Who is an Expert? The Role of Human Expertise in Human-AI Synergy Studies*

Mickey V. Mancenido[1,*], Erin K. Chiou[2]

[1]School of Mathematical and Natural Sciences, Arizona State University, Glendale, AZ USA 85306
[2]The Polytechnic School, School of Computing and Augmented Intelligence, Arizona State University, Mesa, AZ, USA 85212

Abstract

Many human-AI (HAI) synergy studies rely on synthetic test environments (STEs) with recruited participants from the general population, yet participant expertise is rarely considered or operationalized as a design factor. This work provides a systematic review and thematic synthesis of empirical studies on AI decision support systems (AI-DSSs) with humans-in-the-loop. We find that participant expertise in HAI literature is inconsistently addressed, often limited to demographic reporting or acknowledged as a study limitation. This is despite growing evidence that key metrics, such as trust perceptions, trust behaviors, and performance, are contingent on participant expertise. In response, we propose a conceptual framework that (1) introduces a typology of participant archetypes synthesized from existing HAI research, and (2) maps these archetypes to HAI interaction dynamics across three common experimental phases in synergy studies: pre-interaction, in-situ, and post-task. This framework provides a conceptual foundation for embedding participant expertise as a central driver of design decisions in synthetic test environments to support context-aware interpretation of outcomes in HAI studies.

Keywords

sampling, statistics, testbeds, human-AI teaming, methodology, cognitive systems, crowd-sourcing

1. Background

The emerging paradigm promoting AIs as collaborative teammates rather than mere tools [1] has coalesced along with significant interest in human-AI (HAI) synergy metrics such as trust and trust outcomes [2], collaborative decision-making performance [3], and other indicators of successful team dynamics [4]. As a result of the growing interest in the effective design of HAI teams, investigators have resorted to *synthetic test environments* or STEs to manipulate and test design variables that are purported to impact human-AI synergy. STEs – also called simulated task environments, testbeds, microworlds, or sandboxes – are simplified, controlled scenarios that abstract key elements of a real-world task [5, 6, 7] so that researchers can investigate how people perceive, interact with, and respond to artificial intelligence agents in a safe, relevant, and experimental environment. STEs thus offer a middle ground between a tightly-controlled lab experiment and an unwieldy field study—they are intended to provide sufficient realism for participants to experience authentic interactions with AI systems under representative cognitive demands, while retaining the controllability and repeatability required in rigorous, scientific research. However, despite their methodological strengths, an important yet frequently overlooked factor influencing HAI synergy in STE-based studies is the expertise level of recruited participants.

HAI researchers have been increasingly using crowdsourcing platforms (e.g., Prolific, Amazon Mechanical Turk) as a recruitment and deployment tool for STEs. The practice of recruiting general population participants in online HAI research experiments likely increased during the COVID-19 pandemic but the methodological convenience, cost-efficiency, and scalability of deploying human subject experiments online point to their sustained growth [8]. As a result of the broadened access to a general population participant pool, empirical research in human–AI interaction has substantially

SYNERGY - Designing and Building Hybrid Human–AI Systems, HHAI 2025
*Corresponding author.
†These authors contributed equally.
mvmancenido@asu.edu (M. V. Mancenido); erin.chiou@asu.edu (E. K. Chiou)
https://labs.engineering.asu.edu/adapt/ (E. K. Chiou)
0000-0002-3000-8922 (M. V. Mancenido); 0000-0002-7201-8483 (E. K. Chiou)

accelerated, particularly in domains where recruitment of target expert populations are logistically challenging or resource-intensive.

Despite the convenience of recruiting general-population participants (e.g., crowdsourced workers) for HAI synergy experiments, reliance on this participant pool introduces potential threats to the validity and generalizability of study findings. A common issue is *task-participant misalignment* - where the difficulty of tasks is not appropriately matched to participants' expertise or cognitive capacities. This misalignment is particularly problematic in studies investigating high-stakes decision-making contexts, which typically require the specialized knowledge, training, and experience of domain experts [9, 10]. Researchers have addressed this issue by deliberately simplifying or selecting tasks that are closely aligned with the cognitive ability and domain familiarity of general population participants. For example, high-stakes medical decisions have been reduced to simplified analogs, such as choosing between nutrition options instead of clinical treatments [11], while complex professional classification tasks have been reduced from a comprehensive set of professions to a few familiar categories [**?**]. Researchers have also frequently chosen familiar image or text classification tasks, which crowdsourced workers can readily perform without requiring specialized knowledge or training [12].

However, decisions to simplify task environments in STEs are not without trade-offs. When tasks are overly simplistic, participants may under-rely on the AI's input [13], confounding lower measures of trust in the AI. Conversely, tasks that are excessively complex can cause participants to over-rely on AI recommendations, even when the AI is incorrect, confounding higher trust measures in the AI and masking collaborative breakdowns. For example, increasing task difficulty has been shown to result in higher reliance on AI regardless of the AI's accuracy [13] . These findings suggest that relative task difficulty in STEs can cause miscalibrated trust and subsequently inappropriate reliance outcomes, rather than miscalibrated trust being a result of poorly communicated purpose, process, or performance information [14]. The net result is a potential distortion in experimental findings in the search for "optimal" trust—what gets categorized as appropriate or inappropriate trust, or high versus low trust, might actually be an artifact of the task-participant pairing more so than meaningful measures of trust or reliance that are intended to help achieve desired performance goals.

A related criticism of simplified decision contexts in HAI experiments is their inherent lack of *ecological validity* i.e., scenarios often do not sufficiently resemble real-world risk contexts, particularly in high-stakes domains. This limitation contributes to potential overgeneralization of results. Glikson and Woolley [15] highlight a key limitation of empirical studies – "an overreliance on short-term, small-sample experimental studies" – and emphasize that trust in AI tends to develop quite differently in longer-term, high-stakes, real-world contexts. In other words, what researchers learn from a 30-minute task on platforms like Prolific.co might not translate to an expert physician who has worked for months, in multiple decision contexts, with an AI decision aid. Additionally, researchers have pointed out that findings from one domain often fail to translate effectively to another. This is especially true in high-stakes domains such as legal or healthcare, where experts tend to approach AI with different prior expectations, such as exhibiting distrust or caution [16], which contrasts significantly with the findings from simplified lab studies conducted with laypeople who, by and large, tend to show positive trust with new technologies [17].

These methodological criticisms of STEs, especially those involving crowd sourced participants, collectively point to an often overlooked factor in HAI synergy research: participant expertise. This work makes two major contributions. Firstly, we introduce an empirically grounded typology of participants in human-AI studies. This categorization clarifies ambiguity in the literature and supports more precise experimental stratification in human-subjects research. Secondly, we show that even in STEs, trust is not a fixed attitude, but an emergent, dynamic, and context-sensitive process impacted by task-participant alignment, task complexity, and AI framing. Trust must be interpreted relative to these contextual factors.

2. Synthetic Test Environments in HAI Synergy Research

STEs are task-centric platforms in HAI research that are designed to simulate real-world operational environments, with human participants performing tasks with assistance from an AI system. While they do not fully emulate the operational environment, researchers attempt to recreate key task elements at appropriate fidelity to invoke similar decision-making processes [7, 18, 19, 20, 21, 22]. These synthetic environments, often deployed in digital or web-based formats, facilitate the manipulation of variables of interest (e.g., AI reliability, task difficulty) in a safe and controlled setting.

Researchers demonstrate considerable flexibility and creativity in how they design STEs but common implementations include interactive virtual environments [23], factorial surveys [24], and game-based simulations [25]. *Virtual environments* are immersive, interactive computer-based simulations, designed with realistic user interfaces and task dynamics that closely mirror actual operational contexts, such as flight cockpits or virtual emergency rooms [26]. Because these environments can simulate realistic workload, timing constraints, and decision pressures, they are ideal for examining real-time decision-making, trust dynamics, and performance under conditions similar to actual deployments [6]. In such cases where deploying a real AI model is not feasible due to technology limitations or other resource constraints, virtual environments are combined with "Wizard-of-Oz" techniques, which is when a researcher acts as an AI to study human responses or to prototype interactions with hypothetical AI capabilities [27]. In contrast, *factorial surveys* present scenarios, or vignettes, that provide participants with systematically varied hypothetical situations that could involve AI assistance. While low in interaction fidelity, factorial surveys are useful for isolating participants' judgments about trust, reliance, or ethical decision-making in early-stage assessments of attitudes or intended behaviors [28, 29]. Lastly, *game-based simulations* resemble simplified video games or structured collaborative puzzles with clearly defined goals, constraints, and rules. These tasks are well-suited to studying teamwork processes, mutual adaptation, and communication patterns, as the gaming format naturally engages users in cooperative problem-solving and encourages dynamic human–AI interactions [30]. While all three techniques are used to measure and manipulate similar constructs of human-AI synergy, the methodological choice ultimately lies on the researcher's priorities with respect to realism, interactivity, scalability, and specific teaming phenomena.

Many HAI decision-support experiments in synthetic environments often follow the *pre-task→task→post-task* workflow. The *pre-task* stage covers participant on-boarding and baseline measurements. Participants provide informed consent and receive instructions about the scenario and a description of the AI decision support system. Often, a training or familiarization session is included, where the participant learns the interface and possibly practices on a few example tasks [31, 32]. Additionally, researchers frequently administer pre-task questionnaires to gather demographics, domain expertise, or dispositional traits. For example, a participant might fill out a propensity to trust automation survey before interacting with the AI [33]. Such pre-task indicators can later serve as covariates or baseline trust levels. In some studies, a pre-task briefing or scenario narrative is given to provide participants with the operational context, which can help better prepare participants for the task at hand and control for initial state conditions across groups.

In the *task execution* phase, study participants engage in the decision-making tasks with the AI system. The AI's behavior is usually systematically varied to test research hypotheses. Some common manipulations include the AI's performance level (reliability, accuracy), interface elements including data visualization or communication (transparency, explanations), or ability to complete the task relative to the human operator (levels of automation or autonomy). During study task trials, some STEs log directly observable data such as the participant's decisions, response times, behavior in response to AI's errors, etc. In some setups, self-reported measures are also collected concurrent with the task. For example, after each trial, or as interruptions within a trial, participants might be asked to rate their confidence in the AI's prediction. The task execution stage is where the observational focus is on the dynamics of humans-AI interaction and researchers take note of task execution phenomena such as over-reliance (e.g., mindlessly accepting AI suggestions) or skepticism (e.g., superfluously rejecting the AI recommendations).

After completing the decision task(s), participants move to debriefing and evaluation. The standard is to administer post-task questionnaires that probe participants' subjective experiences with the AI system. These often include trust questionnaires [34] and workload assessments [35]. Other common post-task instruments include usability or satisfaction surveys [36] and bespoke surveys about the AI's explainability or the participant's confidence in their own decisions.

The pre-task, task execution, and post-task structure has become a common framework in task-oriented STEs for HAI interaction research. Beyond serving as a study protocol, these phases reflect researcher interest in how trust in AI and related behaviors evolve over time. Stratifying the impact of expertise and other covariates across these study phases supports a clearer analysis of outcomes prior, during, and after exposure to an AI system. This structure also supports temporal analysis while revealing distinct phase-specific patterns, such as differences in expectations, behaviors, or shifts in perceptions that might be common in certain population groups and not others.

3. Who is an Expert?

How people perceive, interact with, and trust expert systems has been shown to depend on their level of expertise, a finding that is well-established in the literature, including trust in automation scholarship. *Expert performance* has been defined as "consistently superior performance on a specified set of representative tasks for a domain" [37], suggesting that expertise is strongly domain-specific. For example, chess Grandmasters have exceptional memory and recognition for chess patterns and movements but not necessarily for arbitrary information [38]. Experience also plays a role: foundational research suggests that it takes roughly about a decade of deliberate practice needed to reach world-class performance in many domains [37]. Graded models of expertise acquisition, such as proficiency scales, have also emerged by synthesizing traditional apprenticeship models with applied cognitive research [39].

While HAI research conceptually adopts these definitions, it often relies on more pragmatic and accessible indicators of expertise in practice. Some operational determinants of categorizing participants as "expert" or "novice" include:

- *Formal training or credentials.* Participants with professional qualifications or domain-specific training or education are typically considered experts. For example, a study on AI-assisted medical diagnosis categorized board-certified radiologists as experts and non-radiology physicians as non-experts [40]. In a study investigating the interpretation of post-hoc AI explanations, researchers operationalized domain expertise by comparing participants' performance across familiar (MNIST) and unfamiliar (Kannada-MNIST) image classification tasks. Since all participants were English-speaking adults, they were considered "deep-experts" in MNIST numerals and "non-experts" in Kannada numerals [41].
- *Experience level (duration or exposure).* Time on the job is a frequently used distinction. Some studies' definition depends on the minimum number of years working in the field [42], objective temporal measure of experience [43], or the amount of domain-relevant cases processed. For example, the U.S. Air Force distinguishes between experienced (expert) and inexperienced pilots (novice) using 500 flight hours as the threshold [44]. In another study on human resource recruitment, years of experience and volume of cases (applications) processed were used to calculate an aggregate measure of expertise [42].
- *Prior measures of performance.* Some studies use tests or past performance to assess expertise when a domain-relevant test is available. This is more common in traditional cognitive experiments such as using rating systems from tournaments to distinguish between chess novices and experts [39]. In other cases, performance-based grouping is applied retroactively; for instance, Bayer, et al. [45] classified participants as novices if they scored below a predefined threshold on the Amsterdam Chess Test, using performance as a post-hoc criterion for expertise.
- *Self-reported expertise or knowledge.* In the absence of objective indicators, studies rely on self-reports. Participants may be asked to rate their familiarity with AI [46] or proficiency with a task

domain, such as the anthropod image classification study in [17]. This approach is a relatively lax criterion but is sometimes adopted for its convenience in accounting for participant expertise.

However, many HAI studies do not formally categorize or measure participant expertise unless explicitly investigating differences between experts and non-experts. Further, the terms "novice" and "general population" are often used interchangeably in HAI literature. Clarifying this distinction is important: a *novice* is an individual who is new to a specific task or domain but may be undergoing training or has adjacent experience, while *general population* participants are people with no assumed background, training, or familiarity with the task context. For example, a crowdsourced worker with no prior training to identity screening tasks would be considered part of the general population, while an individual who has completed professional instruction on face verification procedures but lacks operational experience in real-world screening environments would be considered a novice.

While terminology of participant populations varies across the literature, we propose a more rigorous stratification of participant populations based on how researchers have implicitly or explicitly referred to participants in HAI studies:

- *Laypeople.* Individuals with no formal training, experience, or specialized knowledge in either the task domain or AI systems. This group often includes crowdsourced workers from Mechanical Turk or Prolific, university students, or members of the general public;
- *Technical (AI) novices.* Individuals with some exposure to machine learning, statistics, computer science, or mathematics.
- *Technical (AI) experts.* Individuals with advanced training or professional experience in AI, machine learning, or data science. This includes machine learning professionals or academics, and system developers with substantial experience with AI systems.
- *Domain novices.* Individuals with formal education or structured training in the task domain or an adjacent domain. Examples include general healthcare workers asked to do ultrasound diagnostics [47] or pilots with flight hours below the 500-hour threshold of expertise as defined by the U.S. Air Force.
- *Domain experts.* Individuals with both formal training and substantial real-world experience in the task domain, often possessing highly developed decision-making strategies and tacit knowledge relevant to the task.

Studies that explicitly investigate how different population groups perceive and/or interact with AI systems have presented evidence of cognitive and behavioral differences in decision-making contexts. For example, Ghaffar et al. conducted a comparative analysis of novice and expert optometrists reviewing glaucoma cases [48]. Although no specific AI system was deployed in the comparative observational study, researchers were able to isolate systematic differences in diagnostic concordance, reasoning strategies, and temporal focus: novices relied on structured rules and external comparisons, while experts used pattern recognition, intuitive synthesis, and strategically timed analytical assessments of simulated glaucoma cases. Chavaillaz, et al. compared laypersons (students with very minimal training on X-ray screening) and experts (airport screening professionals) in a simulated baggage screening task with the same diagnostic aid [49]. Only novices experienced performance gains, while experts demonstrated higher compliance and reliance with the decision-support tool, emergent outcomes that suggest they more likely used the AI to confirm rather than guide decisions. Portela, et al. revealed that domain experts (law enforcement professionals) approached an AI-supported recidivism prediction task differently from technical novices and experts (data scientists) or laypersons [16]. While the non-expert groups relied on their own heuristics (e.g., data scientists approached the task as a data-driven statistics problem), domain experts perceived recidivism risk as a negotiated judgment, not a binary prediction.

As more STEs are deployed in HAI research, there is an amplified need for design guidelines that align with the characteristics of the participant pool that is targeted for recruitment. Our objective is to find patterns of similarities or divergences in how different participant groups interact with AI systems in STE-based experiments, and to identify where design adaptations may be needed to support more valid and generalizable inferences.

4. HAI Dynamics as a Function of Participant Expertise Across STE Interaction Phases

Since the early introduction of automated decision-support systems, STE-based research, particularly prominent in aviation, has investigated how expertise affects attitudes and behaviors toward technology. Human factors studies in the 1990's, for example, demonstrated notable differences between novice and veteran pilots with respect to their initial perceptions and interactions with early-warning signals in flight simulators [50]. Such foundational studies established that initial attitudes and mental models – formed through factors such as personal experience and domain knowledge [51] – fundamentally influence how people perceive, respond to, and interact with technology.

Participant expertise continues to be an important factor in HAI research but with new questions arising due to the shift from specialized, deterministic systems to the more broadly available and probabilistic AI models. Earlier decision support systems, such as rule-based automation in pilot workflows, were designed for highly-trained professionals whose expertise levels were determined based on, for example, flight hours or years of experience. By contrast, contemporary AI-based systems are now widely used in many domains, albeit in less specialized functions to start, but with a rapidly increasing marketplace for tools that can allow these systems to be used in a more customized manner [52, 53]. Such systems may be made more widely available to laypeople, or minimally-trained professionals, under the guise that such technologies might allow them to perform more like experts.

In addition, the increasing availability and use of STEs for online studies has precipitated the use of crowd-sourced laypeople to study human-AI synergies. Consequently, how minimally-trained professionals and study participants alike interpret and respond to AI systems may no longer depend primarily on their understanding of the task or the interface, but also by their ability to interpret and act upon probabilistic, inherently uncertain, and frequently opaque outputs from AI models.

4.1. Methodology

Over six years of federally sponsored research supporting four different competitively-awarded projects, our project team has developed at least twelve iterations of four different STEs that emulate AI-based decision-support systems (DSS) in airport security and intelligence analysis decision contexts. These STEs were developed by our project team members — university students and faculty — with close involvement and input from subject matter experts through bi-monthly project meetings. These STEs were then tested with crowd-sourced study participants, volunteer participants, and, in the case of two STEs, with full-time professionals in the corresponding decision domain. This paper reports the preliminary findings of a systematic review of HAI studies in the literature that (1) explicitly compare and contrast HAI interaction dynamics across different participant pools and (2) focus on one participant group but investigate constructs that have yielded conflicting results in other studies with otherwise similar tasks or system features. Combined with our experience in designing STEs for domain experts, novices, and laypersons, our objective for the literature review is to provide HAI researchers with a framework (Figure 1) and actionable guidelines for designing STEs for a given participant population that are aligned with their cognitive strategies, prior knowledge, and task-relevant dispositions to support valid interpretations and generalizable conclusions within that population.

4.2. Our Framework

Our preliminary review and synthesis show that, even in STEs, trust is not simply a static attitude or preference; rather it is far more contextual than how most STE based studies tend to represent it or measure it. Trust emerges, fluctuates, recalibrates, and these dynamics are not uniform across participant populations. Critically, trust behaviors vary systematically with participant expertise, expertise that reflects differences in domain knowledge, task familiarity, and technical fluency.

Figure 1 presents a conceptual framework showing how participant expertise may impact human-AI interaction dynamics by influencing how other manipulable study factors, such as AI framing or task

difficulty, affect key responses often measured in HAI research (e.g., reliance or compliance behaviors). We are particularly interested in empirical work that demonstrate a "two-factor interaction" effect (in statistical terms) between participant expertise and a moderator variable. For an example of such pattern, the same AI framing (e.g., framing the AI as an expert collaborator) may result in higher compliance behaviors among novice operators but have little to no effect – or even the opposite impact – among domain experts. Our preliminary review reveals these trends in recent literature – e.g., Rutjes, et al. showed a statistically significant effect between participant expertise and transparency levels. Domain novices were more likely to align their recommendations with the AI at medium levels of transparency while experts were more likely to provide congruent recommendations under high-transparency conditions [54]. Similarly, in Portela et al., domain experts revised their predictions more often in response to AI suggestions than technical novices/experts or laypersons, based on a pre-post comparison of their own predictions before and after viewing the AI's recommendation [16]. This suggests a likely active two-factor interaction (although not statistically tested) between participant expertise and the presence or absence of AI recommendations. Accordingly, the design and deployment of STEs for HAI research should account for these participant differences, if the goal of this research is to address decision performance factors in operational environments that have to deal with task complexity, system uncertainty, and participant capability across the three common phases of STE experiments: pre-interaction, in-situ interaction, and post-interaction.

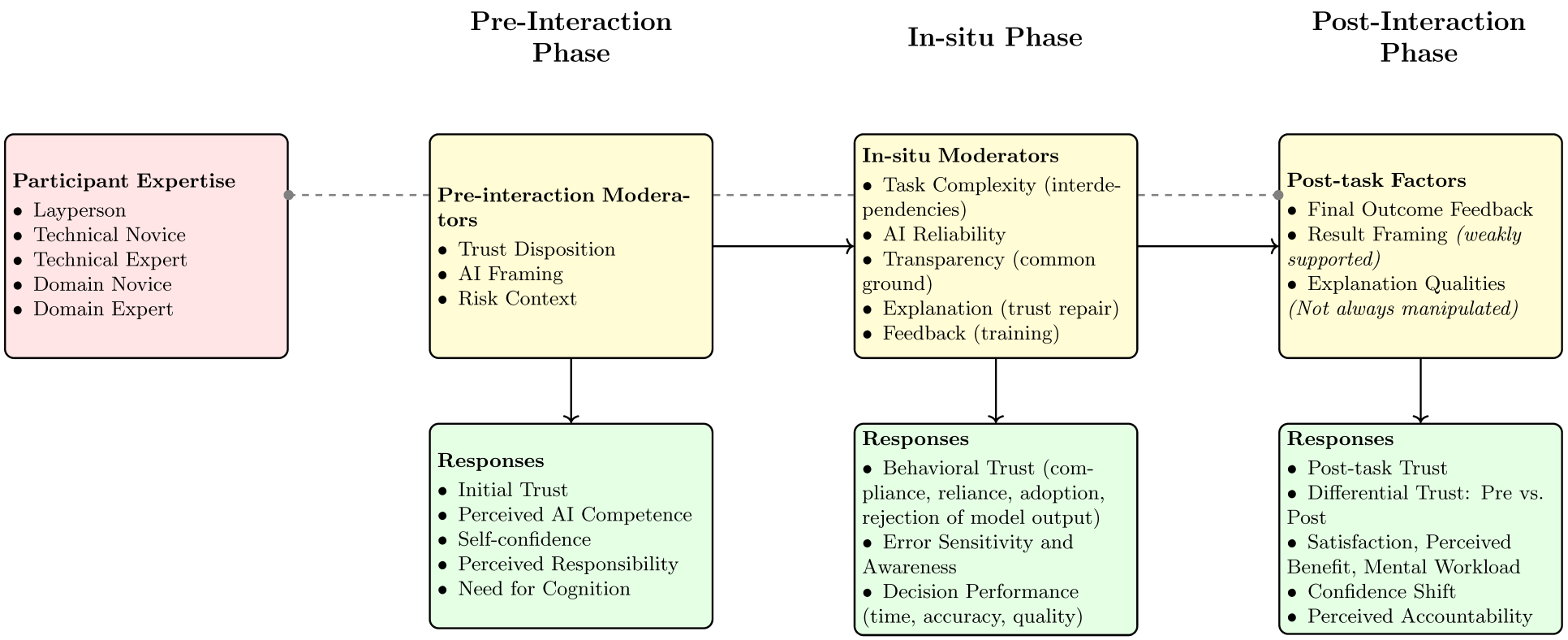

Figure 1: Hypothesized framework illustrating how participant expertise interacts with key STE moderators and influences participant responses across the pre-interaction, in-situ, and post-task phases of STE-based HAI studies.

References

[1] E. National Academies of Sciences, Medicine, Human-AI Teaming: State of the Art and Research Needs, The National Academies Press, Washington, D.C., 2021. URL: https://www.nap.edu/catalog/26355. doi:10.17226/26355.

[2] K. E. Schaefer, J. Y. C. Chen, J. L. Szalma, P. A. Hancock, A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems, Human Factors: The Journal of the Human Factors and Ergonomics Society 58 (2016) 377–400. doi:10/gf4sjg.

[3] D. D. Woods, Cognitive technologies: The design of joint human-machine cognitive systems, AI Magazine 6 (1985) 86–86. doi:10/grpwb9.

[4] R. Iftikhar, Y.-T. Chiu, M. S. Khan, C. Caudwell, Human–agent team dynamics: A review and future research opportunities, IEEE Transactions on Engineering Management 71 (2024) 10139–10154. doi:10/g9dtdf.

[5] B. Brehmer, D. Dörner, Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study, Computers in Human Behavior 9 (1993) 171–184. doi:10/cxzwnm.

[6] W. D. Gray, Simulated task environments: The role of high-fidelity simulations, scaled worlds, synthetic environments, and laboratory tasks in basic and applied cognitive research, Cognitive Science Quarterly 2 (2002) 205–227.

[7] E. Martin, D. R. Lyon, B. T. Schreiber, Designing synthetic tasks for human factors research: An application to uninhabited air vehicles, Proceedings of the Human Factors and Ergonomics Society Annual Meeting 42 (1998) 123–127. doi:10/fzcv88.

[8] E. Peer, L. Brandimarte, S. Samat, A. Acquisti, Beyond the turk: Alternative platforms for crowdsourcing behavioral research, Journal of Experimental Social Psychology 70 (2017) 153–163. doi:10.1016/j.jesp.2017.01.006.

[9] G. Klein, B. Shneiderman, R. R. Hoffman, K. M. Ford, Why expertise matters: A response to the challenges, IEEE Intelligent Systems 32 (2017) 67–73. doi:10/gg98st.

[10] R. J. B. Hutton, G. Klein, Expert decision making, Systems Engineering 2 (1999) 32–45. doi:10/cj6cv7.

[11] Z. Buçinca, M. B. Malaya, K. Z. Gajos, 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 (2021) 1–21. doi:10.1145/3449287.

[12] P. Salehi, E. K. Chiou, M. V. Mancenido, A. Mosallanezhad, M. C. Cohen, A. Shah, Decision deferral in a human-ai joint face-matching task: Effects on human performance and trust, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, volume 65, SAGE Publications, 2021, pp. 638–642. doi:10.1177/1071181321651157.

[13] S. Salimzadeh, G. He, U. Gadiraju, Dealing with uncertainty: Understanding the impact of prognostic versus diagnostic tasks on trust and reliance in human-ai decision making, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–17.

[14] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human Factors 46 (2004) 50–80. doi:10.1518/hfes.46.1.50_30392.

[15] E. Glikson, A. W. Woolley, Human trust in artificial intelligence: Review of empirical research, Academy of Management Annals 14 (2020) 627–660. doi:10.5465/annals.2018.0057.

[16] M. Portela, C. Castillo, S. Tolan, M. Karimi-Haghighi, A. A. Pueyo, A comparative user study of human predictions in algorithm-supported recidivism risk assessment, Artificial Intelligence and Law (2024). URL: https://doi.org/10.1007/s10506-024-09393-y. doi:10.1007/s10506-024-09393-y.

[17] M. Nourani, J. T. King, E. D. Ragan, The role of domain expertise in user trust and the impact of first impressions with intelligent systems, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 8, AAAI Press, 2020, pp. 122–131.

[18] N. J. Cooke, K. Rivera, S. M. Shope, S. Caukwell, A synthetic task environment for team cognition research, Proceedings of the Human Factors and Ergonomics Society Annual Meeting 43 (1999)

303–308. doi:10/fzd6p7.
[19] M. Wong, A. Ezenyilimba, T. Anderson, A. Wolff, E. Chiou, M. Demir, N. Cooke, A remote synthetic testbed for human-robot teaming: An iterative design process, in: 65th Annual Meeting of the Human Factors and Ergonomics Society, Baltimore, Maryland, USA, 2021.
[20] G. J. Lematta, P. B. Coleman, S. A. Bhatti, E. K. Chiou, N. J. McNeese, M. Demir, N. J. Cooke, Developing human-robot team interdependence in a synthetic task environment, Proceedings of the Human Factors and Ergonomics Society Annual Meeting 63 (2019) 1503–1507. doi:10/ghzbvn.
[21] F. Raimondo, A. T. Wolff, A. Hehr, M. Wong, M. A. Peel, M. Demir, N. J. Cooke, E. K. Chiou, Trailblazing roblox virtual synthestic testbed development for human-robot teaming studies, in: Proceedings of the 66th International Annual Meeting of the Human Factors and Ergonomics Society, Atlanta, GA, 2022.
[22] M. C. Cohen, N. Kim, Y. Ba, A. Pan, S. Bhatti, P. Salehi, J. Sung, E. Blasch, M. V. Mancenido, E. K. Chiou, Padthai-mm: Principles-based approach for designing trustworthy, human-centered ai using the mast methodology, AI Magazine 46 (2025) e70000. doi:10/g9fm69.
[23] J. E. Laird, K. Gluck, J. C. Lester, Interactive virtual testbeds for studying human-ai collaboration, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 13567–13575. doi:10.1609/aaai.v34i09.7112.
[24] D. J. Phillips, A. Smart, M. M. Smith, Factorial survey experiments in human-ai interaction research, Journal of Experimental Psychology: Applied 27 (2021) 435–450. doi:10.1037/xap0000358.
[25] J. B. Hamrick, P. Battaglia, J. B. Tenenbaum, Game-based environments as ai-human collaboration testbeds, in: Proceedings of the International Conference on Machine Learning (ICML), 2019, pp. 2312–2321. doi:10.5555/3327757.3327888.
[26] S. Ellis, What are virtual environments?, IEEE Computer Graphics and Applications 14 (1994) 17–22. doi:10/dkggts.
[27] L. D. Riek, Wizard of oz studies in hri: A systematic review and new reporting guidelines, Journal of Human-Robot Interaction 1 (2012) 119–136. doi:10/ggbn82.
[28] K. Auspurg, T. Hinz, Sage Research Methods - Factorial Survey Experiments, Quantitative Applications in the Social Sciences, SAGE Publications, Inc., 2015. URL: https://doi.org/10.4135/9781483398075.
[29] T. Li, M. Vorvoreanu, D. Debellis, S. Amershi, Assessing human-AI interaction early through factorial surveys: A study on the guidelines for human-AI interaction, ACM Trans. Comput.-Hum. Interact. 30 (2023) 69:1–69:45. doi:10/gscn5w.
[30] Q. Zhang, Understanding human-ai teaming dynamics through gaming environments, Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment 19 (2023) 440–443. doi:10/g9ftg2.
[31] P. Salehi, Y. Ba, N. Kim, A. Mosallanezhad, A. Pan, M. C. Cohen, Y. Wang, J. Zhao, S. Bhatti, J. Sung, E. Blasch, M. V. Mancenido, E. K. Chiou, Towards trustworthy ai-enabled decision support systems: Validation of the multisource ai scorecard table (mast), Journal of Artificial Intelligence Research 80 (2024) 1311–1341. URL: https://jair.org/index.php/jair/article/view/14990. doi:10.1613/jair.1.14990.
[32] J. Zhao, Y. Wang, M. V. Mancenido, E. K. Chiou, R. Maciejewski, Evaluating the impact of uncertainty visualization on model reliance, IEEE Transactions on Visualization and Computer Graphics 30 (2024) 4093–4107. URL: https://doi.org/10.1109/TVCG.2023.3251950. doi:10.1109/TVCG.2023.3251950.
[33] D. D. Scholz, J. Kraus, L. Miller, Measuring the propensity to trust in automated technology: Examining similarities to dispositional trust in other humans and validation of the ptt-a scale, International Journal of Human–Computer Interaction 41 (2025) 970–993. doi:10/g9f8fg.
[34] A. Alsaid, M. Li, E. K. Chiou, J. D. Lee, Measuring trust: A text analysis approach to compare, contrast, and select trust questionnaires, Frontiers in Psychology 14 (2023) 1192020. doi:10/gs53sq.
[35] S. G. Hart, Nasa-task load index (nasa-tlx); 20 years later, Proceedings of the Human Factors Society 50th Annual Meeting (2006). doi:10/fzvtd4.
[36] J. R. Lewis, The system usability scale: Past, present, and future, International Journal of Human-

Computer Interaction 34 (2018) 577–590. doi:10.1080/10447318.2018.1455307.
[37] K. A. Ericsson, A. C. Lehmann, Expert and exceptional performance: Evidence of maximal adaptation to task constraints, Annual Review of Psychology 47 (1996) 273–305. doi:10.1146/annurev.psych.47.1.273.
[38] W. G. Chase, H. A. Simon, Perception in chess, Cognitive Psychology 4 (1973) 55–81. doi:10.1016/0010-0285(73)90004-2.
[39] M. T. H. Chi, Two approaches to the study of experts' characteristics, in: K. A. Ericsson, N. Charness, P. J. Feltovich, R. R. Hoffman (Eds.), The Cambridge Handbook of Expertise and Expert Performance, Cambridge University Press, Cambridge, UK, 2006, pp. 21–30. doi:10.1017/CBO9780511816796.004.
[40] S. Gaube, H. Suresh, M. Raue, E. Lermer, T. K. Koch, M. F. C. Hudecek, A. D. Ackery, S. C. Grover, J. F. Coughlin, D. Frey, F. C. Kitamura, M. Ghassemi, E. Colak, Non-task expert physicians benefit from correct explainable ai advice when reviewing x-rays, Scientific Reports 13 (2023) 1383. URL: https://doi.org/10.1038/s41598-023-28633-w. doi:10.1038/s41598-023-28633-w.
[41] C. Ford, M. T. Keane, Explaining classifications to non-experts: An xai user study of post-hoc explanations for a classifier when people lack expertise, in: J.-J. Rousseau, B. Kapralos (Eds.), Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges, Springer Nature Switzerland, Cham, 2023, pp. 246–260.
[42] A. Lacroux, C. Martin-Lacroux, Should i trust the artificial intelligence to recruit? recruiters' perceptions and behavior when faced with algorithm-based recommendation systems during resume screening, Frontiers in Psychology Volume 13 - 2022 (2022). URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.895997. doi:10.3389/fpsyg.2022.895997.
[43] J. B. Lyons, N. T. Ho, A. L. Van Abel, L. C. Hoffmann, G. G. Sadler, W. E. Fergueson, M. A. Grigsby, M. Wilkins, Comparing trust in auto-gcas between experienced and novice air force pilots, Ergonomics in Design 25 (2017) 4–9. doi:10.1177/1064804617716612.
[44] J. S. Miranda, Are You Experienced? A Fresh Look at the Fifth-Generation Fighter Experience Model, Technical Report, Air Command and Staff College, Air University, Maxwell Air Force Base, Alabama, 2016. URL: https://apps.dtic.mil/sti/citations/AD1038494, aU/ACSC/2016.
[45] S. Bayer, H. Gimpel, M. Markgraf, The role of domain expertise in trusting and following explainable ai decision support systems, Journal of Decision Systems 32 (2023) 110–138. doi:10.1080/12460125.2021.1958505.
[46] J. J. Gualda-Gea, L. E. Barón-Miras, M. J. Bertran, A. Vilella, I. Torá-Rocamora, A. Prat, Perceptions and future perspectives of medical students on the use of artificial intelligence based chatbots: an exploratory analysis, Frontiers in Medicine 12 (2025) 1529305. URL: https://www.frontiersin.org/articles/10.3389/fmed.2025.1529305/full. doi:10.3389/fmed.2025.1529305.
[47] C. Baloescu, J. Bailitz, B. Cheema, R. Agarwala, M. Jankowski, O. Eke, R. Liu, J. Nomura, L. Stolz, L. Gargani, E. Alkan, T. Wellman, N. Parajuli, A. Marra, Y. Thomas, D. Patel, E. Schraft, J. O'Brien, C. L. Moore, M. Gottlieb, Artificial intelligence–guided lung ultrasound by nonexperts, JAMA Cardiology 10 (2025) 245–253.
[48] F. Ghaffar, N. M. Furtado, I. Ali, C. Burns, Diagnostic decision-making variability between novice and expert optometrists for glaucoma: Comparative analysis to inform ai system design, JMIR Med Inform 13 (2025) e63109.
[49] A. Chavaillaz, A. Schwaninger, S. Michel, J. Sauer, Expertise, automation and trust in x-ray screening of cabin baggage, Frontiers in Psychology 10 (2019).
[50] K. L. Mosier, L. J. Skitka, S. Heers, M. Burdick, Automation bias: Decision making and performance in high-tech cockpits, Technical Report NASA-CR-201422, NASA Ames Research Center, Moffett Field, CA, 1996.
[51] S. J. Payne, A descriptive study of mental models†, Behaviour Information Technology 10 (1991) 3–21. doi:10/dh8nvs.
[52] E. C. Garrido-Merchán, J. L. Arroyo-Barrigüete, F. Borrás-Pala, L. Escobar-Torres, C. Martínez de Ibarreta, J. M. Ortíz-Lozano, A. Rua-Vieites, Real customization or just marketing: Are customized versions of generative ai useful?, F1000Research 13 (2024) 791. doi:10/g9f845.

[53] V. Shukla, G. G. Parker, Building custom large language models for industries: A comparative analysis of fine-tuning and retrieval-augmented generation techniques, in: 2024 International Conference of Adisutjipto on Aerospace Electrical Engineering and Informatics (ICAAEEI), 2024, p. 1–6. URL: https://ieeexplore.ieee.org/abstract/document/10899129. doi:10/g9f848.

[54] H. Rutjes, M. C. Willemsen, W. A. IJsselsteijn, Tailoring transparency to expertise: Health professionals' need for transparency in representing self-tracking data of clients, in: 4th HUMANIZE Workshop on Transparency and Explainability in Adaptive Systems through User Modeling Grounded in Psychological Theory (at IUI'20), Cagliari, Italy, 2020. URL: https://research.tue.nl/en/publications/tailoring-transparency-to-expertise-health-professionals-need-fo, workshop at the ACM Conference on Intelligent User Interfaces (IUI 2020).