, body-worn video), under the assumption that AI bullshits. Note, Ferguson (2025) highlighted that current audit of AI-generated content was not always available (also see Axon's guidelines; EFF, 2025).

Fourthly, the paper has argued that these error categories are seldom isolated and often co-occur, and it has stated the resulting compounding problem as a sequential Bayesian prediction. Compounding bullshit (i.e., output that is simultaneously a factuality error, a faithfulness error, and/or a task-specific error) exploits a distinct vulnerability that policing is not yet prepared for.

Because review stages rarely produce false alarms, every approval issued by a review process, including one incapable of establishing the error classes actually present, lowers the perceived prior probability of error being carried into the next stage. The predicted consequence is that detection becomes progressively less likely as errors pass through successive stages, cascading and snowballing through the criminal justice system (Dror, 2025). Qualitatively, this again represents a new vector, in which contamination through bullshit is constitutively indifferent to truth, and propagates through decision-making because it presents a synthetic, competent perception of professionalism.

When the compounding problem is situated in policing, a common solution put forward situates the problem in the code (i.e., we code out the bullshit). But, whilst there is some truth to that (e.g., RAG, fine-tuning and RLHF can reduce the rate of certain error types), they cannot eliminate the underlying truth-indifference. More fundamentally, however, the compounding issues outlined in this paper do not refer to the technological vectors; rather, they reflect AI/human-interactionism. Current systems described in the paper directly exploit cognitive vulnerabilities (e.g., confirmation bias, automation bias, anchoring bias) of the practitioners who receive, interpret and act upon these outputs. At present, there is no engineering solution that addresses the downstream [or upstream] impact on policing, given the time-critical, often ambiguous scenarios they operate in. It is reasonable to assume, under the time-critical and often ambiguous conditions that characterise operational policing, that officers are unlikely to simultaneously fact-check, compare and apply domain specific professional judgement to every AI-generated output (e.g., West Midlands Chief Constable, BBC, 2026).

The Policy and Practice Implications section developed above sets out the operational consequences of these arguments across procurement, supervision, documentation, and oversight. What follows from those implications, at a more general level, is that bias mitigation in AI-augmented criminal justice systems cannot be addressed by either technological refinement or procedural safeguard alone. On the one hand, addressing compound errors without addressing cascade allows AI-generated material to enter the system and propagate through subsequent decision points. On the other hand, addressing cascade without addressing compound errors leaves the output failures largely uncovered, allowing factually fabricated, source-unfaithful, and domain-invalid content to enter the criminal justice process before any cascade safeguard could

plausibly intervene. Both must be addressed together if the criminal justice process is to remain accountable to the evidential standards on which it depends. With that in mind, the criminal justice system must now contend not only with the biases of its human actors but with the outputs of systems that are, by design, indifferent to the truth those actors are sworn to pursue. The framework developed in this article is intended to support the prospective governance work that will be required.

The practical implication in this regard, is that bias mitigation in AI-augmented criminal justice systems requires a dual strategy. Firstly, research needs to examine the efficacy of a layered detection mechanism to address compound bullshit within individual AI outputs (e.g., simultaneous fact-checking, source comparison, and expert review). Secondly, these detection mechanisms must be combined with the systemic safeguard. Neither strategy alone is sufficient. Addressing compound errors without addressing cascade allows AI-generated bullshit to enter the system and propagate through every subsequent element, effectively contaminating review processes, prosecutorial decisions, and judicial proceedings. Further, addressing cascading errors without addressing compound errors leaves the output failures themselves uncovered, permitting triple indifference; output that is, in Frankfurt's (2005) terms, constitutively unconcerned with truth across every dimension simultaneously.

A taxonomy-before-technology principle would, perhaps, apply at all levels of error. In other words, a precise classification of how AI errors compound within outputs, and precise understanding of how those compound errors cascade and snowball through the criminal justice system (Dror, 2025; Cuellar et al., 2022; Kassin et al., 2013) are needed, before governance frameworks adequate to the risk, can be developed. The criminal justice system must now contend not only with the biases of its human actors but with the outputs of systems that are, by design, indifferent to the truth those actors are sworn to pursue.

References

**denotes pre-print material; † denotes conference material*

Abdullahi, S., Danyaro, K. U., & Chiroma, H. (2026). The rise of hallucination in large language models: systematic reviews, performance analysis and challenges. Cluster Computing, 29(2), 124.

Alderden, M., Freemon, K., & Rhee, Y. C. (2026). Public perceptions of police use of artificial intelligence to assist in generating police reports: findings from an experimental study. *Policing and Society*, 1-19.

Aljamaan, F., Temsah, M. H., Altamimi, I., Al-Eyadhy, A., Jamal, A., Alhasan, K., ... & Malki, K. H. (2024). Reference hallucination score for medical artificial intelligence chatbots: development and usability study. JMIR Medical Informatics, 12(1), e54345.

Almada, H., Passalacqua, N. V., Congram, D., & Pilloud, M. (2021). As forensic scientists and as people, we must not confuse objectivity with neutrality. Journal of Forensic Sciences, 66(5). https://doi.org/10.1111/1556-4029.14790

*Amiraz, C., Cuconasu, F., Filice, S., & Karnin, Z. (2025). The distracting effect: Understanding irrelevant passages in RAG. arXiv preprint arXiv:2505.06914.

Ask, K., & Granhag, P. A. (2005). Motivational sources of confirmation bias in criminal investigations: The need for cognitive closure. Journal of Investigative Psychology and Offender Profiling, 2(1), 43–63.

Ariso, J. M., & Bannister, P. (2025). 'AI lost the prompt!' Replacing 'AI hallucination' to distinguish between mere errors and irregularities. AI & SOCIETY, 1-13.

*Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024). Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930.

† Bao, G., Zhang, H., Wang, C., Yang, L., & Zhang, Y. (2025, January). How likely do LLMs with CoT mimic human reasoning? In Proceedings of the 31st International Conference on Computational Linguistics (pp. 7831–7850).

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. Journal of Personality and Social Psychology, 71(2), 230.

BBC News. (2026). Police chief admits misleading MPs after AI used in justification for banning Maccabi Tel Aviv fans [Live news]. BBC. https://www.bbc.com/news/live/c394zlr8e12t

*Beck, J., Eckman, S., Kern, C., & Kreuter, F. (2025). Bias in the loop: How humans evaluate AI-generated suggestions. arXiv preprint arXiv:2509.08514.

Belem, C. G., Pezeshkpour, P., Iso, H., Maekawa, S., Bhutani, N., & Hruschka, E. (2025, April). From single to multi: How LLMs hallucinate in multi-document summarization. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 5276–5309).

Bell, B. E., & Loftus, E. F. (1989). Trivial persuasion in the courtroom: The power of (a few) minor details. Journal of Personality and Social Psychology, 56(5), 669.

† Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5185–5198).

Benjamin, R. (Ed.). (2019). Captivating technology: Race, carceral technoscience, and liberatory imagination in everyday life. Duke University Press.

Benjamin, R. (2023). Race after technology. In Social theory re-wired. Routledge.

Berk, H. (2024). Beware of artificial intelligence hallucinations or should we call confabulation? Acta Orthopaedica et Traumatologica Turcica, 58(1), 1.

Biswas, R., Dutta, S., & Werth, D. (2025). Hallucinations in LLMs and resolving them: A holistic approach.

Bittner, E. (1970). The functions of the police in modern society: A review of background factors, current practices, and possible role models.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*(3), 214–234.

Brandon, M., Sidebotham, P., Bailey, S., Belderson, P., Hawley, C., Ellis, C., & Megson, M. (2012). New learning from serious case reviews: A two-year report for 2009-2011. Department for Education.

Brems, R. (2026). Transforming police reporting with generative AI: Insights from the Oklahoma City Police Department. *The Police Journal*, *99*(1), 86-102.

Brender, T. D. (2023). Chatbot confabulations are not hallucinations—Reply. JAMA Internal Medicine, 183(10), 1177–1178.

Brinsa, M. (2026, January). Hallucination rates in 2025 - Accuracy, refusal, and liability. Medium. https://medium.com/@markus_brinsa/hallucination-rates-in-2025-accuracy-refusal-and-liability-aa0032019ca1

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Bruner, J. S., & Minturn, A. L. (1955). Perceptual identification and perceptual organization. The Journal of General Psychology, 53(1), 21–28.

Bull, R., May, B., & Milne, R. (2025). What Is Memory? Investigative Interviewing: Psychology and Practice, 13.

Carlson, K. A., & Russo, J. E. (2001). Biased interpretation of evidence by mock jurors. Journal of Experimental Psychology: Applied, 7(2), 91.

Charlotin, D. (2025). AI hallucination cases database. https://www.damiencharlotin.com/hallucinations/

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. Journal of Medical Internet Research, 26(1), Article e53164.

† Chen, N., Liu, J., Dong, X., Liu, Q., Sakai, T., & Wu, X. M. (2024, December). AI can be cognitively biased: An exploratory study on threshold priming in LLM-based batch relevance assessment. In Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (pp. 54–63).

Chen, S., Gao, M., Sasse, K., Hartvigsen, T., Anthony, B., Fan, L., Aerts, H. J. W. L., & Bitterman, D. S. (2025). When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. npj Digital Medicine, 8(1), 605.

Cheng, M., Gligorić, K., Piccardi, T., & Jurafsky, D. (2024, March). AnthroScore: A computational linguistic measure of anthropomorphism. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 807–825).

† Cheng, V., & Yu, Z. (2023, October). Analyzing ChatGPT's mathematical deficiencies: Insights and contributions. In Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023) (pp. 188–193).

Chi, H., Li, H., Yang, W., Liu, F., Lan, L., Ren, X., & Han, B. (2024). Unveiling causal reasoning in large language models: Reality or mirage? Advances in Neural Information Processing Systems, 37, 96640–96670.

Child Safeguarding Practice Review Panel. (2024). Annual report 2023/24: Patterns in practice, key messages and 2024 to 2025 work programme. Child Safeguarding Practice Review Panel. https://www.gov.uk/government/publications/child-safeguarding-practice-review-panel-annual-report-2023-to-2024

Colbert, S. (Host). (2005, October 17). Truthiness.

*Coleman, S. (2025). Early Adoption of Generative Artificial Intelligence in Law Enforcement: A Mixed-Methods Study of Policy, Practice, and Perception. CrimRxiv.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. Journal of Personality and Social Psychology, 83(6), 1314.

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. Journal of Personality and Social Psychology, 92(6), 1006.

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. Journal of Legal Analysis, 16(1), 64–93.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. Journal of Personality and Social Psychology, 44(1), 20.

De Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? Trends in Cognitive Sciences, 22(9), 764–779.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. Annual Review of Neuroscience, 18(1), 193–222.

Dror, I. E. (2011). The paradox of human expertise: Why experts get it wrong. In N. Kapur (Ed.), The paradoxical brain (pp. 177–188). Cambridge University Press.

Dror, I. E. (2018). Biases in forensic experts. Science, 360(6386), 243-243.

Dror, I. E. (2020). Cognitive and human factors in expert decision making: Six fallacies and the eight sources of bias. Analytical Chemistry, 92, 7998–8004.

Dror, I. E. (2025). Biased and biasing: The hidden bias cascade and bias snowball effects. Behavioral Sciences, 15(4), 490.

Duenas, T., & Ruiz, D. (2024). Detecting and classifying LLM hallucinations: A framework for skill-specific error analysis. *ResearchGate.*

*Du, Y., Tian, M., Ronanki, S., Rongali, S., Bodapati, S., Galstyan, A., & Peng, H. (2025). Context length alone hurts LLM performance despite perfect retrieval. arXiv preprint arXiv:2510.05381.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: Race, crime, and visual processing. Journal of Personality and Social Psychology, 87(6), 876.

Eeden, C. A. J., de Poot, C. J., & van Koppen, P. J. (2018). The forensic confirmation bias: A comparison between experts and novices. Journal of Forensic Sciences, 64 (1), 120–126.

Electronic Frontier Foundation. (2025). Axon's Draft One is designed to defy transparency. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2025/07/axons-draft-one-designed-defy-transparency

† Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018, January). Runaway feedback loops in predictive policing. In Conference on Fairness, Accountability and Transparency (pp. 160–171). PMLR.

Ericson, R. V., & Haggerty, K. D. (1997). Policing the risk society. University of Toronto Press.

† Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., & Koyejo, S. (2025, October). SycEval: Evaluating LLM sycophancy. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (Vol. 8, No. 1, pp. 893–900).

Ferguson, A. G. (2025). Generative Suspicion and the Risks of AI-Assisted Police Reports. Nw. UL Rev., 120, 299.

Frankfurt, H. G. (2005). On Bullshit. Princeton, NJ: Princeton University Press.

Frankfurt, H. G. (2006). On truth. Alfred A. Knopf.

Fredrikzon, J. (2025). Rethinking error: "Hallucinations" and epistemological indifference. Critical AI, 3(1).

Frydl, K., & Skogan, W. (Eds.). (2004). Fairness and effectiveness in policing: The evidence. National Academies Press.

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review, 45, 105681.

Green, D. M. & Swets, J.A. (1966). Signal detection theory and psychophysics.

Halford, E. (2026). Police application of generative artificial intelligence: A scoping review. *Journal of Policing, Intelligence and Counter Terrorism*, 1-30.

Halford, E. (2026). Policing in the age of artificial intelligence: a simulation study modelling workforce automation scenarios. Police Practice and Research, 27(1), 164-187.

Halford, E., & Webster, A. (2024). Using chat GPT to evaluate police threats, risk and harm. International Journal of Law, Crime and Justice, 78, 100686.

*Hämäläinen, M. (2025). On psychology of AI—Does primacy effect affect ChatGPT and other LLMs? arXiv preprint arXiv:2504.20444.

Harcourt, B. E. (2006). Against prediction: Profiling, policing, and punishing in an actuarial age. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/A/bo4101022.html

*Harnad, S. (2024). Language writ large: LLMs, ChatGPT, grounding, meaning and understanding. arXiv preprint arXiv:2402.02243.

Harris, T. R. (2024). The wrong of bullshit. Social Epistemology, 38(4), 413–424.

Hatem, R., Simmons, B., & Thornton, J. E. (2023). A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. Cureus, 15(9).

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection theory: A user's guide*. Routledge.

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. Ethics and Information Technology, 26(38), 1-10.

Home Office. (2025). Microsoft 365 Copilot experiment: Cross-government findings report. UK Government. https://assets.publishing.service.gov.uk/media/683db42bd23a62e5d32680d0/M365_Copilot_Experiment_Findings_Report.pdf

Home Office. (2026). From local to national: A new model for policing. UK Government. https://www.gov.uk/government/publications/from-local-to-national-a-new-model-for-policing

*Hoorn, J. F., & Chen, J. J. Y. (2023). Epistemic considerations when AI answers questions for us. arXiv preprint arXiv:2304.14352.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), 1–55.

*Ibrahim, L., & Cheng, M. (2025). Thinking beyond the anthropomorphic paradigm benefits LLM research. arXiv preprint arXiv:2502.09192.

Järvilehto, L., Sun, Y., Korkman, J., Haginoya, S., & Santtila, P. (2026). Testing large language model capability in building rapport and interviewing children about a witnessed mock event. Journal of Psychology and AI, 2(1), 2608167.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. Journal of Applied Research in Memory and Cognition, 2(1), 42-52.

King County Prosecuting Attorney's Office. (2024, September 25). Memorandum regarding AI-assisted police report narratives [Memorandum]. King County Prosecuting Attorney's Office.

Kensinger, E. A., Garoff-Eaton, R. J., & Schacter, D. L. (2006). Memory for specific visual details can be enhanced by negative arousing content. Journal of Memory and Language, 54(1), 99–112.

Kensinger, E. A., & Schacter, D. L. (2006). Processing emotional pictures and words: Effects of valence and arousal. Cognitive, Affective, & Behavioral Neuroscience, 6(2), 110–126.

*Kim, Y., Jeong, H., Chen, S., Li, S. S., Park, C., Lu, M., & Breazeal, C. (2025). Medical hallucinations in foundation models and their impact on healthcare. arXiv preprint arXiv:2503.05777.

† King, A, & May, B. (2026, March 11–14). Beyond the loop: Why large language models fail to align with the decision-making framework for policing [Conference presentation]. American Psychology-Law Society (AP-LS) Annual Conference, Reno, NV, United States.

† Kleinberg, J., Oren, S., & Raghavan, M. (2017, June). Planning with multiple biases. In Proceedings of the 2017 ACM Conference on Economics and Computation (pp. 567–584).

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, 35, 22199–22213.

*Kovač, G., Perez, J., Portelas, R., Dominey, P. F., & Oudeyer, P. Y. (2025). Recursive training loops in LLMs: How training data properties modulate distribution shift in generated data? arXiv preprint arXiv:2504.03814.

Kücking, F., Hübner, U., Przysucha, M., Hannemann, N., Kutza, J. O., Moelleken, M., & Busch, D. (2024). Automation bias in AI-decision support: Results from an empirical study. In German Medical Data Sciences 2024 (pp. 298–304). IOS Press.

Kukucka, J., & Dror, I. E. (2023). Human factors in forensic science: Psychological causes of bias and error. *In D DeMatteo & K Scherr (eds), The Oxford Handbook of Psychology and Law.* https://doi.org/10.1093/oxfordhb/9780197649138.013.36.

Kukucka, J., Kassin, S. M., Zapf, P. A., & Dror, I. E. (2017). Cognitive bias and blindness: A global survey of forensic science examiners. Journal of Applied Research in Memory and Cognition, 6(4), 452-459.

Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108(3), 480.

*Li, J., Chen, J., Ren, R., Cheng, X., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2024). The dawn after the dark: An empirical study on factuality hallucination in large language models. arXiv preprint arXiv:2401.03205.

Lindo, M. R., & Panes, R. (2024). The hallucination effect: Correlating generative AI usage frequency with source verification habits among Grade 12 STEM researchers. IMCC Journal of Science, 4(2), 9–11.

Lindsay, R. C., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. Journal of Applied Psychology, 70(3), 556.

Lipsky, M. (2010). Street-level bureaucracy: Dilemmas of the individual in public service. Russell Sage Foundation.

† Lithgow-Serrano, O. W., Kanjirangat, V., & Antonucci, A. (2025, November). Causal understanding by LLMs: The role of uncertainty. In Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025) (pp. 208–228).

Liu, C., & Lin, C. (2026). Are AI hallucinations the same as error information? A qualitative exploration of hallucination tolerance and resistance strategies. Online Information Review, 1-19.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12, 157–173.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103.

Lou, J., & Sun, Y. (2026). Anchoring bias in large language models: An experimental study. Journal of Computational Social Science, 9(1), 11.

Lum, K., & Isaac, W. (2016). To predict and serve? Significance, 13(5), 14–19.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. Journal of Personality and Social Psychology, 67(5), 808.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. Journal of Empirical Legal Studies, 22(2), 216–242.

† Maleki, N., Padmanabhan, B., & Dutta, K. (2024, June). AI hallucinations: a misnomer worth clarifying. In 2024 IEEE Conference on Artificial Intelligence (CAI) (pp. 133-138). IEEE.

Manning, P. K. (2010). Democratic policing in a changing world. Routledge.

Marcus, G., & Davis, E. (2020). GPT-3, bloviator: OpenAI's language generator has no idea what it's talking about. *Technology Review*, 294.

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., & Griffiths, T. L. (2024). Large language models predict human sensory judgments across six modalities. Scientific Reports, 14(1), 21445.

May, B. (2025). Are we DrAIfting? The role of large language models in police investigative interviewing. *Investigative Interviewing: Theory to Practice, 15*(1), 1–21.

May, B., Milne, R., & Bull, R. (2025). Applying cognitive psychology to crime investigation. In D. Groome (Ed.), *An introduction to applied cognitive psychology* (3rd ed.). Psychology Press.

May, B., & Jones, M. A. (2026). The taxonomy of bullshit: Implementation challenges for AI in policing. *The Florida Chief*.

*Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. arXiv preprint arXiv:2005.00661.

Ministry of Justice. (2025). Independent review of the criminal courts: Part 1. https://www.gov.uk/government/publications/independent-review-of-the-criminal-courts-part-1

Neal, T., & Brodsky, S. L. (2016). Forensic psychologists' perceptions of bias and potential correction strategies in forensic mental health evaluations. Psychology, Public Policy, and Law, 22(1), 58.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220.

Omar, M., Sorin, V., Collins, J. D., Reich, D., Freeman, R., Gavin, N., & Klang, E. (2025). Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Communications Medicine, 5(1), 330.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410.

Police1. (2026). AI police report turns Heber City officer into a frog. *Police1*. https://www.police1.com/artificial-intelligence/utah-pd-testing-ai-report-writing-software-shares-comical-error-caused-by-the-princess-and-the-frog

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. Psychological Review, 111(3), 781.

Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, 43(4), 565–578.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self-versus others. Personality and Social Psychology Bulletin, 28(3), 369–381.

Pulvino, R., Sutton, D., & Naddeo, J. J. (2025). Hiding in plain sight: An empirical study of prosecutorial bias in AI legal analysis. Science and Technology Law Review, 27(1).

Rahman, A. A., Anwar, S., Usman, M., Ahmad, I., & Mian, A. (2025). DefAn: Definitive answer dataset for LLM hallucination evaluation. Information, 16(11), 937.

*Ravichander, A., Ghela, S., Wadden, D., & Choi, Y. (2025). Halogen: Fantastic LLM hallucinations and where to find them. arXiv preprint arXiv:2501.08292.

*Rawte, V., Sheth, A., & Das, A. (2023). A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. Consciousness and Cognition, 8(3), 338–342.

Richards, L. (2025, September 3). Response to media: The Domestic Abuse, Stalking and Harassment and so-called Honour Based Abuse Risk Identification, Assessment and Management Model (DASH 2009–2025). The Laura Richards.

Reason, J. (1990). Human error. Cambridge University Press.

Schnider, A. (2008). The confabulating mind: How the brain creates reality. Oxford University Press.

*Shao, J., Lu, Y., & Yang, J. (2025). Benford's curse: Tracing digit bias to numerical hallucination in LLMs. arXiv preprint arXiv:2506.01734.

*Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., & Perez, E. (2023). Towards understanding sycophancy in language models. arXiv preprint arXiv:2310.13548.

Shumaker, D. (2024). 'Hallucinate, confabulate, obfuscate': The state of AI today. Information Today, 41(1), 12–14.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. Perception, 28(9), 1059–1074.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5), 991–1006.

Smith, A. L., Greaves, F., & Panch, T. (2023). Hallucination or confabulation? Neuroanatomy as metaphor in large language models. PLOS Digital Health, 2(11), e0000388.

Song, P., Han, P., & Goodman, N. (2026). Large language model reasoning failures. Transactions on Machine Learning Research. https://doi.org/10.48550/arXiv.2602.06176

† Srivastava, T., Chou, J. C., Shroff, P., Livescu, K., & Graziul, C. (2024, December). Speech recognition for analysis of police radio communication. In 2024 IEEE Spoken Language Technology Workshop (SLT) (pp. 906-912). IEEE.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Buber, L., & Smyth, P. (2025). What large language models know and what people think they know. Nature Machine Intelligence, 7(2), 221–231.

Strauß, S. (2021). Deep automation bias: How to tackle a wicked problem of AI? Big Data and Cognitive Computing, 5(2), 18.

† Sui, P., Duede, E., Wu, S., & So, R. (2024, August). Confabulation: The surprising value of large language model hallucinations. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 14274–14284).

Sun, Y., Pang, H., Järvilehto, L., Zhang, O., Shapiro, D., Korkman, J., ... & Santtila, P. (2025). Comparing the performance of a large language model and naive human interviewers in interviewing children about a witnessed mock-event. PloS One, 20(2), e0316317.

*Sun, Z., Yim, W. W., Uzuner, O., Xia, F., & Yetisgen, M. (2025). A scoping review of natural language processing in addressing medically inaccurate information: Errors, misinformation, and hallucination. arXiv preprint arXiv:2505.00008.

Tao, J., Chism, K. A., Luo, F., del Carmen, A., & Bowling, B. (2026). Catalysts of generative AI adoption in policing: the primacy of transformational leadership and organizational capacity. Policing: An International Journal, 49(1), 90-108.

Telles-Correia, D., Moreira, A. L., & Gonçalves, J. S. (2015). Hallucinations and related concepts - their conceptual background. Frontiers in Psychology, 6, 991.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124-1131.

Urai, A. E., De Gee, J. W., Tsetsos, K., & Donner, T. H. (2019). Choice history biases subsequent evidence accumulation. eLife, 8, Article e46331.

*Varotsis, A. (2025). Measuring bias and variance in AI perceptions of policing risk. CrimRxiv.

Vered, M., Livni, T., Howe, P. D. L., Miller, T., & Sonenberg, L. (2023). The effects of explanations on automation bias. Artificial Intelligence, 322, Article 103952.

Walker, S. (2005). The new world of police accountability. Sage Publications.

*Wanner, M., Azzopardi, L., Thomas, P., Dan, S., Van Durme, B., & Craswell, N. (2025). All Claims Are Equal, but Some Claims Are More Equal Than Others: Importance-Sensitive Factuality Evaluation of LLM Generations. arXiv preprint arXiv:2510.07083.

Watts, S., White, M. D., & Malm, A. (2024). Automating body-worn camera footage review through AI: Baseline attitudes from a multisite randomized control trial. Policing: A Journal of Policy and Practice, 18, paae107.

White, M. D., Watts, S., Malm, A., & Navarrete, G. B. (2025). Officer Acceptance and Use of AI-Driven Body-Worn Camera Footage Review. American Journal of Criminal Justice, 1-22.

Woesle, C., Fischer-Brandies, L., & Buettner, R. (2025). A systematic literature review of hallucinations in large language models. IEEE Access.

*Xu, Z., Jain, S., & Kankanhalli, M. (2025). Hallucination is inevitable: An innate limitation of large language models. arXiv preprint arXiv:2401.11817. https://doi.org/10.48550/arXiv.2401.11817

*Yin, H., Vardi, S., & Choudhary, V. (2025). Fragile preferences: A deep dive into order effects in large language models. arXiv preprint arXiv:2506.14092.

*Yuan, J., Murukannaiah, P. K., & Singh, M. P. (2024). Right vs. Right: Can LLMs Make Tough Choices? arXiv preprint arXiv:2412.19926.

*Zheng, C., Zhou, H., Meng, F., Zhou, J., & Huang, M. (2023). Large language models are not robust multiple choice selectors. arXiv preprint arXiv:2309.03882.

Table 1

Exemplar of Factuality Error Sub-types

Sub-type	Definition	Example	Supporting Empirical Evidence
Entity fabrication	Generation of non-existent persons, organisations, locations, or case academic or non-legal references that have no basis in reality.	An LLM assisting with intelligence analysis references a non-existent organised crime group, cites a fabricated previous conviction, or invents a police force policy that does not exist.	Chelli et al. (2024); Watson (2024); Aljamaan et al. (2024); Li et al. (2024); Rahman et al. (2025); Mata v. Avianca, Inc. (2023)
Temporal fabrication	Incorrect dates, times, or sequences of events, which are particularly dangerous where timelines are evidentially important.	An AI system confidently states incorrect dates for legislative changes, sentencing guideline updates, procedural revisions, inverts the sequence of events when processing case materials, or changes time on mobile phone data.	Alansari & Luqman, 2025; Anh-Hoang, Tran, & Nguyen (2025); Huang et al. (2025); Abdullahi et al., 2026; Xu et al. (2024)
Legal fabrication	Invention or misstatement of statutes, case law, sentencing guidelines, or procedural requirements.	An LLM invents non-existent case law, misquotes sentencing guidelines, or fabricates procedural requirements for a jurisdiction that officers then	Dahl et al. (2024); Magesh et al. (2025); Charlotin (2025); Mata v. Avianca, Inc. (2023); Biswas et al. (2025); Ferguson (2025)

Sub-type	Definition	Example	Supporting Empirical Evidence
		follow in operational decision-making.	
Quantitative fabrication	Generation of fictitious statistics, measurements, or numerical data presented with unwarranted confidence.	Fabricated crime statistics, CompStat reports, intelligence reports, response time data, or performance metrics enter analytical reports and briefing documents without detection, particularly when recipients lack statistical expertise.	Shao et al. (2025); Ravichander et al. (2025); Chen & Yu (2023); Belem et al. (2025); Omar et al. (2025)
Attribution fabrication	Attributing statements, actions, or characteristics to the wrong individual; individual elements may each be accurate but relationships between them are fabricated.	An AI-generated police report attributes actions to an officer who was not at the scene, or assigns a witness's statement to the wrong individual in a multi-suspect investigation.	EFF (2025); Huang et al. (2025); May, 2025; Ferguson (2025); Heber City incident (2026); King County Prosecutor's Office (2024)

Table 2

Exemplar of Faithfulness Error Sub-types

Sub-type	Definition	Example	Supporting Empirical Evidence
Instruction non-compliance	The system deviates from explicit task parameters, ignoring specified frameworks, omitting required categories, or applying definitions inconsistent with the instructed methodology.	A prompt specifying that the LLM should assess risk using a specific methodology but yields output that uses an alternative assessment structure, omits required vulnerability categories, or applies threat definitions inconsistent with the policing framework.	Halford and Webster (2024); Young, Gillins, and Matthews (2025); Sharma et al. (2023); Zhou et al. (2023); Ouyang et al. (2022); Zhang et al. (2023)
Context degradation	The system loses track of earlier information in long documents or multi-turn interactions, producing outputs that contradict or fail to account for	An LLM processing an extensive case file for an investigation fails to connect early witness statements with later forensic evidence, or contradicts information provided at the beginning of a	Liu et al. (2025); Wan et al. (2025); Leng et al. (2024); Kim et al. (2025)

Sub-type	Definition	Example	Supporting Empirical Evidence
	previously provided context.	lengthy briefing document when generating conclusions at the end.	
Source material embellishment	The system adds plausible but unsupported details when processing source documents; the equivalent of an officer adding details to a witness statement that the witness did not provide.	An LLM summarising a witness interview transcript adds contextual details (e.g., the suspect appeared agitated) that were not stated in the original transcript, or strengthens hedged language (e.g., 'possibly involved' becomes 'involved') in ways that alter the evidentiary value of the account.	Peters and Chin-Yee (2025); Huang et al. (2023, 2025); Alansari et al. (2025)
Sycophantic validation	The system validates user assumptions or framing rather than faithfully processing the input, reinforcing errors or biases present in the prompt rather than correcting them against the source material.	An officer drafts a report framing a suspect as aggressive; the LLM, rather than faithfully reflecting the body-worn camera transcript which shows a calm interaction, reinforces the officer's characterisation by selecting and emphasising compatible details while downplaying contradictory evidence.	Sharma et al. (2023); Dahl et al. (2024); Chen et al. (2025); Fanous et al. (2025)
Input boundary confusion	The system fails to distinguish between relevant input and irrelevant noise in the provided source material, integrating extraneous content into outputs as though it formed part of the substantive input.	An AI report-writing system processing body-worn camera audio incorporates dialogue from a television programme playing in the background of a domestic incident, weaving fictional narrative content into an official police report as though it were part of the reported events.	Amiraz et al., 2025; Liu et al. (2025); Electronic Frontier Foundation (2025); Ferguson (2025); King County Prosecutor's Office (2024)

Table 3

Exemplar of Task-Specific Error Sub-types

Sub-type	Definition	Example	Supporting Empirical Evidence
Methodological misapplication	The system produces output that uses domain terminology and formatting conventions correctly but applies the underlying methodology incorrectly, misweighting factors, omitting required inferential steps, or applying assessment criteria in ways that violate the logic of the professional framework.	An LLM conducting a risk assessment (e.g., pre-sentence report) correctly identifies and lists the relevant risk factors but treats them additively (summing ticked boxes) rather than applying the required inferential judgment about patterns of coercive control, stalking behaviour, and escalation, producing an invalid risk classification.	Halford and Webster (2024); Turner, Medina, and Brown (2022); Richards (2025) (Note: Richards does not specifically discuss LLMs but emphasised that risk is not additive and requires professional inference about coercive control and escalation patterns — a judgment structure that, we argue, LLMs are not designed to replicate.)
Causal reasoning failure	The system identifies statistical associations or co-occurrences in data but fails to reason about causal mechanisms, producing analyses that confuse correlation with causation, miss confounding variables, or generate spurious causal narratives from coincidental patterns.	An LLM analysing intelligence data identifies that two persons of interest were in the same geographic area during three separate incidents and generates an analytical product asserting a collaborative relationship, when the co-occurrence reflects shared residential proximity rather than coordinated activity.	Bao et al. (2025); Chi et al. (2024); Sun et al. (2025); Kim et al. (2025)
Precedential analysis error	The system generates legal or procedural analysis that correctly identifies relevant authorities but mischaracterises the relationships between them, misrepresenting holdings, inventing precedential chains, or applying legal tests in ways that would not withstand professional scrutiny.	An LLM assisting with case file preparation correctly identifies relevant statutes and case law but mischaracterises the holding of a key precedent, states that Case A established a test that was actually established by Case B, or invents a precedential relationship between two unrelated decisions (i.e., errors invisible to non-legal reviewers).	Dahl et al. (2024); Magesh et al. (2025)

Sub-type	Definition	Example	Supporting Empirical Evidence
Risk quantification distortion	The system generates numerical risk scores, probability estimates, or categorical risk levels that appear precise and authoritative but reflect neither valid probabilistic reasoning nor the structured professional judgment that the assessment methodology requires.	An LLM assessing missing person risk generates a numerical risk score of 7.2/10 with apparent precision, when the underlying assessment framework requires categorical professional judgment (low/medium/high) informed by specific vulnerability indicators, producing output that feels more rigorous than expert assessment but lacks methodological validity.	Varotsis (2025); Halford and Webster (2024)
Safeguarding inference failure	The system correctly identifies individual risk indicators but fails to make the inferential connections between them that constitute professional safeguarding judgment, missing escalation patterns, failing to recognise that apparently low-level indicators collectively signal serious risk, or treating risk factors as independent when they are diagnostically interdependent.	An LLM reviewing a multi-agency case (e.g., MAPPA) file correctly identifies a safeguarding issue, but fails to recognise these as the signature pattern of child sexual exploitation, listing each as a separate low-level issue rather than identifying the inferential connection that would trigger a safeguarding referral.	Child Safeguarding Practice Review Panel (2024); Brandon et al. (2012)

Table 4

Decision Rule for Classifying AI-Generated Errors

Coding question	Error class	Review resource required	What cannot be established
Does any claim in the output conflict with verifiable external evidence, such as entities,	Factuality error	Comparison of the claim against an authoritative external source, such as	It cannot establish where the claim originated, and it cannot establish whether an

Coding question	Error class	Review resource required	What cannot be established
dates, statutes, statistics, or attributions?		records, databases, or legislation.	accurate claim misapplies a professional framework.
Does the output diverge from the task-defined input through additions, omissions, altered meaning, or the integration of extraneous material?	Faithfulness error	Side-by-side comparison of the generated output against the original input, instructions, and source materials.	It cannot on its own establish whether a divergence is material, which may require domain expertise, and it cannot establish whether the diverging content happens to be true.
Does the output misapply the professional framework, methodology, or inferential logic governing the task, irrespective of its accuracy and fidelity?	Task-specific error	Evaluation by a practitioner or expert competent in the relevant framework.	It cannot rule out fabrication or source divergence, both of which may be invisible without the external source or the original input, though an expert may notice either incidentally.