Meaningful Control in Human-AI Systems: Trusting AI Agents

Erin K. Chiou, Jessica R. Lee
Arizona State University
USA

erin.chiou@asu.edu jrlee21@asu.edu

ABSTRACT

Past research has established that understanding human trust can help with understanding what drives behaviors like reliance and compliance that affect performance with automated systems. However, increasingly capable and interactive AI-enabled systems have challenged the paradigm of high-performing machines in well-defined task environments with a supervising human in the loop. The promise of AI-enabled capabilities has also ushered in scholarly perspectives like human-AI teaming for achieving superior system performance. As a result, designing for meaningful human control remains an open area of research that has long been the focus of trust in automation scholarship. Recent conceptual work investigating trust in these new contexts posits that a relational view of trust in automation may be a useful direction for achieving meaningful control in complex environments. This relational view suggests that simply providing information to either foster or calibrate people's trust in AI will be insufficient for human-AI teaming success in the long run. Success in the long run will involve going beyond established performance qualities like reliability, dependability, and predictability to also consider relational qualities like responsivity and its effect on operator engagement in complex and under-defined task environments. This paper presents recent work applying this relational lens to operationalize responsivity as a measurable concept and describes how responsivity might help broaden and direct recent human-AI teaming research objectives for more meaningful human control.

1.0 INFORMATIONAL OR RELATIONAL

Recent scholarly discourse on trust in automation proposes a more relational view of trust in AI that considers how situation structures in interactive task environments shape how trust evolves over time [1]. This relational view expands upon previous approaches to trust and empirical studies that presume a paradigmatic relational structure for human-machine systems known as supervisory control. Supervisory control describes a human operator supervising a highly capable automated system, and the primary outcome of interest is usually signal detection performance. In the context of supervisory control, work on improving human-machine performance tends to focus on either 1) factors in the work environment that can affect human performance like workload, task complexity, interruptions or 2) the amount, quality, or timing of information transmitted by the machine that can affect human performance. Along these lines, past research has shown that information about a system's purpose, process, and performance can affect a person's trust and subsequent decisions to rely or comply with the machine [2]. This is referred to as the information processing view of human-machine interactions, in which a human operator is treated like an information processor with limited resources, and the primary resolution for performance issues related to trust is about figuring out when to present the right information at the right time to maintain situation awareness, and other measures of meaningful control.

While focusing on the information available and transmitted between human-machine counterparts may in fact be the best way to improve work environments that demand better information transparency [3], [4], and relatedly, qualities of effective communication remains a core focus of human-AI teaming literature [5], [6], [7], the ideal of optimizing information transfer between interacting counterparts is addressing a relatively narrower problem compared to the scope of problems that characterize working together effectively in complex task environments that consistently require human judgment. Rather, complex task environments that require human judgment are mired in persistent information asymmetry, uncertainty, and at times needing to make difficult decisions despite the information communicated. Therefore, limiting the scope of our problem to the information transmitted is not a panacea for maintaining meaningful human control. With bounded rationality [8], people can be overwhelmed by many options, and too much information can lead to decision fatigue, which may inspire offloading work and corresponding feelings of responsibility to an available machine counterpart.

Like the information processing view, the relational view supports system operators having the right information at the right time; there is not much to debate about the need to process and communicate information. However, a relational view also addresses the interdependencies between heterogenous agents, and is an especially useful lens for social situations, which are all situations that involve at least one person [9], [10]. A relational view takes the dyad as the smallest unit of analysis, addressing problems that cannot be addressed by simply evaluating the aggregate of individual performance or perceptions, cataloguing acceptable ranges, and then coordinating known task dependencies [11]. Rather, a dyadic unit of analysis helps to address meso-level [12] human systems issues that are essential for achieving competitive advantage in fast paced and evolving landscapes. These human systems issues include the need to address how human-machine teams can be flexible, adaptable, adaptive, and cooperative. Optimizing for information speed and scale remain important capabilities that automation and AI-enabled systems can help to address, but so is the ability to navigate, direct, and pivot quickly between the tasks that can be optimized. How to support this ability is the subject of understanding relational trust to improve the design of AI-enabled agents, and what it means to have meaningful control in complex task environments.

While a relational worldview can help to reconceptualize the trust problem in a more robust way, one criticism is that it needs to be further operationalized for human-machine system design. Effective concepts, as operational models, should be testable and falsifiable [13], which is difficult to do with abstract concepts. One potential approach to operationalizing the relational worldview is to develop the concept of responsivity. To be relational, there must be a responsive other [14], and because responsivity is a quality, rather than a worldview, this means it can be more easily operationalized as items for evaluation.

2.0 RESPONSIVITY

Responsivity refers to the ability to anticipate and provide useful responses in varied situations within specific goal environments. Although the concept of responsivity is relatively new in the human-machine systems discourse, it is also a theoretically robust concept rooted in social psychological theory [15], reflected in outcome-driven approaches like the technology acceptance model [16] and is considered a critical element in establishing the mutual understanding and adaptability that underpin successful long-term collaborations [17], [18]. As a unifying concept, responsivity is promising for understanding and designing for meaningful control in human-machine teams [19] because teaming is not just about information transfer performance but also understanding how social dynamics can affect and predict productive interactions in the long run. Responsivity goes beyond usability and usefulness to include interdependent goals and decision situations that are relevant for complex tasks involving multiple agents. Moreover, responsivity in meaningful human control involves tracking and responding to the reasons and intentions of human agents [20]. Yet, to our knowledge there is no delineation of what responsivity entails for human-machine systems, nor is there a specific instrument to assess responsivity, indicating a need for further operationalization.

3.0 OPERATIONALIZING RESPONSIVITY – PRELIMINARY WORK

In a pilot project called “Testing Responsivity as a Unifying Signal of Trust Measurement and Evaluation” with support from the U.S. Department of Defense’s Chief Digital and Artificial Intelligence Office (CDAO) we are developing a responsivity evaluation tool to reflect the current state-of-the-art in relational trust theory, and to address the practical challenges that afflict current trust assessment methods, a topic out of scope for the current paper. Our ongoing efforts have resulted in five preliminary criteria that will be cross-checked for content validity within the empirical literature and then operationalized as a rubric with individual items and response levels. The resulting rubric will be tested and evaluated in a human participant study to assess construct validity, criterion-related validity, and to refine the overall instrument into a more agile tool. Here we present our preliminary work to develop and validate criteria for assessing machine responsivity in human systems. To address the essential gaps that the relational view first identified, we first focus on the smallest possible unit of analysis for responsivity, a human-machine dyad teaming within a shared work context.

3.1 Developing and Validating Responsivity Criteria

To develop our initial responsivity criteria, our project team members comprising subject matter experts compiled a list of issues that are known to arise in human-machine work systems, and that could potentially affect an operator's trust in a machine counterpart. We ultimately identified 15 relevant items. To synthesize our ideation and to evaluate for their potential overlap, we then organized those 15 items into five main criteria, with the 15 items comprising sub-criteria.

To assess content validity of our criteria, we are conducting a confirmation review of literature to investigate whether there are empirical studies that explicitly support or detract from the responsivity criteria we developed. Results from 16 different Boolean searches that combined key terms like "Adaptability AND Trust* AND (Automation OR AI OR "artificial intelligence" OR Computer OR Autonomy OR Robot OR Machine OR Technology) AND (People OR Human OR Person) AND Interact*" in Scopus, IEEE, Engineering Village, and PsycINFO databases, with reference scavenging from key articles from Google Scholar, returned 2,981 unique titles of relevant articles. Based on their titles, and whether an article was duplicated across the different databases indicating potentially higher relevance to both engineering and psychology fields, we further catalogued these articles into different levels of relevance, and the highly relevant articles (11% or 315 articles) are being prioritized for full review to identify potential gaps, alignment, and misalignments with our criteria.

Our five working criteria for assessing responsivity have been summarized as: Ability, Human-Centered, Collaboration, Contextualization, and Prioritization. These criteria are described in more detail below, drawing from their respective sub-criteria:

- *Ability* refers to the machine having access to appropriate resources, including task information, adequate response time, and essential hardware or software components for a function. The machine should also have the necessary privacy protections and interaction styles that would move a human counterpart to interact with it. If the machine lacks the resources required to perform a task to expectation, the machine should appropriately communicate this in a timely manner.
- *Human-Centered* refers to the machine adequately anticipating its human counterpart needs and preferences, predicting and noticing when the person misses a relevant signal and acting on or notifying appropriately. Human-Centered also includes a history of interactions, e.g., where more familiarity with personality and work style can better predict unnoticed signals.
- *Collaboration* refers to the task allocation between the human and machine counterparts being made clear, with the machine being capable of proactivity that is appropriate for the situation. The resulting task allocation should be informed by, or justified against, learned human preferences, such as interaction style.
- *Contextualization* refers to the machine effectively understanding its environment and context for its task and broader work system. The machine should actively monitor, perceive, and interpret new signals within this context, and signals can include changes to the environment, feedback from the environment (including from the human counterpart), or changes in the machine's own capabilities.

- *Prioritization* refers to the machine prioritizing shared goals (e.g., the commander's goals), even when they conflict with the human counterpart's goals (i.e., local goals). The machine should recognize these distinctions, and act complementarily to the shared goals. If appropriate, the machine can provide an uncertainty level to estimate its alignment with the shared goals.

We have found substantial support for each of our initial criteria in the extant literature and have refined our sub-criteria descriptions to better align with the literature. Several references also connect the responsivity concept studied with trust responses or trust related outcomes. We have not found literature that has identified gaps in our criteria; however, some of our sub-criteria, like the importance of feedback and learned preferences, are much easier to find support for than others. This seems to indicate that our responsivity criteria might not be a direct reflection of what is being written about most often in empirical studies of human-machine systems, but that the criteria robustly capture key qualities that enable productive human-machine relationships, with some notable support that many of these qualities relate to trust. Operationalizing these criteria into a clear, testable rubric that can generalize across AI systems will be crucial for further validating these criteria empirically and systematically.

4.0 RESPONSIVITY FOR ADVANCING HUMAN-AI TEAMING

Just as the concept of situation awareness bakes in the knowledge that an operator's attention and decision performance in information-rich environments is affected by the informational qualities of other human or machine agents, we posit that the concept of responsivity bakes in the knowledge that a person's trust and subsequent interactions with others depends on relational factors in the social and task environment. Both situation awareness and responsivity concepts are concerned with improving human system performance through system design and evaluation. As a result, like the impact that situation awareness has had in improving information-based performance across multiple time scales and levels of a work system [21], [22], [23], our responsivity criteria addresses multiple time scales and levels of a work system and can also be used to guide future research objectives towards more meaningful control in human-AI teams.

For example, we can apply the concept of responsivity to the top research priorities identified at a recent workshop held at the U.S. Naval Information Warfare Center-Pacific. The top research priorities identified for advancing human-AI teaming were: *1) advancing human-AI team effectiveness metrics, 2) advancing human-AI team testbeds, 3) establishing the efficacy of novel human-AI team task sharing paradigms, 4) developing AI awareness of the human teammate, and 5) forming development teams focused on human-AI teams* [24]. These research priorities were selected from a larger list of 57 research objectives identified in a National Academies of Sciences, Engineering, and Medicine consensus study supported by the U.S. Air Force Research Laboratory 711th Human Performance Wing [25]. Whereas the consensus study reviewed current state-of-the-art research areas in human-AI teaming and identified future research objectives in these areas, the workshop focused on how to further prioritize these objectives for the Navy given workshop participants' familiarity with current capabilities and opportunities for collaboration. However, both efforts generally leave up to the research community how to go about pursuing these objectives.

Responsivity may serve as a starting point to meaningfully broaden the scope of the research objectives as described in the consensus study, while providing direction for several of the workshop's prioritized objectives. For example, our project to develop a responsivity assessment tool addresses research priority *1) advance human-AI team effectiveness metrics,* by serving as an integrative measure of both machine trustworthiness and team effectiveness that goes beyond simply looking at team task performance metrics and perception measures. The responsivity criteria help to achieve this by focusing the research community's attention on the relationship between quantifiable performance measures (e.g., AI response time or accuracy within certain bounds) and other critical perceptions of AI that can influence trust, such as the *appropriateness* of the AI's response timing, or *how* the AI system addresses being used outside its performance boundaries. Fast AI response times could also mean that the response is disruptive unless responsivity is considered, e.g., the timing of the response results in a deleterious interruption. Responsivity also considers the need to communicate about potential limitations in-the-moment, rather than simply demonstrating superior performance in restricted cases and attaching a terms of use document to address the remaining cases.

Responsivity can also be used to inform what types of data and situation structures would be valuable to include for research priority *2) advancing human-AI team testbeds*. Our *Ability* criterion for responsivity would show that it is insufficient to invest solely in testbeds that demonstrate whether AI agents reliably or robustly fulfill a function. Rather, the ability to clearly communicate limitations and accommodate human preferences and needs is crucial for the sustained success of human-machine teams and should be incorporated into modern testbeds if responsivity is a concern. Therefore, testbeds mainly used to demonstrate AI superiority over human performance in specific use cases, and not also how the AI system communicates and adapts to different people or situations, may be limited in their evaluation of whole system abilities and limited in predicting the likelihood of AI adoption in the field.

Our responsivity criteria can also help direct the third prioritized research objective, *3) establishing the efficacy of novel human-AI team task sharing paradigms,* a reflection of the perceived inadequacy of relying on the HABA-MABA (humans-are-better-at; machines-are-better-at) function allocation paradigm for work system design. Though HABA-MABA works in well-planned and controlled task environments like the factory mass production lines in the early 20th century, the most advantageous teaming capabilities specifically address work environments that function allocation alone cannot handle. These teaming capabilities include adaptation and resilience in operational environments laden with uncertainty, that require many in-the-moment decisions against well-laid plans, and human judgment in the face of pluralistic values and long-range goals. This type of environment tempts conflict among interacting agents and invites miscommunication and misinformation. Effectively adapting the whole system to coordinate well and gain competitive advantage will require the ability to build and repair trust, to foster honesty and cooperation. It will be difficult to achieve these abilities without responsivity among the interacting agents within the system.

Although *4)* and *5)* were categorized as far-term research priorities from the workshop, both offer support for the need to develop responsivity as a concept while underscoring the technical and organizational challenges of achieving responsivity. In *4) developing AI awareness of the human teammate,* having this awareness would clearly enable the AI to be more responsive to human counterparts' needs and preferences. Correspondingly, AI context or human awareness is a common theme across many of the responsivity criteria. Understanding what is important to be aware of is something that we expect the responsivity criteria can help to identify, although the technical challenges of how to achieve sufficient awareness remains. This brings us to *5) forming development teams focused on human-AI teams,* a slight rephrase by the workshop from the original consensus study (research objective 10-3 [25]) that focused more on developing new approaches to forming research teams with multi-disciplinary competencies. The rephrased priority highlights how diverse areas of expertise and persistent human effort will be needed to continuously improve the efficacy of AI teammates because it is infeasible to anticipate all possible situations and needs that will arise in the field. These far-term priorities reflect the major intellectual and practical challenges that remain in operationalizing responsivity for system design.

5.0 RESPONSIVITY AS A MECHANISM FOR MEANINGFUL CONTROL

As human-AI teaming research objectives aim to help realize high performing systems, responsivity promises to address challenges in realizing meaningful control for these systems. In complex work environments characterized by uncertainty and information asymmetry, mechanisms like tracking and tracing are widely recognized as essential to ensure meaningful human control over AI systems [26]. Tracking ensures AI systems align their actions with human moral reasoning, adjusting decisions to reflect ethical considerations. Tracing, in turn, provides transparency by linking AI decisions to human oversight, supporting design, deployment, and operational accountability. Together, these concepts aim to establish clear lines of responsibility and AI system alignment with human values and intentions [20], [27], [28]. However, while important, tracking and tracing do not fully address the in-the-moment issues that arise at the interactional, situational, and relational levels [1], [18]. This is where responsivity can be particularly useful.

Building on the previous discussion about the limitations of the information processing view, and the need for a relational approach to trust in human-machine interactions, responsivity offers a practical mechanism to bridge these gaps. It acknowledges that complete control is unattainable in open-world environments and posits that a scientific understanding of how people respond to AI agents leads to more meaningful control – beyond merely imposing accountability mechanisms like supervision [19]. Defined by criteria such as Ability, Human-Centeredness, Collaboration, Contextualization, and Prioritization, responsivity accounts for how people will engage relationally with technology and proffers how AI systems can engage relationally with people. By operationalizing these criteria, as we have begun to do, we address the gaps identified by the relational view, focusing on the human-machine dyad within a shared work context.

Unlike tracking and tracing, which are high-level evaluation criteria for assessing how effectively a system embodies meaningful control [27], responsivity goes further by considering how trust in other entities depends not just on accountability mechanisms, the information presented, or prior experience, but also on how the entity responds to an individual's desires, intentions, and reasonings in the moment. This aligns with our emphasis on the need for AI systems to be flexible, adaptable, and cooperative – qualities that are essential in environments requiring frequent human judgment. Operationalizing responsivity to include individual, group, and organizational concerns further ensures that AI actions remain aligned with evolving situations, human values, and ethical standards – meeting people where they are individually and situationally. Embedding responsivity into AI systems enables adaptable and trust-based interactions. As a result, AI systems can enhance team decision-making, uphold ethical engagement, and support meaningful control in complex work environments.

6.0 REFERENCES

[1] E. K. Chiou and J. D. Lee, "Trusting automation: Designing for responsivity and resilience," Hum. Factors, Apr. 2021, doi: 10/gjvcr2.

[2] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," Hum. Factors, p. 31, 2004, doi: 10/dr6jf9.

[3] C. Hurter et al., "Usage of more transparent and explainable conflict resolution algorithm: air traffic controller feedback," Transp. Res. Procedia, vol. 66, pp. 270–278, Jan. 2022, doi: 10/grpnbr.

[4] Y. Zou and C. Borst, "Investigating transparency needs for supervising unmanned air traffic management systems," presented at the 13th SESAR Innovation Days, Sevilla, Spain, 2023.

[5] N. J. Cooke, J. C. Gorman, C. W. Myers, and J. L. Duran, "Interactive team cognition," Cogn. Sci., vol. 37, no. 2, pp. 255–285, Mar. 2013, doi: 10/gf69qb.

[6] S. Zhou and J. C. Gorman, "The impact of communication timing and sequencing on team performance: A comparative study of human-AI and all-human teams," Proc. Hum. Factors Ergon. Soc. Annu. Meet., p. 10711813241275090, Aug. 2024, doi: 10/g2f9bv.

[7] C. Liang, J. Proft, E. Andersen, and R. A. Knepper, “Implicit communication of actionable information in human-AI teams,” in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, in CHI ‘19. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–13. doi: 10.1145/3290605.3300325.

[8] H. A. Simon, “Bounded rationality,” in Utility and Probability, J. Eatwell, M. Milgate, and P. Newman, Eds., London: Palgrave Macmillan UK, 1990, pp. 15–18. doi: 10.1007/978-1-349-20568-4_5.

[9] B. Reeves and C. I. Nass, The media equation: How people treat computers, television, and new media like real people and places. in The media equation: How people treat computers, television, and new media like real people and places. New York, NY, US: Cambridge University Press, 1996, pp. xiv, 305.

[10] C. Nass and Y. Moon, “Machines and mindlessness: Social responses to computers,” J. Soc. Issues, vol. 56, no. 1, pp. 81–103, Jan. 2000, doi: 10/cqzrs6.

[11] K. D. Williams, “Dyads can be groups (and often are),” Small Group Res., vol. 41, no. 2, pp. 268–274, Apr. 2010, doi: 10/d6msv6.

[12] B.-T. Karsh, P. Waterson, and R. J. Holden, “Crossing levels in systems ergonomics: A framework to support ‘mesoergonomic’ inquiry,” Appl. Ergon., vol. 45, no. 1, pp. 45–54, Jan. 2014, doi: 10/ghzbs4.

[13] M. Poornikoo and K. I. Øvergård, “Model evaluation in human factors and ergonomics (HFE) sciences; case of trust in automation,” Theor. Issues Ergon. Sci., vol. 25, no. 4, pp. 416–452, Jul. 2024, doi: 10.1080/1463922X.2023.2233591.

[14] I. Schröder, O. Müller, H. Scholl, S. Levy-Tzedek, and P. Kellmeyer, “Can robots be trustworthy?” Ethik Med., vol. 35, no. 2, pp. 221–246, Jun. 2023, doi: 10/gs6bxz.

[15] J. A. Simpson, “Psychological foundations of trust,” Curr. Dir. Psychol. Sci., vol. 16, no. 5, pp. 264–268, 2007, doi: 10/cgg4k6.

[16] Venkatesh, Morris, Davis, and Davis, “User acceptance of information technology: Toward a unified view,” MIS Q., vol. 27, no. 3, p. 425, 2003, doi: 10/gc8zn2.

[17] M. E. Bratman, “Shared cooperative activity,” Philos. Rev., vol. 101, no. 2, pp. 327–341, 1992, doi: 10/cp95rd.

[18] N. Cila, “Designing human-agent collaborations: Commitment, responsiveness, and support,” in CHI Conference on Human Factors in Computing Systems, New Orleans LA USA: ACM, Apr. 2022, pp. 1–18. doi: 10/gp8d2c.

[19] A. Tsamados, L. Floridi, and M. Taddeo, “Human control of AI systems: From supervision to teaming,” AI Ethics, May 2024, doi: 10/gtzk8m.

[20] F. Santoni de Sio and J. van den Hoven, “Meaningful human control over autonomous systems: A philosophical account,” Front. Robot. AI, vol. 5, Feb. 2018, doi: 10/gf597h.

[21] A. K. Gardner, M. Kosemund, and J. Martinez, “Examining the feasibility and predictive validity of the SAGAT tool to assess situation awareness among medical trainees,” Simul. Healthc. J. Soc. Simul. Healthc., p. 1, Aug. 2016, doi: 10/ggqf2q.

[22] M. S. Crozier et al., “Use of human patient simulation and validation of the team situation awareness global assessment technique (TSAGAT): A multidisciplinary team assessment tool in trauma education,” J. Surg. Educ., vol. 72, no. 1, pp. 156–163, Jan. 2015, doi: 10/ggqfwk.

[23] J. Y. C. Chen, S. G. Lakhmani, K. Stowers, A. R. Selkowitz, J. L. Wright, and M. Barnes, “Situation awareness-based agent transparency and human-autonomy teaming effectiveness,” Theor. Issues Ergon. Sci., vol. 19, no. 3, pp. 259–282, May 2018, doi: 10/ggqft9.

[24] J. Wong, E. K. Chiou, R. Gutzwiller, M. Cook, and C. Fallon, “Human-artificial intelligence teaming for the U.S. Navy: Developing a holistic research roadmap,” presented at the ASPIRE, 2024.

[25] National Academies of Sciences, Engineering, and Medicine, Human-AI Teaming: State of the Art and Research Needs. Washington, D.C.: The National Academies Press, 2021. doi: 10.17226/26355.

[26] S. Robbins, “The many meanings of meaningful human control,” AI Ethics, Jul. 2023, doi: 10/g232vj.

[27] L. Cavalcante Siebert et al., “Meaningful human control: Actionable properties for AI system development,” AI Ethics, vol. 3, no. 1, pp. 241–255, Feb. 2023, doi: 10/gp6zkx.

[28] H. M. Roff and R. Moyes, “Meaningful human control, artificial intelligence and autonomous weapons,” in Briefing paper prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, Geneva, Apr. 2016. Accessed: Sep. 20, 2024. [Online]. Available: https://article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf