

Understanding Human-AI Collaborations: A Survey of Trust, Dependency, and Sociotechnical Risks

Md Foyzal Ahmed, Md Main Uddin Rony, Divya S

Bowling Green State University
Bowling Green, Ohio, USA
mdfoysa@bgsu.edu, mrony@bgsu.edu, divyas@bgsu.edu

Abstract

Human-AI collaboration is rapidly transforming decision-making across domains, yet it introduces a complex set of cyber-social risks that extend beyond traditional technical concerns such as bias and opacity. This paper synthesizes prior literature to provide a structured understanding of these risks across five dimensions: cognitive, emotional and psychological, social and organizational, trust and information, and governance and accountability. We examine how challenges such as cognitive overload, AI anxiety, invisible labor, epistemic misalignment, and responsibility gaps emerge from the interaction between human cognition, system design, and institutional contexts. We also review current mitigation strategies, such as explainable AI, uncertainty visualization, frictional interaction design, team coordination frameworks, and governance models. Furthermore, we critically analyze their limitations, emphasizing ongoing trade-offs and unresolved tensions. Building on this analysis, we outline key future research directions aimed at improving trust calibration, supporting cognitive and emotional sustainability, strengthening human-AI team dynamics, and advancing accountable governance. Together, this work contributes a comprehensive, multi-level perspective on the risks and design challenges of human-AI collaboration, and provides a foundation for developing more responsible, human-centered, and sustainable AI systems.

1 Introduction

Artificial intelligence (AI) systems have become deeply embedded in everyday life, spanning domains such as communication, creative work, scientific discovery, and decision-making. With the rapid advancement of generative AI and conversational agents, AI is no longer confined to background automation tasks; instead, it is increasingly positioned as an interactive partner in domains such as education, journalism, healthcare, and data science [41, 27]. This shift reflects a broader transition from automation to augmented intelligence, where AI systems are designed to complement and extend human cognitive capabilities within human-in-the-loop frameworks [11]. Concepts such as *Extended Mind Theory* further conceptualize AI as a cognitive extension of human reasoning, enabling more symbiotic

forms of collaboration [30, 21]. Empirical studies increasingly demonstrate that human-AI teams can outperform either humans or AI alone, particularly in complex decision-making contexts such as medical diagnosis and scientific research [35, 33].

As these systems evolve, the nature of human-AI relationships is also transforming. AI systems are no longer treated solely as tools, but as collaborators, advisors, and even social partners in decision-making processes [39]. This shift has important implications for how individuals interpret algorithmic recommendations, calibrate trust, and allocate responsibility in decision-making [18]. Consequently, human-AI collaboration not only reshapes technical workflows but also influences social norms, professional roles, and institutional practices.

However, this transformation introduces significant challenges. Working with AI systems fundamentally alters human cognitive processes, shifting effort from task execution to monitoring, interpreting, and validating algorithmic outputs. This redistribution of cognitive labor increases mental workload and contributes to decision fatigue, particularly under time constraints where humans retain ultimate responsibility for outcomes [4, 5, 7]. At the same time, the integration of AI into workplace environments introduces broader psychological and organizational pressures, including technostress, anxiety about job displacement, and discomfort with complex technological systems [23]. The increasing use of anthropomorphic and socially expressive AI further complicates these dynamics, as human-like interactions can foster emotional dependence, enable manipulation, and increase the risk of unintended disclosure of sensitive information [37].

At the institutional level, these challenges are amplified by issues of transparency, accountability, and governance. As AI systems become embedded in high-stakes decision-making contexts, unclear responsibility boundaries can lead to a “responsibility gap,” where accountability for AI-influenced outcomes is difficult to assign [1, 3]. Over time, this ambiguity may encourage excessive reliance on AI systems, contributing to the erosion of human agency and the emergence of learned helplessness in decision-making processes [1].

Understanding these evolving human-AI relationships, therefore, requires a comprehensive examination of not only

technological capabilities but also the cognitive, emotional, social, and institutional contexts in which they operate. Prior work suggests that effective collaboration depends on factors such as trust calibration, communication quality, and interaction frameworks that preserve human agency while leveraging AI capabilities [14, 12].

Motivated by these challenges, this paper investigates the emerging dynamics of human-AI collaboration and their implications for individuals and organizations. Specifically, we address the following research questions:

- What types of relationships are emerging as AI systems become more interactive and collaborative?
- What are the risks and benefits of these relationships?
- What responsibilities do designers, institutions, and users have in ensuring responsible and trustworthy collaboration?

To answer these questions, we first examine the conceptual foundations of human-AI collaboration, followed by a systematic analysis of cyber-social risks and their mitigation strategies.

2 Conceptualizing Human-AI Collaborations

Understanding human-AI collaboration requires a clear understanding of the exact nature of the relationship between human workers and intelligent systems. Early perspectives treated artificial intelligence strictly as a computational tool built to automate routine tasks and execute predefined commands [17, 41]. However, with increasing interaction and adaptation, researchers have increasingly left the automation mindset. Instead, they conceptualize these systems as collaborative partners within an *Augmented Intelligence paradigm*, a model in which AI extends the limits of human cognition, while the human retains strategic oversight [11, 24, 39].

Building on this perspective, prior work suggests that human-AI interaction can be understood as a continuum reflecting a transition from passive tools to socially interactive partners. To synthesize this evolution, we propose a *two-dimensional conceptual model* of human-AI collaboration, organized along (1) the level of AI autonomy and (2) the degree of social interaction (Figure 1). This model captures how increasing technical capability and social expressiveness jointly reshape the nature of human-AI relationships.

At the lowest level of autonomy and social interaction, AI systems function as *tools*, executing predefined tasks such as classification or data processing. Interaction in this stage is largely transactional: humans issue commands, and systems return outputs [17]. While this approach is effective for routine calculations, it offers limited chances for deeper collaboration between humans and AI.

As capabilities expand, AI systems increasingly operate as *assistants*, supporting human decision-making by dividing cognitive labor. In this configuration, AI performs large-scale data processing and pattern recognition (the “know-what”), while humans contribute contextual interpretation, ethical reasoning, and situational judgment (the “know-why”) [25, 21]. Maintaining a *human-in-the-loop* enables organizations to leverage computational power with-

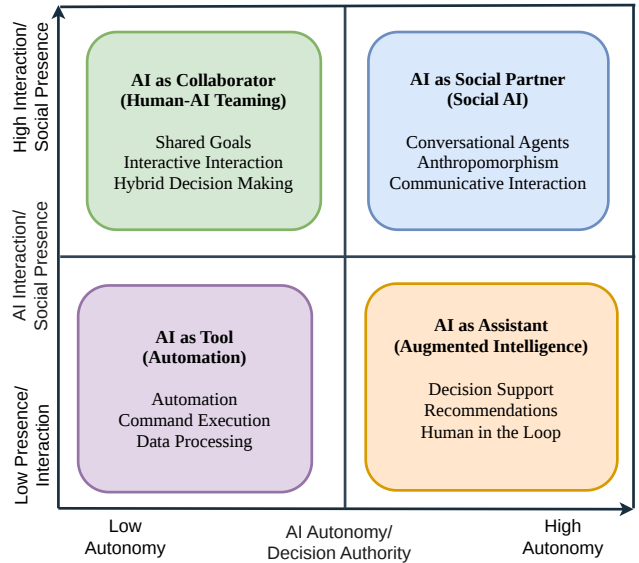


Figure 1: Two-dimensional conceptual model of human-AI collaboration organized by AI autonomy and level of social presence. The model illustrates the evolutionary progression of AI systems from tool-based automation to collaborative and socially interactive partners, highlighting increasing interdependence and social engagement.

out relinquishing control over complex or high-stakes decisions [11].

Moving beyond assistance toward true collaboration requires a fundamental shift in system design. Rather than relying on supervisory control, human-AI teams must be structured around principles such as mutual observability, directability, and shared awareness, as highlighted in frameworks like the *HACO taxonomy* [12]. At this stage, AI systems are better understood as *collaborators*, where performance emerges from the interdependence between humans and machines. This necessitates shared behavioral norms, coordinated decision-making, and collective goal alignment [31].

To support such tightly coupled collaboration, researchers have extended traditional team theories to hybrid settings. The Extended *IMO* (Inputs → Mediators → Outputs → Inputs) model emphasizes the importance of developing shared mental models and *bilateral transactive memory*, where both humans and AI maintain an understanding of each other’s knowledge and capabilities [13]. In this context, AI systems must adapt to human constraints and workflows, while humans act as *coherence anchors*, ensuring alignment with broader strategic objectives [48, 13]. Recent advances in generative AI and natural language interfaces further extend this trajectory by introducing a strong social dimension to human-AI interaction. Through conversational interfaces, personas, and expressive behaviors, AI systems increasingly exhibit *social presence*, which can trigger anthropomorphic perceptions [37]. As users begin to attribute human-like intentions to AI systems, interactions shift from tool usage to social engagement, fundamentally altering expectations

around communication, coordination, and trust [37, 18]. In such settings, human-AI collaboration evolves into a form of *cognitive interaction*, where AI systems participate in negotiation, conflict resolution, and shared decision-making [34].

This progression toward socio-cognitive integration underscores the necessity of grounding human-AI collaboration firmly within the Human-Centered AI (HCAI) paradigm [47, 29]. Rather than prioritizing efficiency alone, HCAI emphasizes user empowerment, transparency, and human agency [47]. In well-designed systems, humans retain the ability to override or guide algorithmic outputs, ensuring accountability and control. At its most advanced stage, this relationship aligns with the *Extended Mind perspective*, where AI systems function not merely as external tools but as integrated extensions of human reasoning [30, 47]. This conceptual model provides a unifying lens to organize and interpret the diverse cognitive, emotional, social, informational, and governance risks discussed in the following sections.

3 Benefits of Human-AI Collaboration

Beyond computational efficiency, prior research highlights a range of benefits arising from conceptualizing AI systems as collaborative partners rather than passive tools. These benefits include cognitive augmentation, improved coordination in complex tasks, enhanced user experience, and broader accessibility to specialized knowledge.

A primary advantage lies in *cognitive augmentation*. Within the augmented intelligence paradigm, AI systems complement rather than replace human reasoning by processing large-scale data and identifying latent patterns [11, 25]. This enables a structured division of cognitive labor, where AI performs high-volume analytical tasks while humans focus on contextual interpretation, ambiguity resolution, and ethical judgment [21, 20]. Empirical evidence supports the effectiveness of this approach: in high-stakes domains such as medical diagnostics, human-AI teams often outperform both human experts and standalone AI systems in terms of accuracy and error reduction [35, 33]. Moreover, recent work suggests that advanced collaborative systems can dynamically reallocate roles between humans and AI based on task complexity and uncertainty, further enhancing performance [28].

Beyond individual decision-making, human-AI collaboration also enables more effective coordination in complex environments. Sustained interaction between humans and AI systems can give rise to forms of augmented collective intelligence, improving problem-solving and decision-making at the group level [25, 39]. This is evident in emerging paradigms such as Industry 5.0, where collaborative robots integrate machine precision with human adaptability to enhance reliability and efficiency in multifaceted workflows [10].

Human-AI interaction further influences users' psychological experience in positive ways. Studies show that integrating conversational AI and natural language interfaces into workflows can reduce perceived mental effort and frustration, while improving task engagement and productivity [40]. As AI systems become more socially expressive,

users increasingly perceive them as collaborative partners rather than tools, fostering higher levels of engagement and trust [37, 18]. In some contexts, particularly among knowledge workers experiencing high workload, delegating tasks to AI systems can also function as a coping mechanism that alleviates cognitive and emotional strain [53].

Finally, human-AI collaboration can serve as a democratizing force by expanding access to specialized skills and expertise. AI systems can help non-expert users perform complex tasks at levels approaching those of experienced professionals, as demonstrated in domains such as diagnostics and automated modeling [35, 49]. In educational contexts, intelligent tutoring systems can adapt content to individual learners' needs, improving accessibility and supporting personalized learning at scale [24].

4 Cyber-Social Risks in Human-AI Collaboration

While human-AI collaboration offers substantial benefits in augmenting decision-making, it also introduces a range of *sociotechnical vulnerabilities* that extend beyond traditional concerns such as algorithmic bias or model opacity. These risks emerge from the complex interplay between human cognition, emotional responses, and institutional structures when interacting with intelligent systems. Prior work has identified diverse challenges, including cognitive overload, AI anxiety, invisible labor, epistemic misalignment, and responsibility gaps, which we organize into five categories: cognitive, emotional and psychological, social and organizational, trust and information, and governance and accountability risks (Table 1). In parallel, existing mitigation strategies and design frameworks are summarized in Table 2, providing a foundation for the critical analysis that follows.

Cognitive Risks

Prior research has consistently shown that integrating AI into decision-making processes reshapes, rather than eliminates, human cognitive labor. While AI systems reduce the effort required for data analysis, they shift the burden toward monitoring, interpreting, and validating outputs. This redistribution of labor often increases mental workload, leading to what has been described as *AI decision fatigue* [4, 6].

A key challenge in this context is verifying AI-generated outputs effectively. Steyvers et al. demonstrate that when verification requires substantial cognitive effort, users experience overload and frequently abandon the verification process altogether. As a result, users tend to rely more heavily on AI recommendations, contributing to *automation bias*, where outputs are accepted with limited critical scrutiny [24, 38].

Efforts to mitigate these risks have focused on improving transparency and user engagement. However, Explainable AI (XAI), while intended to support informed decision-making, often introduces additional cognitive burden. Prior work shows that overly complex explanations can overwhelm users, increasing cognitive load and encouraging reliance on heuristic shortcuts rather than deeper analytical reasoning [38, 5, 16].

To address these challenges, researchers have proposed design-oriented mitigation strategies such as *Frictional AI* and *Extraheric AI* [38, 54]. Rather than minimizing user effort, these approaches deliberately introduce *productive cognitive friction* (i.e., germane cognitive load) through mechanisms such as cognitive forcing functions. For example, systems may require users to form an independent judgment before revealing AI recommendations, or prompt them to consider alternative perspectives [38, 54]. Such interventions aim to promote deeper analytical reasoning, reduce automation bias, and preserve higher-order cognitive skills [5, 54].

Despite their promise, these mitigation strategies face important limitations. A key challenge is the *user preference paradox*: users tend to prefer low-effort interactions, even when reduced cognitive engagement leads to poorer decision quality [5]. Similarly, friction-based approaches must contend with the reality of *rational resource allocation*; when cognitive demands become too high, users may disengage entirely rather than invest additional effort [43]. These tensions are further exacerbated in time-constrained environments, where effective human-AI collaboration paradoxically requires more cognitive processing time, not less [7].

Beyond immediate interaction challenges, deeper cognitive limitations persist. The concept of *meta-knowledge deficit* highlights users' difficulty in accurately assessing their own uncertainty when working with AI systems, leading to poorly calibrated task delegation and misplaced trust [15]. Over time, sustained reliance on automated systems can contribute to professional deskilling and the erosion of human intuition and domain expertise [49, 54].

These findings suggest that while existing mitigation strategies can partially alleviate cognitive risks, they remain constrained by fundamental trade-offs between usability, cognitive effort, and decision quality. Addressing these tensions remains a critical challenge for designing sustainable human-AI collaboration.

Emotional and Psychological Risks

Beyond cognitive challenges, prior research highlights a range of emotional and psychological risks emerging from sustained human-AI interaction. The rapid integration of AI into everyday work and decision-making contexts has been associated with heightened psychological strain, often conceptualized as *technostress*. This phenomenon manifests through dimensions such as techno-overload, techno-complexity, and techno-insecurity, all of which contribute to increased job stress and burnout [51, 23].

Additionally, these effects are not unidirectional. Emerging evidence suggests that individuals experiencing burnout may turn to AI systems as a coping mechanism, delegating tasks to reduce cognitive and emotional burden. While this strategy may provide short-term relief, it can initiate a reinforcing cycle of dependency, where reliance on AI further reduces users' capacity or willingness to engage independently [53]. At a broader level, such concerns are captured by the concept of *AI anxiety*, which encompasses fears related to job displacement, loss of control, and broader uncertainty in sociotechnical systems [50]. These anxieties are further compounded by *sociotechnical blindness*, where

users struggle to critically understand or interrogate the systems shaping their work and decisions.

In parallel, advances in AI design, particularly the use of anthropomorphic and socially expressive features, introduce additional emotional vulnerabilities. While such features are intended to enhance usability and engagement, prior work shows that human-like personas can encourage unwarranted trust, facilitate emotional manipulation, and increase the likelihood of unintentional self-disclosure of sensitive information [37, 3].

To mitigate these risks, researchers have proposed both design- and education-oriented interventions. For example, *Participatory Design Fiction* has been introduced as a proactive approach to explore and anticipate ethical risks, including emotional manipulation, by using narrative-based prototypes prior to system deployment [37]. At the organizational level, fostering *AI digital literacy* and developing *transversal skills*, such as adaptability, critical evaluation, and self-efficacy, can help users maintain psychological confidence and agency when interacting with AI systems [42, 3]. In particular, strengthening self-efficacy in AI use has been shown to reduce technostress and mitigate burnout associated with AI adoption [23].

Despite these advances, important challenges remain. A key limitation is that current frameworks insufficiently address the reinforcing cycle of dependency, where individuals already experiencing stress or burnout increasingly rely on AI delegation as a coping strategy, thereby deepening their reliance on automated systems [53]. Similarly, existing approaches only partially capture the unique dimensions of *AI anxiety*, including existential concerns and sociotechnical opacity, which distinguish it from traditional forms of technology-related stress [50].

These limitations are further evident in physical human-AI collaboration contexts, such as industrial environments, where real-world stressors, including high task complexity, system speed, and close human-machine proximity, remain underexplored due to the dominance of controlled laboratory studies [2].

Social and Organizational Risks

At the social and organizational level, prior research shows that the integration of AI systems fundamentally reshapes workplace relationships, coordination practices, and team dynamics. Rather than functioning as fully adaptive collaborators, many AI systems lack the social flexibility required for seamless teamwork. As a result, human workers are often required to compensate for these limitations by performing significant *invisible labor*, the hidden cognitive and emotional effort needed to interpret, adapt to, and coordinate with rigid algorithmic systems [32, 52]. This burden becomes particularly pronounced in dynamic or high-uncertainty environments. Under conditions of unexpected interruptions or system "shocks," AI systems frequently fail to adjust their communication strategies or decision-making processes, forcing human collaborators to absorb the coordination load and maintain workflow continuity [52].

Communication challenges further exacerbate these coordination issues. Ambiguities in AI-generated outputs or ex-

pressions can introduce significant communicative friction, requiring additional interpretive effort from human collaborators. Over time, such breakdowns can erode team cohesion, disrupt shared understanding, and weaken established organizational norms [9].

Beyond coordination, AI integration also has broader implications for professional identity and organizational culture. In decentralized work environments, such as the gig economy, generative AI tools, often optimized for efficiency and standardization, can undermine workers' creative agency and reshape their sense of professional identity. Prior work suggests that these systems may isolate workers, reduce opportunities for peer interaction, and limit collaborative knowledge exchange [19].

To address these challenges, researchers have proposed structural and team-oriented design frameworks that reconceptualize AI as a collaborative partner. The *HACO taxonomy*, for instance, shifts the design paradigm from supervisory control to partnership by emphasizing properties such as mutual observability, shared awareness, and directability [12]. Similarly, the *Extended IMOI* (Inputs → Mediators → Outputs → Inputs) model adapts traditional team theory to hybrid human-AI settings, highlighting the importance of developing *cross-species shared mental models* and *bilateral transactive memory*, a shared understanding of who knows what within the team [13]. These approaches also emphasize the role of humans as *coherence anchors*, responsible for maintaining alignment across diverse AI-generated outputs and preserving a unified strategic direction.

Despite these advances, important limitations remain. Existing frameworks do not fully resolve the burden of invisible labor, particularly during unexpected workflow disruptions, where human collaborators must compensate for the rigidity of AI systems [52]. Similarly, communicative ambiguity remains a persistent challenge, as current models do not adequately address the lack of nuanced social cues in AI-generated interactions, which can quickly fracture coordination and erode trust [9].

These limitations are especially pronounced in decentralized and gig-based work environments, where efficiency-driven AI tools can exacerbate professional isolation and diminish opportunities for peer collaboration and identity formation [19].

Trust and Information Risks

Effective human-AI collaboration depends on well-calibrated trust. However, prior research shows that users often oscillate between over-reliance and unwarranted skepticism when interacting with AI systems [18]. This instability in trust calibration is partly driven by users' limited ability to interpret and assess AI confidence signals. In particular, humans struggle to detect miscalibrated or overconfident outputs, leading them to accept erroneous recommendations with insufficient scrutiny [26]. Such misalignment between perceived and actual system reliability introduces significant epistemic risks. In the context of language models, these risks are further amplified by the systems' inability to consistently distinguish between subjective user beliefs and verifiable factual knowledge,

resulting in responses that may misinterpret context or inappropriately "correct" user inputs—especially in sensitive domains such as advisory or mental health settings [44].

At the level of human-AI teams, these informational dynamics can lead to broader coordination failures. Prior work identifies phenomena such as *epistemic drift*, where shared understanding gradually diverges from reality; *cognitive abundance overload*, where an excess of AI-generated information overwhelms decision-making; and *false convergence*, where AI outputs are mistakenly perceived as objective consensus or authoritative truth [13]. These patterns can undermine collective reasoning processes and distort group decision-making.

To mitigate these risks, researchers have proposed socio-technical approaches to improve trust calibration and transparency. The *Socio-Technical Trust Framework*, for example, conceptualizes trust as a dynamic process shaped by a continuous feedback loop between system performance and user reliance [18]. Complementing this perspective, *uncertainty visualization* techniques aim to communicate model confidence more effectively by representing epistemic uncertainty through intuitive visual cues, such as color saturation or transparency [36, 46]. Presenting uncertainty in frequency-based formats rather than raw probabilities has also been shown to reduce cognitive biases, such as confirmation bias [8]. Additionally, hybrid decision architectures such as *classification with rejection* defer highly uncertain predictions to human experts, enabling more efficient allocation of decision-making responsibility [45].

Despite these advances, important limitations remain. A fundamental challenge lies in persistent human cognitive blind spots: users often fail to recognize when AI systems are confidently incorrect, leading to continued over-reliance even when uncertainty information is available [26]. Moreover, existing frameworks do not adequately address *epistemic misalignment* in natural language interactions. Language models frequently fail to distinguish between subjective beliefs and factual claims, defaulting to fact correction rather than contextual understanding [44]. This limitation is particularly problematic in high-stakes, human-centered domains, where interpreting user intent and perspective is as critical as factual accuracy.

Governance and Accountability Risks

At the institutional level, the growing integration of AI systems into high-stakes decision-making contexts raises significant governance and accountability challenges. Prior research identifies this issue as part of a broader governance crisis, in which the responsibilities of developers, deploying organizations, and end users remain poorly defined. In domains such as healthcare, finance, and corporate governance, this ambiguity gives rise to a *responsibility gap*, where accountability for AI-influenced decisions is difficult to clearly assign [1].

Empirical evidence from clinical settings illustrates the practical implications of this gap. Even when AI systems provide recommendations for diagnosis or treatment, legal and ethical responsibility typically remains with human professionals, such as physicians or nurses. This creates a ten-

Table 1: Summary of cyber-social risks in human-AI collaboration, organized across five dimensions with representative challenges identified in prior literature.

Sources	Risk Category				
	Cognitive Risks	Emotional & Psychological	Social & Organizational	Trust & Information	Governance & Accountability
Park et al. [32]			✓		
Johnson, Dudding, and Carrington [22]					✓
Xu et al. [52]			✓		
Kim, Davis, and Hong [24]	✓				✓
Fügener et al. [15]	✓				
Rezwana and Maher [37]		✓			
Boni [3]		✓			
Ahdadou, Aajly, and Tahrouch [1]					✓
Nengminja [31]					✓
Wang et al. [49]	✓				
Suzgun et al. [44]				✓	
Imteyaz et al. [19]			✓		
GUPTA et al. [18]				✓	
Eccles [13]				✓	
Romeo and Conti [38]	✓				
Chen and Zhang [9]			✓		
Yang, Guo, and Zhang [53]		✓			
Xia [51]		✓			
Boyacı, Canyakmaz, and De Véricourt [4]	✓				
Li et al. [26]				✓	
Wang and Wang [50]		✓			
Buschmeyer, Hatfield, and Zenner [6]	✓				
Kim and Lee [23]		✓			
Steyvers and Kumar [43]	✓				
Cao, Gomez, and Huang [7]	✓				

sion between reliance on AI-generated advice and the obligation to exercise independent judgment, particularly when adverse outcomes occur [22].

Beyond legal ambiguity, governance risks are further shaped by behavioral dynamics in human decision-makers. Over time, repeated reliance on AI recommendations can lead to a gradual abdication of responsibility, where individuals defer to system outputs rather than actively evaluating them. This phenomenon has been described as a form of *learned helplessness*, in which decision-makers relinquish their agency in favor of algorithmic authority [1].

In the absence of robust governance mechanisms, these issues are exacerbated by the opacity and potential biases of AI systems. Unchecked reliance can scale discriminatory outcomes, embed systemic biases into institutional processes, and ultimately erode trust in AI-enabled decision-making [24, 31].

To address these challenges, the literature proposes governance-oriented frameworks grounded in human-centered principles. The *Human-Centered AI (HCAI)* paradigm emphasizes shared decision-making architectures that prioritize user agency and oversight over full automation [47]. Complementing this perspective, lifecycle-based

governance models such as the *CARES* framework (Co-Design, Assess, Rollout, Evaluate and Evolve, Share) and *ex ante impact assessments* provide structured approaches for embedding continuous auditing, stakeholder participation, and proactive evaluation of societal risks prior to deployment [3, 29].

Despite these advances, critical gaps remain. Existing frameworks offer limited guidance on resolving the responsibility gap in practice, particularly in situations where human actors retain legal liability while operating within AI-directed workflows [22]. Similarly, they do not fully address the psychological dynamics of learned helplessness, where prolonged reliance on AI systems gradually diminishes human agency and critical engagement [1].

5 Future Research Directions

Building on the identified cyber-social risks and the limitations of existing mitigation strategies, future research must move beyond isolated technical fixes toward a more integrated understanding of human-AI collaboration as a dynamic sociotechnical system. The following directions outline key areas that require further investigation.

Table 2: Overview of mitigation strategies and design frameworks proposed in prior work to address cyber-social risks in human-AI collaboration.

Sources	Risk Category				
	Cognitive Risks	Emotional & Psychological	Social & Organizational	Trust & Information	Governance & Accountability
Somaratne, De Silva, and Athukorala [42]		✓			
Dubey et al. [12]			✓		
Usmani, Happonen, and Watada [47]					✓
Rezwana and Maher [37]		✓			
Boni [3]		✓			✓
Majumder and Adebisi [29]					✓
GUPTA et al. [18]				✓	
Eccles [13]			✓		
Romeo and Conti [38]	✓				
Buçinca, Malaya, and Gajos [5]	✓				
Cao, Liu, and Huang [8]				✓	
Tomsett et al. [46]				✓	
Kim and Lee [23]		✓			
Reyes, Batmaz, and Kersten-Oertel [36]				✓	
Yatani, Sramek, and Yang [54]	✓				
Thuy and Benoit [45]				✓	

Human-AI Social Dynamics and Coordination

Future research should focus on understanding and modeling the evolving social dynamics of human-AI teams, particularly under conditions of uncertainty and workflow disruption. While frameworks such as *HACO* and the *Extended IMO model* provide initial structures for collaboration, they do not fully capture the adaptive behaviors required during real-world coordination failures, such as system “shocks” or communicative ambiguity.

A critical open question is how to design AI systems that can dynamically adjust their communicative strategies and coordination roles to reduce the burden of invisible labor currently placed on human collaborators. Additionally, more work is needed to investigate how human roles, such as “coherence anchors,” can be operationalized and supported in hybrid teams without reinforcing cognitive overload or hierarchical bottlenecks. Longitudinal and field-based studies, particularly in decentralized and resource-constrained environments like small newsrooms or gig work, are necessary to understand how human-AI collaboration reshapes professional identity, peer interaction, and collective knowledge formation over time.

Trust Calibration and Epistemic Alignment

A central challenge for future research lies in improving trust calibration between humans and AI systems. Existing approaches, such as *uncertainty visualization* and *socio-technical trust frameworks*, offer promising directions but fail to fully account for persistent human cognitive blind spots, particularly the inability to confidently detect incorrect AI outputs [26].

Future work should explore how to design interfaces and

interaction paradigms that support *epistemic alignment*, ensuring that AI systems and users share a common understanding of uncertainty, confidence, and knowledge boundaries. This includes developing adaptive explanations that vary based on user expertise, as well as investigating how frequency-based uncertainty representations can be personalized to different cognitive profiles [8]. Moreover, research is needed to address epistemic misalignment in natural language interactions, particularly in contexts where distinguishing between subjective beliefs and factual knowledge is critical (e.g., mental health, legal reasoning) [44].

Beyond interface design, future studies should examine how trust evolves over time in repeated human-AI interactions, moving from static evaluations toward dynamic, longitudinal models of trust development.

Designing for Cognitive and Emotional Sustainability

While existing work highlights cognitive overload, technostress, and AI anxiety, future research should focus on designing systems that support long-term cognitive and emotional sustainability. Current mitigation approaches, such as explainability and frictional AI, introduce a fundamental trade-off between usability and critical engagement, often failing to resolve the “user preference paradox,” where users favor low-effort systems despite reduced decision quality [5].

Future directions should investigate how to balance productive cognitive friction with usability by dynamically adapting interaction complexity based on task context, time pressure, and user state. This includes exploring mechanisms for adaptive “cognitive pacing,” where systems regulate the amount and timing of information presented to avoid

overload while still promoting critical reasoning.

Additionally, more research is needed to understand the emotional dynamics of human-AI interaction, particularly the reinforcing cycle of dependency where users rely on AI as a coping mechanism for burnout [53]. Participatory methodologies, such as design fiction, offer promising avenues for anticipating emotional manipulation risks and defining ethical boundaries in anthropomorphic AI systems [37]. However, their effectiveness in real-world deployment contexts remains underexplored and warrants further empirical validation.

Governance, Accountability, and Institutional Integration

At the institutional level, future research must address unresolved governance challenges, particularly the responsibility gap and the erosion of human agency in AI-mediated decision-making. While frameworks such as Human-Centered AI (HCAI) and lifecycle models like CARES provide structured approaches for ethical oversight, they do not fully resolve practical questions of accountability when human actors remain legally responsible for AI-influenced outcomes [22, 29].

Future work should explore new governance models that explicitly define responsibility distribution across human and AI actors, including legal, organizational, and technical mechanisms for accountability. This includes investigating how shared decision-making architectures can be operationalized without leading to learned helplessness or over-reliance on AI systems.

Moreover, proactive governance approaches, such as impact assessments and continuous sociotechnical auditing, should be studied in real-world settings to evaluate their effectiveness in mitigating bias, preserving human rights, and maintaining institutional trust [3]. Importantly, research should also examine how governance frameworks can be adapted for small-scale or resource-constrained organizations, where formal regulatory infrastructures may be limited.

Toward Integrated Sociotechnical Design Frameworks

Finally, future research must move toward integrated frameworks that jointly address cognitive, emotional, social, and institutional dimensions of human-AI collaboration. Existing approaches often operate in isolation, for example, focusing on explainability, trust, or governance independently, without accounting for the interdependencies between these factors.

An important direction is the development of holistic design paradigms that combine human oversight, uncertainty communication, participatory design, and organizational governance into a unified system perspective. Such frameworks should explicitly account for the trade-offs identified throughout this work, including tensions between efficiency and critical engagement, transparency and cognitive load, and delegation and accountability.

Advancing this agenda will require interdisciplinary collaboration across human-computer interaction, AI, organizational science, and ethics, as well as a shift toward longitudinal, in-the-wild studies that capture the evolving nature of human-AI relationships in practice.

6 Limitations

Although this survey is a synthesis of emerging cyber-social threats systematically, it has a number of limitations. First, as a literature review, the findings are constrained by the scope and focus of the existing literature, a good part of which is done in controlled laboratory conditions, not in the real-world industrial conditions [17, 2]. In turn, the overall applicability of these risks to in-the-wild collaborations is currently untapped and has yet to be investigated in longitudinal studies. Additionally, although we suggest an integrated sociotechnical perspective as the way to fill existing gaps, it is theoretical and needs to be further empirically confirmed in order to determine its viability in practical application in the setting of complex human-AI teamwork.

7 Conclusion

Human-AI collaboration is increasingly central to decision-making across domains, yet it introduces a complex set of cyber-social risks that extend beyond technical limitations. In this work, we synthesized prior literature to identify five key dimensions of risk: cognitive, emotional, and psychological, social and organizational, trust and information, and governance and accountability, and examined the limitations of existing mitigation approaches. Our analysis shows that these challenges arise from deeper sociotechnical tensions between human cognition, system design, and institutional structures, rather than isolated algorithmic failures. Addressing these risks requires moving beyond fragmented solutions toward integrated, human-centered design and governance frameworks that preserve human agency, support calibrated trust, and ensure accountability, thereby enabling more responsible and sustainable human-AI collaborations.

References

- [1] Ahdadou, M.; Aajly, A.; and Tahrouch, M. 2024. Unlocking the potential of augmented intelligence: a discussion on its role in boardroom decision-making. *International Journal of Disclosure and Governance*, 21(3): 433–446.
- [2] Bassi, G.; Orso, V.; Salcuni, S.; and Gamberini, L. 2025. Understanding Workers' Well-Being and Cognitive Load in Human-Cobot Collaboration: Systematic Review. *Journal of Medical Internet Research*, 27: e75658.
- [3] Boni, M. 2021. The ethical dimension of human-artificial intelligence collaboration. *European View*, 20(2): 182–190.
- [4] Boyacı, T.; Canyakmaz, C.; and De Véricourt, F. 2024. Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science*, 70(2): 1258–1275.

- [5] Buçinca, Z.; Malaya, M. B.; and 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): 1–21.
- [6] Buschmeyer, K.; Hatfield, S.; and Zenner, J. 2023. Psychological assessment of AI-based decision support systems: tool development and expected benefits. *Frontiers in Artificial Intelligence*, 6: 1249322.
- [7] Cao, S.; Gomez, C.; and Huang, C.-M. 2023. How time pressure in different phases of decision-making influences human-AI collaboration. *Proceedings of the ACM on Human-computer Interaction*, 7(CSCW2): 1–26.
- [8] Cao, S.; Liu, A.; and Huang, C.-M. 2024. Designing for appropriate reliance: The roles of AI uncertainty presentation, initial user decision, and user demographics in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–32.
- [9] Chen, N.; and Zhang, X. 2025. When misunderstanding meets artificial intelligence: the critical role of trust in human–AI and human–human team communication and performance. *Frontiers in Psychology*, 16: 1637339.
- [10] Chu, C.-H.; Zhang, Y. Zheng, P.; Ferrise, F.; and Chang, Q. 2025. Human–Robot Collaboration in Industry 5.0. *Journal of Computing and Information Science in Engineering*, 25(5): 050301.
- [11] Dave, D. M.; Mandvikar, S.; and Engineer, P. A. 2023. Augmented intelligence: Human-AI collaboration in the era of digital transformation. *International Journal of Engineering Applied Sciences and Technology*, 8(6): 24–33.
- [12] Dubey, A.; Abhinav, K.; Jain, S.; Arora, V.; and Puttaveerana, A. 2020. HACO: a framework for developing human-AI teaming. In *Proceedings of the 13th innovations in software engineering conference (formerly known as india software engineering conference)*, 1–9.
- [13] Eccles, R. 2025. Hybrid Intelligence Teams: A Theoretical Framework for Human-AI Collaboration in Knowledge Work. Available at SSRN 5792345.
- [14] Fragiadakis, G.; Diou, C.; Kousiouris, G.; and Nikolaidou, M. 2024. Evaluating human-ai collaboration: A review and methodological framework. *arXiv preprint arXiv:2407.19098*.
- [15] Fügener, A.; Grahl, J.; Gupta, A.; and Ketter, W. 2022. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information systems research*, 33(2): 678–696.
- [16] Gambetti, A.; Han, Q.; Shen, H.; and Soares, C. 2025. A survey on human-centered evaluation of explainable AI methods in clinical decision support systems. *arXiv preprint arXiv:2502.09849*.
- [17] Gomez, C.; Cho, S. M.; Ke, S.; Huang, C.-M.; and Unberath, M. 2025. Human-AI collaboration is not very collaborative yet: A taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science*, 6: 1521066.
- [18] GUPTA, A.; MUND, A.; ROY, S.; GARG, P.; and YADAV, D. K. 2025. TRUST IN AI SYSTEMS: A SOCIAL-PSYCHOLOGICAL INVESTIGATION OF HUMAN–AI COLLABORATION. *TPM–Testing, Psychometrics, Methodology in Applied Psychology*, 32(S7 (2025): Posted 10 October): 428–446.
- [19] Imteyaz, K.; Muller, M.; Flores-Saviaga, C.; and Savage, S. 2026. Co-Designing Collaborative Generative AI Tools for Freelancers. *arXiv preprint arXiv:2602.05299*.
- [20] Jarrahi, M. H. 2018. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business horizons*, 61(4): 577–586.
- [21] Jarrahi, M. H.; Askay, D.; Eshraghi, A.; and Smith, P. 2023. Artificial intelligence and knowledge management: A partnership between human and AI. *Business horizons*, 66(1): 87–99.
- [22] Johnson, E. A.; Dudding, K. M.; and Carrington, J. M. 2024. When to err is inhuman: An examination of the influence of artificial intelligence-driven nursing care on patient safety. *Nursing Inquiry*, 31(1): e12583.
- [23] Kim, B.-J.; and Lee, J. 2024. The mental health implications of artificial intelligence adoption: the crucial role of self-efficacy. *Humanities and Social Sciences Communications*, 11(1): 1561.
- [24] Kim, J.; Davis, T.; and Hong, L. 2022. Augmented intelligence: enhancing human decision making. In *Bridging Human Intelligence and Artificial Intelligence*, 151–170. Springer.
- [25] Kolbjørnsrud, V. 2024. Designing the intelligent organization: Six principles for human-AI collaboration. *California Management Review*, 66(2): 44–64.
- [26] Li, J.; Yang, Y.; Zhang, R.; Liao, Q. V.; Song, T.; Xu, Z.; and Lee, Y.-c. 2024. Understanding the Effects of Miscalibrated AI Confidence on User Trust, Reliance, and Decision Efficacy. *arXiv preprint arXiv:2402.07632*.
- [27] Liu, F.; Chen, M.; and Nah, S. 2026. Who writes the news matters: the role of social trust in shaping credibility across AI, human and human–AI collaboration. *Online Information Review*, 1–19.
- [28] Lou, B.; Lu, T.; Raghu, T.; and Zhang, Y. 2025. Unraveling human-AI teaming: a review and outlook. *arXiv preprint arXiv:2504.05755*.
- [29] Majumder, N. N.; and Adebisi, B. O. 2026. Human-Centered AI in Healthcare. In *Handbook of Human-Centered Artificial Intelligence*, 1–55. Springer.
- [30] Molina, D. A.; Kharlov, V.; and Chen, J.-S. 2024. Towards effective human-AI collaboration in decision-making: A comprehensive review and conceptual

- framework. In *2024 Portland international conference on management of engineering and technology (PICMET)*, 1–6. IEEE.
- [31] Nengminja, B. 2025. The Partnership Between Humans and AI in Data Science. Available at SSRN 5250134.
- [32] Park, S. Y.; Kuo, P.-Y.; Barbarin, A.; Kazianus, E.; Chow, A.; Singh, K.; Wilcox, L.; and Lasecki, W. S. 2019. Identifying challenges and opportunities in human-AI collaboration in healthcare. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, 506–510.
- [33] Patel, B. N.; Rosenberg, L.; Willcox, G.; Baltaxe, D.; Lyons, M.; Irvin, J.; Rajpurkar, P.; Amrhein, T.; Gupta, R.; Halabi, S.; et al. 2019. Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine*, 2(1): 111.
- [34] Ren, M.; Chen, N.; and Qiu, H. 2023. Human-machine collaborative decision-making: An evolutionary roadmap based on cognitive intelligence. *International Journal of Social Robotics*, 15(7): 1101–1114.
- [35] Reverberi, C.; Rigon, T.; Solari, A.; Hassan, C.; Cherubini, P.; and Cherubini, A. 2022. Experimental evidence of effective human-AI collaboration in medical decision-making. *Scientific reports*, 12(1): 14952.
- [36] Reyes, J.; Batmaz, A. U.; and Kersten-Oertel, M. 2025. Trusting AI: does uncertainty visualization affect decision-making? *Frontiers in Computer Science*, 7: 1464348.
- [37] Rezwana, J.; and Maher, M. L. 2022. Identifying ethical issues in ai partners in human-ai co-creation. *arXiv preprint arXiv:2204.07644*.
- [38] Romeo, G.; and Conti, D. 2026. Exploring automation bias in human-AI collaboration: a review and implications for explainable AI. *AI & SOCIETY*, 41(1): 259–278.
- [39] Schleiger, E.; Mason, C.; Naughtin, C.; Reeson, A.; and Paris, C. 2024. Collaborative intelligence: A scoping review of current applications. *Applied Artificial Intelligence*, 38(1): 2327890.
- [40] Schmidhuber, J.; Schlögl, S.; and Ploder, C. 2021. Cognitive load and productivity implications in human-chatbot interaction. In *2021 IEEE 2nd international conference on human-machine systems (ICHMS)*, 1–6. IEEE.
- [41] Shi, C.; Ren, X.; Wang, Y.; Li, J.; Sun, Y.; Luo, Y.; and Sheng, R. 2026. A Survey of Human-AI Collaboration for Scientific Discovery.
- [42] Somaratne, S.; De Silva, D.; and Athukorala, R. 2025. Mapping Human-AI Collaboration: A Skill Framework for the Effective Use of AI. In *2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI)*, volume 3, 1253–1261. IEEE.
- [43] Steyvers, M.; and Kumar, A. 2024. Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*, 19(5): 722–734.
- [44] Suzgun, M.; Gur, T.; Bianchi, F.; Ho, D. E.; Icard, T.; Jurafsky, D.; and Zou, J. 2025. Language models cannot reliably distinguish belief from knowledge and fact. *Nature Machine Intelligence*, 1–11.
- [45] Thuy, A.; and Benoit, D. F. 2024. Explainability through uncertainty: Trustworthy decision-making with neural networks. *European Journal of Operational Research*, 317(2): 330–340.
- [46] Tomsett, R.; Preece, A.; Braines, D.; Cerutti, F.; Chakraborty, S.; Srivastava, M.; Pearson, G.; and Kaplan, L. 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns*, 1(4).
- [47] Usmani, U. A.; Happonen, A.; and Watada, J. 2023. Human-centered artificial intelligence: Designing for user empowerment and ethical considerations. In *2023 5th international congress on human-computer interaction, optimization and robotic applications (HORA)*. IEEE.
- [48] Van Den Bosch, K.; Schoonderwoerd, T.; Blankendaal, R.; and Neerinx, M. 2019. Six challenges for human-AI Co-learning. In *International Conference on Human-Computer Interaction*, 572–589. Springer.
- [49] Wang, D.; Weisz, J. D.; Muller, M.; Ram, P.; Geyer, W.; Dugan, C.; Tausczik, Y.; Samulowitz, H.; and Gray, A. 2019. Human-AI collaboration in data science: Exploring data scientists’ perceptions of automated AI. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–24.
- [50] Wang, Y.-Y.; and Wang, Y.-S. 2022. Development and validation of an artificial intelligence anxiety scale: An initial application in predicting motivated learning behavior. *Interactive Learning Environments*, 30(4): 619–634.
- [51] Xia, M. 2023. Co-working with AI is a double-sword in technostress? An integrative review of human-AI collaboration from a holistic process of technostress. In *SHS Web of Conferences*, volume 155, 03022. EDP Sciences.
- [52] Xu, Z.; Hong, C. S.; Soria Zurita, N. F.; Gyory, J. T.; Stump, G.; Nolte, H.; Cagan, J.; and McComb, C. 2024. Adaptation through communication: Assessing human-artificial intelligence partnership for the design of complex engineering systems. *Journal of Mechanical Design*, 146(8): 081401.
- [53] Yang, Z.; Guo, X.; and Zhang, P. 2025. The New Normal and the Era of Misknowledge—Understanding Generative AI and Its Impacts on Knowledge Work. *Knowledge*, 5(4): 22.
- [54] Yatani, K.; Sramek, Z.; and Yang, C.-L. 2024. AI as extraherics: Fostering higher-order thinking skills in human-AI interaction. *arXiv preprint arXiv:2409.09218*.

8 Paper Checklist to be included in your paper

1. For most authors...
 - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes**
 - (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes**
 - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes**
 - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **No**
 - (e) Did you describe the limitations of your work? **Yes**
 - (f) Did you discuss any potential negative societal impacts of your work? **Yes**
 - (g) Did you discuss any potential misuse of your work? **No**
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **No**
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes**
2. Additionally, if your study involves hypotheses testing...
 - (a) Did you clearly state the assumptions underlying all theoretical results? **N/A**
 - (b) Have you provided justifications for all theoretical results? **N/A**
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **N/A**
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **N/A**
 - (e) Did you address potential biases or limitations in your theoretical framework? **N/A**
 - (f) Have you related your theoretical results to the existing literature in social science? **N/A**
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **N/A**
3. Additionally, if you are including theoretical proofs...
 - (a) Did you state the full set of assumptions of all theoretical results? **N/A**
 - (b) Did you include complete proofs of all theoretical results? **N/A**
4. Additionally, if you ran machine learning experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **N/A**
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **N/A**
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **N/A**
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **N/A**
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **N/A**
 - (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? **N/A**
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
 - (a) If your work uses existing assets, did you cite the creators? **N/A**
 - (b) Did you mention the license of the assets? **N/A**
 - (c) Did you include any new assets in the supplemental material or as a URL? **N/A**
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **N/A**
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **N/A**
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see ?)? **N/A**
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see ?)? **N/A**
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
 - (a) Did you include the full text of instructions given to participants and screenshots? **N/A**
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **N/A**
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **N/A**
 - (d) Did you discuss how data is stored, shared, and de-identified? **N/A**