The Real Danger Is Not Artificial Intelligence but Human Stupidity: An Analytical Study of AI Misuse and Human Decision-Making

A Research Paper Submitted in the Field of Human–Computer Interaction and Artificial Intelligence Ethics

Author

Youssef Al-Amgad Rafat

AI Engineer, Instructor, and Researcher

youssefalamgad6@gmail.com

Independent Research Study

2026

Abstract

Artificial intelligence (AI) systems are now used in healthcare, transportation, recruitment, education, and cybersecurity, among many other fields. Public debate about AI risk usually focuses on the technology itself: how powerful it is, how autonomous it might become, and what it could theoretically do if left unchecked. This paper argues that this framing is incomplete. Many documented AI failures, from biased hiring tools to fatal autonomous vehicle crashes, occurred not because the underlying algorithms were mysterious or uncontrollable, but because the humans who built, deployed, or relied on them misunderstood, over-trusted, or failed to verify what the system produced. Using an interdisciplinary literature review, the paper develops a conceptual framework linking four variables: AI literacy, trust in AI, verification behavior, and decision quality, with automation bias treated as a moderating factor that weakens the link between having information and acting on it critically. Six real-world case studies (healthcare, autonomous vehicles, recruitment, generative AI and misinformation, education, and cybersecurity) are examined through this lens. A methodology is proposed for an empirical study of 150 participants, using a 22-item Likert-scale questionnaire and standard statistical procedures such as correlation, regression, and analysis of variance. No experiment has yet been conducted; all findings described here are expected outcomes based on the reviewed literature, clearly labeled as such. The paper concludes that AI risk management requires as much investment in human AI literacy, critical verification habits, and organizational oversight as it does in technical safety research. The central claim is not that humans are foolish, but that human factors, not the technology alone, determine whether AI is used safely or harmfully.

Keywords: artificial intelligence risk, human decision-making, automation bias, AI literacy, trust in AI, AI misuse, human-AI interaction, verification behavior

Contents

Introduction

Artificial intelligence has moved from research laboratories into everyday decision-making. Hospitals use predictive models to flag high-risk patients, banks use algorithms to score loan applications, companies use automated tools to screen job candidates, and millions of people use generative AI systems to write, summarize, and search for information. This rapid adoption has been accompanied by a familiar public narrative: that AI itself is the danger, a technology that could become too powerful, too autonomous, or too unpredictable for humans to control.

This paper takes a different, though not contradictory, position. It argues that most of the AI-related harms documented so far did not occur because a model became uncontrollable. They occurred because people misunderstood what the model could do, trusted its output more than they should have, skipped the step of checking its results, or deployed it in a setting for which it was never designed. In other words, the technology supplied the tool, but human decisions supplied the failure.

The title of this paper, borrowed loosely from a common saying about technology and human error, should not be read literally. It is not an argument that people are unintelligent. It is an academic shorthand for a specific and measurable set of human factors: low AI literacy, excessive or misplaced trust in automated systems, automation bias, insufficient verification of AI outputs, weak understanding of the limitations of a given model, misuse of AI outside its intended purpose, absent human supervision, and, more broadly, poor decision-making under uncertainty. Each of these factors has an established research literature behind it, reviewed in the sections that follow.

The purpose of this paper is threefold. First, it reviews the literature on AI risk, human error, automation bias, AI literacy, trust, human-AI interaction, and AI misuse, and organizes this literature into a single conceptual framework. Second, it examines six real-world domains where AI has caused documented harm, and asks in each case how much of the harm can be traced to the technology and how much to the humans who used it. Third, it proposes a methodology, including a questionnaire, for empirically testing the relationships in the framework, so that future researchers, or the same author in a later study, can move from literature-based argument to measured evidence.

The paper does not claim that AI is risk-free or that all responsibility lies with users. Some risks are technical: models can be biased by their training data, can hallucinate false information, and can behave unpredictably outside their training distribution. The paper's claim is narrower and, hopefully, more useful: that technical risk and human risk interact, and that a large share of the harm attributed to "AI" in public discussion is better explained as a joint failure of technology design and human judgment. Addressing AI safety, therefore, requires attention to both sides of that interaction.

Research Problem

Public and policy discussions about AI risk tend to concentrate on the capabilities of the technology: how accurate a model is, how large it is, or how close it might be to some hypothetical general intelligence. This focus is understandable, but it can obscure a simpler and more immediate problem. Many of the harms already caused by AI systems, wrongful arrests from facial recognition errors, biased hiring recommendations, medical algorithms that under-served minority patients, fatal crashes involving semi-autonomous vehicles, did not require a superintelligent or malicious system. They required an ordinary, imperfect model combined with a human operator or organization that trusted it too much, understood it too little, or failed to build in adequate checks.

This creates a research problem worth investigating directly: to what extent are AI-related harms explained by human factors, as opposed to purely technical ones, and which human factors matter most? Without a clear answer, organizations may continue to invest heavily in technical safety measures while underinvesting in training, oversight structures, and verification procedures that address the human side of the equation. The research problem addressed in this paper is therefore the need for a structured, evidence-based account of how AI literacy, trust, automation bias, and verification behavior combine to shape the quality of decisions made with AI assistance.

Research Questions

This study is organized around the following questions:

1. What role do human factors, specifically AI literacy, trust in AI, automation bias, and verification behavior, play in the occurrence of AI-related harms?
2. How does the level of a person's AI literacy relate to the degree of trust they place in AI-generated outputs?
3. How does trust in AI relate to the likelihood that a person will verify AI outputs before acting on them?
4. Does automation bias weaken the relationship between having relevant knowledge and engaging in verification behavior?
5. How do these human factors combine to influence the overall quality of decisions made with AI assistance?
6. How do these dynamics differ across domains such as healthcare, autonomous vehicles, recruitment, generative AI, education, and cybersecurity?

Research Objectives

The objectives of this paper are:

1. To review and synthesize the academic literature on AI risk, automation bias, AI literacy, trust in AI, human-AI interaction, and AI misuse.
2. To develop a conceptual framework linking AI literacy, trust in AI, verification behavior, and decision quality, with automation bias as a moderating factor.
3. To analyze six real-world case studies in which human factors contributed to AI-related harm.
4. To propose a methodologically sound, ethically appropriate study design, including a validated-style Likert questionnaire, capable of testing the framework empirically.
5. To generate practical recommendations for organizations, educators, and policymakers on reducing AI-related harm through improvements in human factors rather than technology alone.

Hypotheses

Based on the literature reviewed in the following sections, this paper proposes the following hypotheses for future empirical testing. These are stated as directional hypotheses consistent with prior findings on automation and trust; they are not results, since no data collection has taken place.

- **H1:** Higher AI literacy is associated with more appropriately calibrated (neither excessively high nor excessively low) trust in AI systems.
- **H2:** Higher trust in AI is associated with lower rates of independent verification of AI-generated outputs.
- **H3:** Higher levels of automation bias are associated with lower verification behavior, independent of AI literacy.
- **H4:** Verification behavior is positively associated with the quality of decisions made using AI assistance.
- **H5:** AI literacy has an indirect, positive effect on decision quality that operates through trust calibration and verification behavior.
- **H6:** The strength of these relationships varies by domain, with higher-stakes domains (for example healthcare) showing a stronger relationship between verification behavior and decision quality than lower-stakes domains.

Literature Review

AI Risks

The literature on AI risk generally separates two categories of concern. The first is long-term or systemic risk: the possibility that increasingly capable systems could behave in ways their designers did not intend, at a scale that is difficult to reverse. The second, and the focus of this paper, is near-term, well-documented risk arising from how current AI systems are built and used. Floridi and Cowls (2019) proposed a widely cited framework of five ethical principles for AI in society, beneficence, non-maleficence, autonomy, justice, and explicability, arguing that most harms trace back to a failure to operationalize one of these principles rather than to the technology being inherently dangerous. A related multi-stakeholder report, Floridi et al. (2018), reached a similar conclusion: AI risk is best understood as a socio-technical problem, not a purely technical one.

Weidinger et al. (2021) offered one of the more complete taxonomies of harm associated with large language models, including discrimination, information hazards, misinformation, and malicious use. Several of their harm categories are explicitly behavioral rather than computational: they depend on how a human chooses to use or interpret the system's output. Ji et al. (2023) documented the tendency of generative language models to produce fluent but false content, a phenomenon commonly called hallucination. Whether hallucination becomes harmful in practice depends heavily on whether a human downstream checks the claim before acting on it, which links the technical problem back to human verification behavior.

Human Error and Decision-Making

Long before AI became widespread, researchers studying human decision-making under uncertainty documented systematic biases in human judgment. Kahneman's (2011) account of two modes of thinking, a fast, intuitive system and a slower, effortful one, remains foundational for understanding why people often accept a plausible-looking answer without checking it. This matters for AI because outputs from a well-designed AI interface are typically presented in a fluent, confident style, which tends to trigger fast, intuitive acceptance rather than careful evaluation.

Automation Bias

Automation bias refers to the tendency of people to over-rely on automated systems, accepting their recommendations even when contradicting evidence is available, and to under-monitor systems that appear to be functioning correctly. Early empirical work by Skitka et al. (1999) found that participants using an automated decision aid missed events they would have caught unaided, because they had stopped actively verifying the system's suggestions. Cummings (2004) extended this concept

to safety-critical, time-pressured settings, arguing that automation bias becomes more dangerous, not less, as the perceived reliability of a system increases. Parasuraman and Manzey (2010) showed that automation-induced complacency and automation bias are related but distinct: complacency is a failure to monitor, while automation bias is an active preference for automated advice over one's own judgment.

Goddard, Roudsari, and Wyatt (2012) conducted a systematic review of automation bias specifically in clinical decision support systems and found that the effect was consistent across a range of healthcare settings, with correct clinical judgments sometimes overturned after receiving incorrect automated advice. Wickens, Clegg, Vieane, and Sebok (2015) showed experimentally that automation bias increases as automation reliability increases, a counterintuitive but well-replicated finding: the better a system usually performs, the less likely a human is to catch the rare cases where it fails. More recently, automation bias has been documented directly in AI-assisted image reading, where clinicians shown incorrect AI suggestions changed correct assessments to incorrect ones at a measurable rate (Pinto dos Santos et al., 2023).

AI Literacy

AI literacy has emerged as a distinct research area addressing the competencies people need to interact critically with AI systems. Long and Magerko (2020) defined AI literacy as a set of competencies that allow a person to critically evaluate AI technologies, communicate and collaborate effectively with them, and use them appropriately across different settings. Ng, Leung, Chu, and Qiao (2021) reviewed the emerging AI literacy literature and identified four recurring components: knowing and understanding AI, applying AI, evaluating and creating with AI, and understanding the ethical implications of AI use. Both bodies of work converge on the idea that AI literacy is not simply technical knowledge of how algorithms work; it also includes the practical judgment needed to know when a system's output should be questioned.

Trust in AI

Trust is the variable that connects AI literacy to behavior. Lee and See (2004) proposed an influential model of trust in automation, arguing that trust should be "calibrated": neither too high, which produces overreliance, nor too low, which produces underuse of a genuinely useful tool. Hoff and Bashir (2015) extended this model, distinguishing dispositional trust (a person's general tendency to trust machines), situational trust (trust shaped by the specific task and system), and learned trust (trust built or eroded through direct experience with a system over time). Jacovi, Marasović, Miller, and Goldberg (2021) argued that trust in AI should be understood as contractual: appropriate trust requires that the system's actual behavior match the expectations a human has formed about it, and

that mismatches, often invisible to the user, are a major source of misplaced trust.

Two contrasting findings are worth noting. Dietvorst, Simmons, and Massey (2015) found that people who observe an algorithm make even a single visible error often abandon it afterward, a pattern they termed algorithm aversion, even when the algorithm still outperforms human judgment on average. Logg, Minson, and Moore (2019), by contrast, found that in many judgment tasks people show algorithm appreciation, preferring algorithmic advice to human advice, particularly when the task is perceived as objective. Longoni, Bonezzi, and Morewedge (2019) found a domain-specific pattern in healthcare: patients are often more resistant to medical recommendations when they know they came from an algorithm rather than a human physician, even when the recommendation is identical. Together, these studies show that trust in AI is not a single, stable trait; it depends on domain, framing, and prior experience, which is why AI literacy and calibrated trust need to be studied together rather than assumed to move in the same direction.

Human-AI Interaction

The human-computer interaction literature has developed practical guidance for designing AI systems that support, rather than undermine, good human judgment. Amershi et al. (2019) proposed eighteen design guidelines for human-AI interaction, several of which address how a system should communicate uncertainty and support user oversight. Shneiderman (2020) argued for a "human-centered AI" approach that keeps meaningful human control over consequential decisions, combining high automation with high human oversight rather than treating them as trade-offs. Bansal et al. (2021) found experimentally that adding explanations to AI recommendations does not reliably reduce overreliance and can sometimes increase it, because explanations can make an incorrect recommendation feel more credible rather than prompting scrutiny. Buçinca, Malaya, and Gajos (2021) tested "cognitive forcing functions," interface features that require a user to make an initial judgment before seeing the AI's suggestion, and found that these functions reduced overreliance more effectively than explanations alone. This literature is directly relevant to the present paper because it shows that verification behavior can be designed for, not just hoped for.

AI Misuse

AI misuse refers to the deployment of a system outside its intended scope, without adequate testing, or without the safeguards its use case requires. Obermeyer, Powers, Vogeli, and Mullainathan (2019) analyzed a widely used healthcare risk-prediction algorithm and found that it systematically underestimated the health needs of Black patients because it used healthcare cost as a proxy for healthcare need, a design choice that reflected historical inequities in access to care rather than a flaw in the statistical technique itself. The harm arose from an assumption made by the humans

who selected the proxy variable, not from the algorithm failing to do what it was built to do. Dastin (2018) reported that a major technology company scrapped an internal AI recruiting tool after discovering it had taught itself to downgrade résumés containing the word "women's," reflecting biased patterns in the historical hiring data used to train it. Dressel and Farid (2018) evaluated a widely used criminal risk-assessment tool (COMPAS) and found its predictive accuracy was statistically indistinguishable from that of untrained members of the public making the same predictions from a short list of facts, raising questions about why the tool was treated as authoritative in judicial settings. Bender, Gebru, McMillan-Major, and Shmitchell (2021) warned that large language models can produce fluent, confident text without any grounding in truth, and cautioned against deploying such systems where users are likely to mistake fluency for reliability. Vaccaro, Sandvig, and Karahalios (2020) found that a lack of transparency about opaque, automated content-moderation decisions left users unable to meaningfully contest or understand decisions that affected them.

Theoretical and Conceptual Framework

The literature reviewed above suggests a chain of relationships rather than a single cause of AI-related harm. This paper organizes that chain into a conceptual framework with four core variables and one moderating variable.

AI Literacy is treated as the starting point. It reflects a person's understanding of what a given AI system can and cannot do, how it was trained, and what kinds of errors it is prone to make.

Trust in AI is shaped by AI literacy but not determined by it. Following Lee and See (2004) and Hoff and Bashir (2015), trust is treated here as a variable that can be well calibrated (matching the system's actual reliability) or poorly calibrated (too high or too low relative to actual reliability).

Verification Behavior is the practical, observable step of checking an AI output against another source, one's own expertise, or a second method, before acting on it. This is the behavioral variable most directly implicated in the case studies reviewed later in this paper.

Decision Quality is the outcome variable: the accuracy, appropriateness, and safety of the final decision made with AI assistance.

Automation Bias is modeled as a moderating factor that can weaken the relationship between having relevant knowledge (AI literacy) and acting on it critically (verification behavior). Even a person with good AI literacy can still exhibit automation bias under time pressure, high workload, or repeated exposure to a system that has performed reliably in the past (Parasuraman & Manzey, 2010; Wickens et al., 2015).

Figure 1 shows this framework. The main path runs from AI literacy through trust in AI and verifi-

cation behavior to decision quality. Automation bias is shown as a separate box that can weaken the links between trust and verification, and between verification and decision quality, without being reducible to low AI literacy alone.

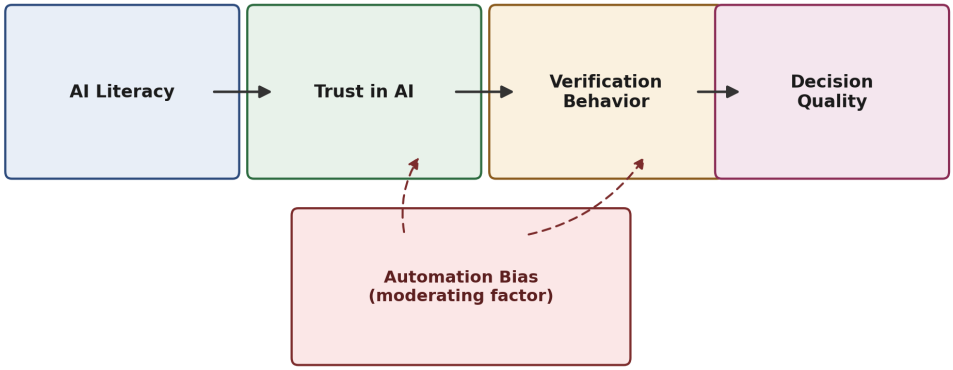

Figure 1: Conceptual Framework of the Study

This framework does not claim that AI literacy is the only input that matters. Organizational culture, time pressure, workload, and the design of the AI interface itself all plausibly influence trust and verification behavior. These are treated as contextual factors in the discussion section rather than as core variables, both to keep the proposed study manageable and because the four core variables already have the strongest and most consistent support in the literature reviewed above.

Human Factors Behind AI Risks

Building on the framework above, this section summarizes the specific human factors most often implicated in AI-related harm, drawing directly on the reviewed literature.

Low AI literacy. Users who do not understand basic facts about how a model was trained, what data it saw, or what "confidence" in a generated answer actually means are poorly positioned to judge when to trust it (Long & Magerko, 2020; Ng et al., 2021).

Excessive trust in AI. When trust exceeds a system's actual reliability, users stop applying independent judgment. This is especially likely when a system communicates its outputs with unwarranted confidence or fluency (Bender et al., 2021; Jacovi et al., 2021).

Automation bias. Independent of general trust, automation bias describes a specific behavioral tendency to accept automated recommendations and to under-monitor system performance, an effect that has been shown to increase, not decrease, as perceived system reliability rises (Skitka et

al., 1999; Wickens et al., 2015).

Failure to verify AI outputs. Verification is the behavioral safeguard that can catch errors regardless of their source. Its absence has been documented directly in radiology (Pinto dos Santos et al., 2023) and clinical decision support (Goddard et al., 2012), and is plausible in any domain where AI outputs are acted on quickly.

Poor understanding of AI limitations. Many AI failures occur when a system is used for a task that falls outside the distribution of data it was trained on, or in a context that differs from its intended use case, a distinct problem from simply trusting the system too much within its proper scope.

AI misuse. This refers to organizational or individual decisions to deploy a system without adequate testing, oversight, or fit to the task, as in the recruiting tool case documented by Dastin (2018) or the proxy-variable problem documented by Obermeyer et al. (2019).

Lack of human supervision. Several documented harms occurred in settings where no meaningful human review step existed between the AI output and the consequential action, removing any opportunity to catch an error before it caused harm.

Poor decision-making under uncertainty. Finally, general human decision-making tendencies, such as preferring fast, intuitive judgments over slower, effortful checking (Kahneman, 2011), interact with all of the factors above and make careful verification effortful even when a person knows, in principle, that they should verify.

Real-World Case Studies

This section examines six domains in which AI-related harm has been documented, and asks in each case what combination of technical and human factors contributed to the outcome.

Healthcare

The health-risk prediction algorithm analyzed by Obermeyer et al. (2019) is one of the most cited examples in this literature. The algorithm was designed to identify patients who would benefit from extra care management, and it performed exactly as it was built to perform: it predicted healthcare costs accurately. The harm arose because the humans who designed and adopted the system treated healthcare cost as an acceptable proxy for healthcare need, without adequately testing whether that assumption held across patient groups. It did not, because Black patients had historically faced barriers to accessing care, and therefore had lower recorded costs than white patients with comparable health needs. Separately, automation bias in clinical decision support (Goddard et al., 2012) and in AI-assisted mammography reading (Pinto dos Santos et al., 2023) shows that even accurate systems

can produce harm if clinicians defer to an incorrect suggestion rather than exercising independent judgment. In both strands of this case, the technology performed as designed; the harm traces to a proxy-variable decision and to a lack of consistent verification by human reviewers.

Autonomous Vehicles

The 2018 fatal collision between a self-driving test vehicle and a pedestrian in Tempe, Arizona, investigated by the National Transportation Safety Board (2019), is frequently cited as a case of AI failure. The Board's investigation found that the vehicle's system did detect the pedestrian, but classified her inconsistently and did not brake in time, partly because emergency braking had been disabled in that mode of operation, a decision made by the humans responsible for the test program. The investigation also found that the human safety driver, whose role existed specifically to catch this kind of system failure, was not watching the road at the critical moment. This case illustrates the framework directly: a technical limitation (imperfect object classification) combined with an organizational decision (disabling automatic braking) and a supervision failure (an inattentive driver) to produce a fatal outcome that no single factor would have caused alone.

Recruitment

The AI recruiting tool described by Dastin (2018) was trained on ten years of historical résumés submitted to a technology company, a period in which the industry was male-dominated. The system learned to associate male-coded language and experience with successful candidates, and downgraded résumés containing terms like "women's," as in "women's chess club captain." The technology performed as machine learning systems generally do: it found patterns in the training data and reproduced them. The human failure was in the choice of training data and in the absence of adequate testing for discriminatory patterns before the tool was used on real candidates. The company ultimately discontinued the tool after discovering the bias, which is itself an example of the kind of verification step that should ideally occur before deployment rather than after.

Generative AI and Misinformation

Generative language models are prone to hallucination: producing fluent, plausible-sounding statements that are factually incorrect (Ji et al., 2023). Bender et al. (2021) warned that this risk is compounded by the fluent, human-like style of these systems' output, which can make false statements feel more credible than they would if presented in a more obviously uncertain format. The resulting harms, from students citing invented sources to professionals submitting AI-drafted material containing fabricated facts, generally require a second step beyond the model's error: a human choosing not to verify the output before using or publishing it. The technical limitation (hallucination) is real and well documented, but its real-world impact depends heavily on whether verification

behavior occurs downstream.

Education

In education, the same technology can be used constructively or carelessly. Long and Magerko (2020) and Ng et al. (2021) argue that AI literacy determines much of this difference: students and educators with a clear understanding of how generative AI systems work are better positioned to use them as drafting or brainstorming aids while still applying independent judgment, whereas those with lower AI literacy are more likely to treat AI output as a finished, authoritative answer. Concerns about over-reliance on AI for writing and problem-solving in educational settings mirror the automation bias literature discussed earlier: the risk is not that AI tools are used, but that they are used without the verification and critical evaluation that meaningful learning requires.

Cybersecurity

Cybersecurity presents a two-sided version of this case. On one hand, AI-based detection tools can miss novel attacks that fall outside their training patterns, or can be evaded by attackers who deliberately craft inputs to avoid detection. On the other hand, organizations that treat an AI-based security tool as a complete replacement for human security analysts, rather than as an aid that still requires monitoring, testing, and periodic review, create exactly the kind of automation bias and supervision gap documented in other domains (Parasuraman & Manzey, 2010; Cummings, 2004). Attackers, meanwhile, have also begun using generative AI to make phishing and social engineering attempts more convincing, which shifts part of the risk back onto human recipients' verification habits: a well-crafted AI-generated message is more likely to succeed against a reader who does not pause to verify sender identity or check for other warning signs.

Human Responsibility versus AI Responsibility

The case studies above show that responsibility for AI-related harm is rarely located entirely on one side of the human-technology relationship. It is useful, however, to distinguish the kinds of responsibility that fall more naturally on each side, since this distinction has practical implications for where remedies should be targeted.

Responsibility that falls primarily on the technology and its developers includes: the statistical properties of a model (its accuracy, its error patterns, its behavior on data unlike its training set), the transparency of its outputs (whether it communicates uncertainty honestly), and the presence or absence of safeguards such as content filters, confidence indicators, or built-in limitations on high-stakes use.

Responsibility that falls primarily on human users and organizations includes: the decision to deploy a system in a given context, the choice of proxy variables and training data, the design of oversight and review processes, the training given to end users, and the individual decision, in any given moment, to verify an output before acting on it.

Zerilli, Knott, Maclaurin, and Gavaghan (2019) discuss this boundary directly, arguing that concerns about a supposed "responsibility gap" in algorithmic decision-making are often overstated, because meaningful human responsibility, for deployment decisions, oversight design, and final action, typically remains identifiable even when the algorithm's internal reasoning is not fully transparent. The European Commission's High-Level Expert Group on Artificial Intelligence (2019) reached a similar conclusion in its ethics guidelines, recommending that human oversight be treated as a core requirement for trustworthy AI rather than an optional add-on. This paper's position follows the same logic: treating AI risk purely as a technical problem, or purely as a human problem, both miss the interaction that actually produces harm. The practical implication is that risk mitigation efforts should be split roughly evenly between improving system design (better uncertainty communication, better testing before deployment) and improving human factors (AI literacy training, verification habits, and organizational oversight structures), rather than concentrating resources on one side alone.

Proposed Methodology

Research Design

This paper proposes a quantitative, cross-sectional survey design intended to test the relationships shown in Figure 1. The design is correlational rather than experimental, since the goal is to measure how AI literacy, trust, automation bias, and verification behavior naturally relate to one another and to self-reported decision quality among people who already use AI tools in their work or studies. This design is realistic for a first study in this area and can be extended later with an experimental component (see Future Work).

Participants

The proposed study would recruit approximately 150 participants (a feasible range of 100 to 200), drawn from adults who use AI tools regularly in at least one of the six domains discussed in this paper: healthcare, transportation or engineering, human resources or recruitment, general knowledge work involving generative AI, education, or cybersecurity or IT. A mixed sample across domains and experience levels would allow the domain-based comparison implied by Hypothesis 6. Recruitment would use a combination of professional networks, university mailing lists, and online

professional communities, with quota sampling used to ensure that no single domain makes up more than roughly 30 percent of the sample. Participation would be voluntary and anonymous, with informed consent obtained before the questionnaire is administered, consistent with standard research ethics requirements for studies involving human participants.

Because the study involves only anonymous self-report data, minimal-risk ethical review would typically apply. Participants would be informed of the study's purpose, told participation is voluntary, and given the option to withdraw at any point before submitting the questionnaire. No deception is involved.

Measures

Four latent variables would be measured using multi-item Likert scales (see the proposed questionnaire below): AI literacy, trust in AI, automation bias, and verification behavior. Decision quality would be measured using a small set of self-report items asking participants to reflect on recent AI-assisted decisions in their domain, since directly observing objective decision quality across six different domains is not feasible within a single survey-based study. This is a limitation acknowledged explicitly in the Limitations section.

Proposed Statistical Analysis

The following analyses are proposed, all standard and appropriate for the type of data this design would generate:

- **Descriptive statistics** (means, standard deviations, frequency distributions) to summarize the sample and each scale.
- **Reliability analysis** (Cronbach's alpha) for each multi-item scale, to confirm internal consistency before further analysis.
- **Correlation analysis** (Pearson's r) to examine the bivariate relationships between AI literacy, trust in AI, automation bias, verification behavior, and decision quality.
- **Multiple regression analysis** to test whether AI literacy and automation bias predict verification behavior, and whether verification behavior predicts decision quality, controlling for relevant demographic variables such as years of AI tool use.
- **Moderation analysis** to test whether automation bias moderates the relationship between AI literacy and verification behavior, consistent with Hypothesis 3.
- **Independent-samples t-tests** to compare verification behavior and decision quality between participants with high versus low self-reported AI literacy (using a median split).
- **One-way ANOVA** to compare mean verification behavior and decision quality scores across the six domains, testing Hypothesis 6.

All analyses would be conducted using standard statistical software (for example, SPSS, R, or Python's statsmodels library), with an alpha level of .05 used as the conventional threshold for statistical significance.

Proposed Questionnaire

The following 22-item questionnaire is proposed for the empirical study described above. All items would use a five-point Likert scale (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree), unless otherwise noted. Items marked (R) would be reverse-scored during analysis.

Section A: AI Literacy (6 items)

No.	Item
1	I understand, in general terms, how the AI tools I use are trained.
2	I can explain what an AI system's "confidence score" or "probability" actually means.
3	I know that AI systems can produce incorrect or fabricated information.
4	I understand that an AI tool's performance can vary depending on the type of task.
5	I feel confident identifying situations where an AI tool is not appropriate to use.
6	I regularly seek to learn more about how AI tools work.

Section B: Trust in AI (5 items)

No.	Item
7	I generally trust the outputs produced by AI tools I use.
8	I would follow an AI system's recommendation even if it seemed slightly unusual to me.
9	I believe AI systems are usually more accurate than my own judgment in the tasks I use them for.
10	My trust in a specific AI tool has increased the more I have used it.

No.	Item
11	I trust AI-generated information as much as information from a human expert in the same field.

Section C: Automation Bias (5 items)

No.	Item
12	When an AI tool gives me an answer, I usually accept it without checking further.
13	I tend to assume that an AI system has already considered information I might otherwise check myself.
14	Under time pressure, I am more likely to accept an AI recommendation without question.
15	I sometimes notice that I stop paying close attention once an AI tool has been reliable for a while.
16	I find it mentally easier to accept an AI answer than to work through the problem myself.

Section D: Verification Behavior (5 items)

No.	Item
17	I check AI-generated information against another source before relying on it.
18	I ask AI tools to explain their reasoning when the task is important.
19	I compare AI output with my own independent judgment before making a final decision. (R)
20	I have caught an AI tool making a mistake because I checked its output.
21	I have a consistent habit of reviewing AI-assisted work before submitting or acting on it.

Section E: Decision Quality (self-report, 1 item + 1 open item)

No.	Item
22	Looking back at recent decisions where I used AI assistance, I am confident those decisions were sound.

An optional open-ended item would also be included: "Please briefly describe one situation in which an AI tool you used produced an incorrect or misleading result, and how you responded." This qualitative item is not part of the Likert scale scoring but would provide illustrative context for the quantitative results.

A short demographic section (age range, domain of AI use, years of experience using AI tools, and frequency of AI tool use) would precede the main questionnaire, to support the domain-based ANOVA and regression control variables described above.

Expected Findings and Implications

No data have been collected for this paper, and the following statements describe expected findings based on the patterns reported in the literature reviewed above, not actual results. Any numbers given below are hypothetical illustrations, clearly labeled as such, intended to show what a plausible pattern of results might look like.

Based on the automation bias and trust literature (Skitka et al., 1999; Wickens et al., 2015; Parasuraman & Manzey, 2010), it is expected that AI literacy and verification behavior will show a positive correlation, while automation bias and verification behavior will show a negative correlation. As a purely illustrative example only, a pattern such as "AI literacy correlated positively with verification behavior (example: r = .35), while automation bias correlated negatively with verification behavior (example: r = -.40)" would be consistent with prior research; these values are hypothetical placeholders, not findings.

It is also expected that the moderation analysis will show that automation bias weakens the positive relationship between AI literacy and verification behavior, consistent with Hypothesis 3: even participants with strong AI literacy may show reduced verification behavior when their automation bias score is high, echoing the finding that automation bias can persist even among experienced or knowledgeable users (Cummings, 2004; Wickens et al., 2015).

Domain comparisons (Hypothesis 6) are expected to show higher verification behavior in higher-stakes domains such as healthcare, consistent with the emphasis on checking procedures documented in clinical decision support research (Goddard et al., 2012), compared to lower-stakes domains such as general knowledge work.

If these expected patterns are confirmed, the practical implication is that AI literacy training alone may be insufficient to improve safety outcomes, because automation bias can suppress verification behavior even among literate users. This would support interventions that change behavior directly, such as the cognitive forcing functions described by Buçinca et al. (2021), rather than relying on education alone.

Discussion

The literature reviewed in this paper, together with the six case studies, supports a central and fairly simple claim: the risk associated with AI in practice is jointly produced by the properties of the technology and the behavior of the humans who build, deploy, and use it. This is not a novel philosophical position, several of the sources cited here (Floridi & Cowls, 2019; Zerilli et al., 2019; the European Commission's High-Level Expert Group, 2019) make versions of the same argument, but it is a claim that deserves to be stated plainly and tested empirically, because public discussion of AI risk still leans heavily toward framing the technology itself as the primary actor.

The proposed framework is useful because it breaks a vague concept, "human error," into measurable components. AI literacy, trust, automation bias, and verification behavior can each be assessed with reasonably well-validated survey instruments, and the relationships between them can be tested with standard statistical tools. This matters because it turns "be more careful with AI" from a vague piece of advice into a set of specific, addressable targets: literacy training, trust calibration, verification habits, and interface designs that support all three.

It is also worth being honest about the limits of this argument. Not every AI harm is reducible to a human factor. Some risks are genuinely technical, for example, a model's vulnerability to adversarial inputs, or its tendency to produce different outputs for logically equivalent prompts, and no amount of user verification will fully compensate for a poorly designed or poorly tested system. The claim of this paper is not that humans are always at fault, but that human factors are underweighted relative to technical factors in current public and organizational risk management, and that correcting this imbalance would likely reduce real-world harm.

Recommendations

Based on the literature and framework developed in this paper, the following recommendations are proposed for different stakeholder groups.

For organizations deploying AI systems: Build mandatory human review steps into any workflow where an AI output leads to a consequential action, particularly in healthcare, hiring, and safety-

critical settings. Test systems for bias and failure modes using data and scenarios that reflect the population the system will actually affect, not only the population it was trained on, following the lesson of Obermeyer et al. (2019) and Dastin (2018).

For interface and product designers: Apply established human-AI interaction guidelines (Amershi et al., 2019) and consider cognitive forcing functions or similar design patterns that prompt independent judgment before revealing an AI recommendation (Buçinca et al., 2021), since explanations alone have not been shown to reliably reduce overreliance (Bansal et al., 2021).

For educators and AI literacy programs: Teach not only how to use AI tools but also their known failure patterns, including hallucination, bias from training data, and the conditions under which automation bias is most likely to occur, so that learners recognize these situations in their own behavior.

For individual users: Adopt a consistent verification habit for any AI output used in a consequential decision, particularly checking outputs against an independent source when the task is unfamiliar, high-stakes, or time-pressured, the exact conditions under which automation bias has been shown to increase (Cummings, 2004).

For policymakers and regulators: Support requirements for human oversight in high-risk AI applications, consistent with the trustworthy AI guidelines proposed by the European Commission's High-Level Expert Group (2019), and fund further research connecting human factors to measurable safety outcomes.

Limitations

This paper has several limitations that should be acknowledged directly. First, and most importantly, no empirical data have been collected; the proposed study design, questionnaire, and expected findings are based on a synthesis of existing literature and have not been tested. Second, the self-report nature of the proposed decision quality measure is a genuine constraint: participants' own judgments about whether their AI-assisted decisions were sound may not match objective outcomes, and a fully rigorous test of the framework would ideally include objective outcome data where feasible. Third, the six case studies were selected because they are well documented in the existing literature, which means they may not be representative of the full range of AI applications, particularly less publicized or less severe incidents that do not receive the same media or academic attention. Fourth, the proposed cross-sectional design can establish correlations among the framework's variables but cannot establish causal direction; a longitudinal or experimental design would be needed to confirm, for example, that low AI literacy causes poor verification behavior rather than the reverse. Finally, the framework itself, while grounded in the literature, simplifies a com-

plex reality into four core variables and one moderator; contextual factors such as organizational culture, time pressure, and interface design plausibly also matter and were treated only briefly in the discussion.

Future Work

Several directions for future work follow naturally from these limitations. The proposed questionnaire could be administered to test the framework empirically, ideally alongside objective measures of decision quality where a domain allows for it, such as diagnostic accuracy in a healthcare simulation task. An experimental extension could manipulate automation bias directly, for example by varying the reliability or framing of an AI tool shown to participants, to test causal claims that the current correlational design cannot support. A longitudinal design could track how trust in a specific AI tool changes over repeated use, addressing the "learned trust" dimension identified by Hoff and Bashir (2015). Finally, future research could test specific interface interventions, such as the cognitive forcing functions studied by Buçinca et al. (2021), directly against the human factors measured in this framework, to identify which interventions most effectively improve verification behavior without unnecessarily reducing legitimate, appropriate trust in AI tools that are, in fact, reliable.

Conclusion

This paper has argued that the risks most commonly associated with artificial intelligence are, in a large share of documented cases, produced jointly by the technology and by the humans who build, deploy, and use it. Reviewing the literature on AI risk, automation bias, AI literacy, trust in AI, human-AI interaction, and AI misuse, together with six real-world case studies spanning healthcare, autonomous vehicles, recruitment, generative AI, education, and cybersecurity, supports a conceptual framework in which AI literacy shapes trust, trust and automation bias shape verification behavior, and verification behavior shapes the quality of the final decision. None of this diminishes the importance of technical AI safety research; some risks genuinely originate in model design and cannot be fixed by better human behavior alone. But it does suggest that current efforts to manage AI risk are incomplete if they focus on the technology while treating human understanding, trust calibration, and verification habits as secondary concerns. The title of this paper is best read not as a judgment on human intelligence, but as a call to take human factors as seriously as technical ones: the tools people build matter, but so does how carefully people learn to use them.

References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complacency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445717

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), Article 188. https://doi.org/10.1145/3449287

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. In *AIAA 1st Intelligent Systems Technical Conference*. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2004-6313

Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women.* Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

European Commission High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI.* European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 624–635). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445923

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), Article 248. https://doi.org/10.1145/3571730

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

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. https://doi.org/10.1016/j.obhdp.2018.12.005

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

National Transportation Safety Board. (2019). *Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018* (Highway Accident Report NTSB/HAR-19/03). https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Pinto dos Santos, D., Giese, D., Brodehl, S., Chon, S. H., Staab, W., Kleinert, R., Maintz, D., & Baeßler, B. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology, 307*(4), Article e222176. https://doi.org/10.1148/radiol.222176

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction, 36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Vaccaro, K., Sandvig, C., & Karahalios, K. (2020). "At the end of the day Facebook does what it wants": How users experience contesting algorithmic content moderation. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), Article 167. https://doi.org/10.1145/3415238

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from language models* (arXiv:2112.04359). arXiv. https://doi.org/10.48550/arXiv.2112.04359

Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and automation bias in the use of imperfect automation. *Human Factors, 57*(5), 728–739. https://doi.org/10.1177/0018720815581940

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic decision-making and the control problem. *Minds and Machines, 29*(4), 555–578. https://doi.org/10.1007/s11023-