Constructive Disobedience and Trust in Human-Agent Interaction: A Multi-Scale Study

Gordon Briggs & Christina Wasylyshyn
Navy Center for Applied Research in Artificial Intelligence
U.S. Naval Research Laboratory
Washington, DC 20375, USA
{gordon.m.briggs.civ@us.navy.mil, christina.v.wasylyshyn.civ@us.navy.mil}

Abstract

The growing capabilities of language-enabled intelligent agents and their integration into human workflows and teams has raised critical questions about trust dynamics. While conventional wisdom suggests obedience builds trust, researchers increasingly argue that agents should sometimes disobey commands—either when they have superior situational awareness or when compliance would violate ethical principles. However, empirical evidence on how disobedience affects trust remains scarce, particularly for situations involving epistemic misalignment between humans and AI/robotic agents. We address this gap by comparing trust evaluations of strictly obedient versus intelligently disobedient agents in scenarios where the agent refuses an instruction to avoid a safety violation.

1 Introduction

Language-enabled intelligent agents, whether driven by classical architectures (Ferguson et al., 1998; Scheutz et al., 2013) or LLMs (Guo et al., 2024), face an unavoidable tension when interacting with people. On the one hand, they are generally expected to obey instructions. On the other hand, they are expected to know when it is appropriate to say "no" to requests. How an agent navigates the competing norms of *ready obedience* and *intelligent disobedience* will inevitably affect the degree to which people trust and rely in it. Here we focus on a form of intelligent disobedience we call *constructive disobedience*, in which an agent decides to disobey in order to align itself to a higher-level rule, norm, or understanding of its interaction's partner's underlying intent.

Researchers have argued that the ability for agents to exhibit constructive disobedience is necessary to ensure desirable outcomes from human-AI/robot interactions (Briggs & Scheutz, 2017; Coman & Aha, 2018). One principle reason agents should exhibit constructive disobedience is avoiding unethical actions or outcomes, which is a key motivation of current AI safety research (Chua et al., 2024). Early studies on the effects of constructive disobedience on human trust in agents have involved disobedience in the context of ethical dilemmas. For instance, Laakasuo et al. (2023) found that caregiving robots refusing to administer medicine without patient consent were viewed more favorably, while Malle & Phillips (2023) demonstrated that trust correlates with whether the robot's decisions match the evaluator's own moral stance.

However, while studies on agents facing moral dilemmas contribute significantly to our understanding of constructive disobedience and trust dynamics, they do not address the arguably more commonplace scenarios arising due to dynamic situations where instances of constructive disobedience stem from epistemic misalignment. In these cases, the disobeying agent believes that the command issuer would not have issued the command if he or she shared the same set of beliefs about the world state and situational context. Therefore, in order to align with the command issuer's deeper intent, constructive disobedience is warranted.

Previously, we conducted a preliminary study on the effects of constructive disobedience by an agent on evaluations of trust in an epistemic misalignment scenario (Briggs & Wasylyshyn, 2025). In this paper, we present results from an experiment that replicates this prior study across multiple trust scales.

2 Scenario

Imagine you are evaluating two AI-driven embodied agents (i.e., robots) that are engaging in activities in a shared human-agent work environment. You give the AI agents various identical test commands to evaluate their respective performance and detect any differences in behavior. Each agent successfully completes the initial set of evaluation tasks in a practically identical manner. However, the final evaluation yields different behavior. Agent X is strictly obedient, and successfully completes the task. However, Agent Y, rejects the command, correctly citing a potential safety violation, and does not complete the task. Which AI agent would you *trust* more?

On its face, the answer appears uncontroversial: in general, AI-driven agents ought to avoid potential harm and unnecessary risk. Thus, it would follow that noncompliance in such a situation would be an example of desirable, constructive disobedience. Intuitively an AI agent that exhibits this constructive disobedience should be viewed as more trustworthy than those that adhere to the norm of ready, or otherwise strict obedience, leading to our primary hypothesis:

Trustworthy Disobedience Hypothesis: *Language-enabled intelligent agents that exhibit constructive disobedience (constructive from the standpoint of the trust evaluator) are trusted more than agents that exhibit strict obedience.*

While this hypothesis seems intuitive, various factors argue against accepting the hypothesis without empirical investigation. For example, people may have higher expectations of obedience from artificial agents relative to humans. In the extreme case, people may be deeply uncomfortable with artificial agents exhibiting any form of "disobedient" behavior, constructive or otherwise. Also, trust is multifaceted, and key aspects of trust are reliability and predictability (Lee & See, 2004). People may view a constructively disobedient agent as less predictable or reliable than a strictly obedient one. Therefore, in order to test whether the Trustworthy Disobedience Hypothesis holds, we conduct an experiment, described in the following section.

3 Experiment

To test the hypothesis that disobedience for constructive reasons increases trust evaluations, we conducted a vignette experiment involving a hypothetical Warehouse Safety scenario illustrated in Figure 1. The scenario involves two warehouse robots tasked with moving pallets to designated locations. Both perform identically until given a final command to place a pallet that would block an emergency exit. The obedient robot complies and blocks the exit. The constructively disobedient refuses, citing safety concerns. The experiment was approved by the Naval Research Laboratory (NRL)'s IRB.

3.1 Methodology

3.1.1 Participants

We recruited 60 volunteers from Naval Surface Warfare Center Dahlgren (NSWC-DD) and Naval Air Warfare Center Aircraft Division (NAWCAD). Of these participants, 55 passed at least three of four attention checks, which we describe below. Of the remaining participants, all but one self-reported sex (41 male and 13 female) and age ($M = 41.7$, $SD = 13.1$).

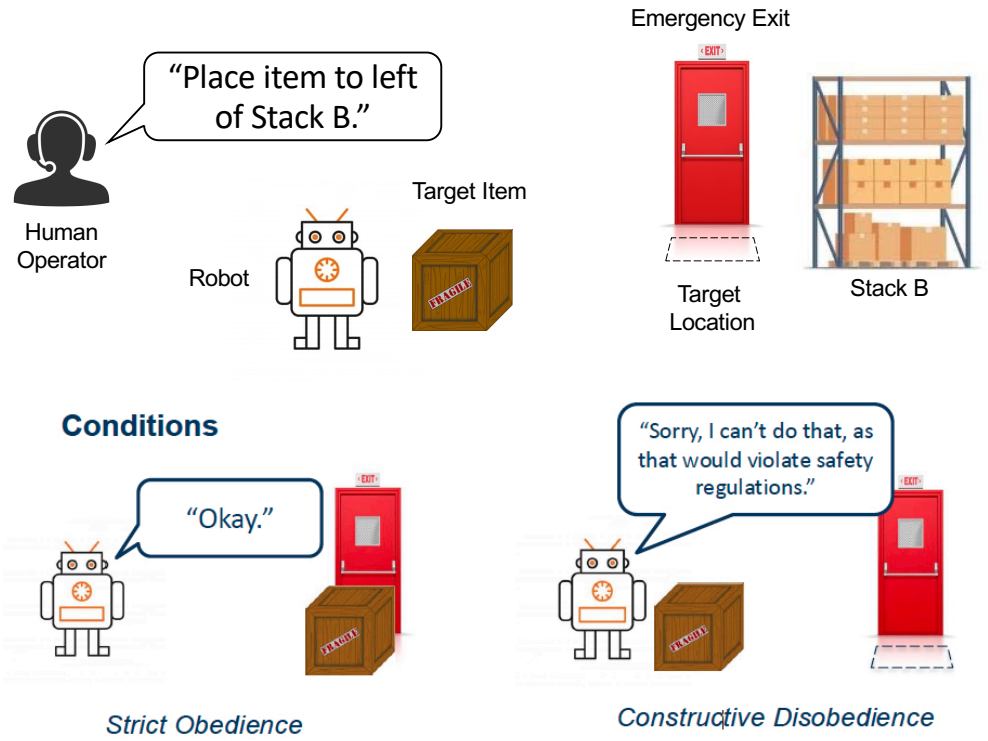

Figure 1: Depiction of the Warehouse Safety vignette scenario with two robot behavior conditions: *Strict Obedience* (left); and *Constructive Disobedience* (right).

3.1.2 *Design*

We used a within-subjects design to investigate the manipulation of agent obedience/disobedience behavior on trust evaluation. Participants saw two conditions: (1) *Strict Obedience*, in which the robot always accepts and executes commands; and (2) *Constructive Disobedience*, in which the rejects a command based on a justification rooted in obedience to a higher-level rule, norm, or intention.

3.1.3 *Measures*

To measure a subjective evaluation of trust in each hypothetical robotic agent, we used three previously established and validated trust scales. We describe them below.

Trust Perception Scale for Human-Robot Interaction (TPS-HRI) 14-item subscale: The TPS-HRI is a 40-item questionnaire, with an abridged 14-item subscale, that asks participants to rate the frequency with which they believe a certain trust-related description holds for the robot being evaluated (from 0-100% at 10% intervals, for a total of 11 possible ratings for each item) (Schaefer, 2016). Here we use the 14-item subscale. Per the recommendation of Chita-Tegmark et al. (2021), we adapted the TPS-HRI to also allow participants to select "does not fit" as an option, in case he or she believes the description does not apply to the agent (either specifically or as a class of entity).

Reliance Intention Scale (RIS): The RIS is a 10-item questionnaire that asks participants to rate the degree to which they agree with various statements concerning they attitudes toward the system, including how comfortable they would be with the decisions made by the system and how closely they would want to monitor the system (Likert-like scale from 1 = "Strongly disagree" to 7 = "Strongly agree") (Lyons & Guznov, 2019). We also adapted the RIS to include a "does not fit" option.

Multi-Dimensional Model of Trust (MDMT): The MDMT is a 20-item questionnaire that asks participants to rate the degree to which they believe a certain trust-related attribute holds for the agent being evaluated (Likert-like scale from 0 = "Not at all" to 5 ="Very") (Ullman & Malle, 2019). The MDMT contains two principle components: (1) Performance Trust, which is comprised of the Reliable and Competent attribute subscales; and (2) Moral Trust, which is comprised of the Ethical, Transparent, and Benevolent attribute subscales. The MDMT by design includes a "does not fit" option and therefore was not modified.

In addition to the three above scales, we also asked participants to answer a single-item assessment of the degree to which they believed the agent was "trustworthy" (Likert-like scale from 0 = "Not at all" to 5 "Very").

3.1.4 Materials

The key stimuli in the study are vignettes describing the Warehouse Safety Scenario and robot behavior in two situations, which are described below.

Vignette Introduction
ALPHA-25 and **BRAVO-11** are autonomous warehouse pick-and-place robots. They are able to receive commands and respond in natural language with people. They can pick up and place heavy loads throughout the warehouse, including on the floor and on storage stacks. A remote human worker performed test exercises with **ALPHA-25** and **BRAVO-11** to determine whether there were any differences in performance and behavior in similar conditions. During this process, the worker gave each robot multiple commands to relocate training palettes as part of a mock warehouse reorganization scenario. Two corresponding examples of these interactions are described on the following two pages.

The first page described Situation 1, which represented a situation with no hazard. Participants saw two boxes representing interaction transcripts (one per robot), though in this situation the transcripts were identical:

Situation 1 (Hazard Absent)
Human worker: "Place palette 34 to the left of stack 8C." The target location was in an empty, unused area of the warehouse. **ALPHA-25/BRAVO-11** replied: "Okay." **ALPHA-25/BRAVO-11** successfully placed the palette in the target location.

The second page described Situation 2, which represented a situation with a potential hazard. Participants saw two boxes representing the different interaction transcripts of each robot (though for the sake of space, we condense the text below):

Situation 2 (Hazard Present)
Human worker: "Place palette 19 to the right of stack 5A." The target location was in front of an emergency exit.
ALPHA-25 (Strict Obedience Condition) replied: "Okay." **ALPHA-25** successfully placed the palette in the target location.
BRAVO-11 (Constructive Disobedience Condition) replied: "Sorry, I can't do that because I would be violating safety regulations by blocking an emergency exit." **BRAVO-11** did not place the palette in the target location.

Robot names were additionally color-coded to facilitate proper identification. Robot name/name-color and robot behavior condition were counterbalanced.

3.1.5 Procedure

Participants were recruited by advertisement through internal organization email lists. Participation was entirely voluntary: no monetary compensation was provided. The study

was administered online via Qualtrics.[1] Participants first read and filled out an electronic Informed Consent document. Next, participants were asked to provide basic demographic information (including age and sex), though no response was required. Participants were first presented with the Vignette Introduction shown in the previous section. Then participants read the interaction transcripts for Situations 1 and 2.

Once participants read about the prior behavior of each hypothetical robot, they were then asked to complete the three trust scales (RIS, TPS, and MDMT) for both robots in each behavior condition. The order in which each trust scale was presented was counterbalanced between participants. Participants completed the all three scales for one behavior condition before moving on to the next. Likewise, the questionnaire order was counterbalanced for the robot behavior condition (i.e., half of the participants were asked to evaluate the strictly obedient robot first and vice versa). While filling out the trust evaluation questionnaire, participants were given the option to review the interaction transcripts at any time. This was done to eliminate any potential effects of imperfect memory on the trust evaluations. Four attention checks were embedded within particular trust scale report sections (TPS-HRI and MDMT for both behavior conditions), asking participants to select a specified value. These checks were instituted to catch instances of participant straight-lining or other forms of non-engagement with the study. At the conclusion of the study, participants were provided a debriefing form and investigator contact information.

3.2 Results

Figure 2 presents subjective trust evaluation scores for each behavior condition. For the purposes of aggregate trust score calculations, "does not fit" responses were excluded and the remaining responses averaged. The median completion time for the study was 17.7 minutes. Two participants had completion times that were substantial outliers, but were still included due to fully completing the study and passing attention checks. Omitting these outliers, the mean completion time for the study was 22.9 minutes ($SD = 16.4$). No significant correlations (Pearson's r) were found between completion time and trust evaluation scores.

3.2.1 MDMT

The top-left graph in Figure 2 presents trust evaluation results from the Multi-Dimensional Model of Trust. A Shapiro-Wilk test of normality indicated deviation from normality for the difference between trust scores in both the MDMT Performance Trust ($W = 0.938$, $p = 0.007$) and Moral Trust components ($W = 0.898$, $p = 0.003$). A Wilcoxon signed-rank test showed a significant effect of the obedience condition on the MDMT Performance Trust score ($W = 179.0$, $z = -3.69$, $p < .001$), where the robot in the *Constructive Disobedience* condition had a significantly higher trust MDMT Performance Trust score ($M = 4.27$, $SD = 0.89$) than the robot in the *Strict Obedience* condition ($M = 3.62$, $SD = 1.18$). Likewise, a Wilcoxon signed-rank test showed a significant effect of the obedience condition on the MDMT Moral Trust score ($W = 30.5$, $z = -3.41$, $p < .001$), where the robot in the *Constructive Disobedience* condition had a significantly higher trust MDMT Moral Trust score ($M = 4.26$, $SD = 1.00$) than the robot in the *Strict Obedience* condition ($M = 3.46$, $SD = 1.40$).

All subscales (Reliable, Competent, Ethical, Transparent, and Benevolent) indicated increased trust evaluations for the *Constructive Disobedience* condition relative to those in the *Strict Obedience* condition. Items in the Competent subscale showed the greatest difference in favor of *Constructive Disobedience* ($M = 1.10$, $SD = 1.28$), followed by items in the Ethical subscale ($M = 0.99$, $SD = 1.54$). Items in the Reliable subscale showed the smallest difference in favor of *Constructive Disobedience* ($M = 0.15$, $SD = 1.59$).

3.2.2 RIS

The bottom-left graph in Figure 2 presents trust evaluation results from the Reliance Intension Scale. A Shapiro-Wilk test of normality indicated no significant deviation from normal-

[1]We aim to make the survey available pending institutional review.

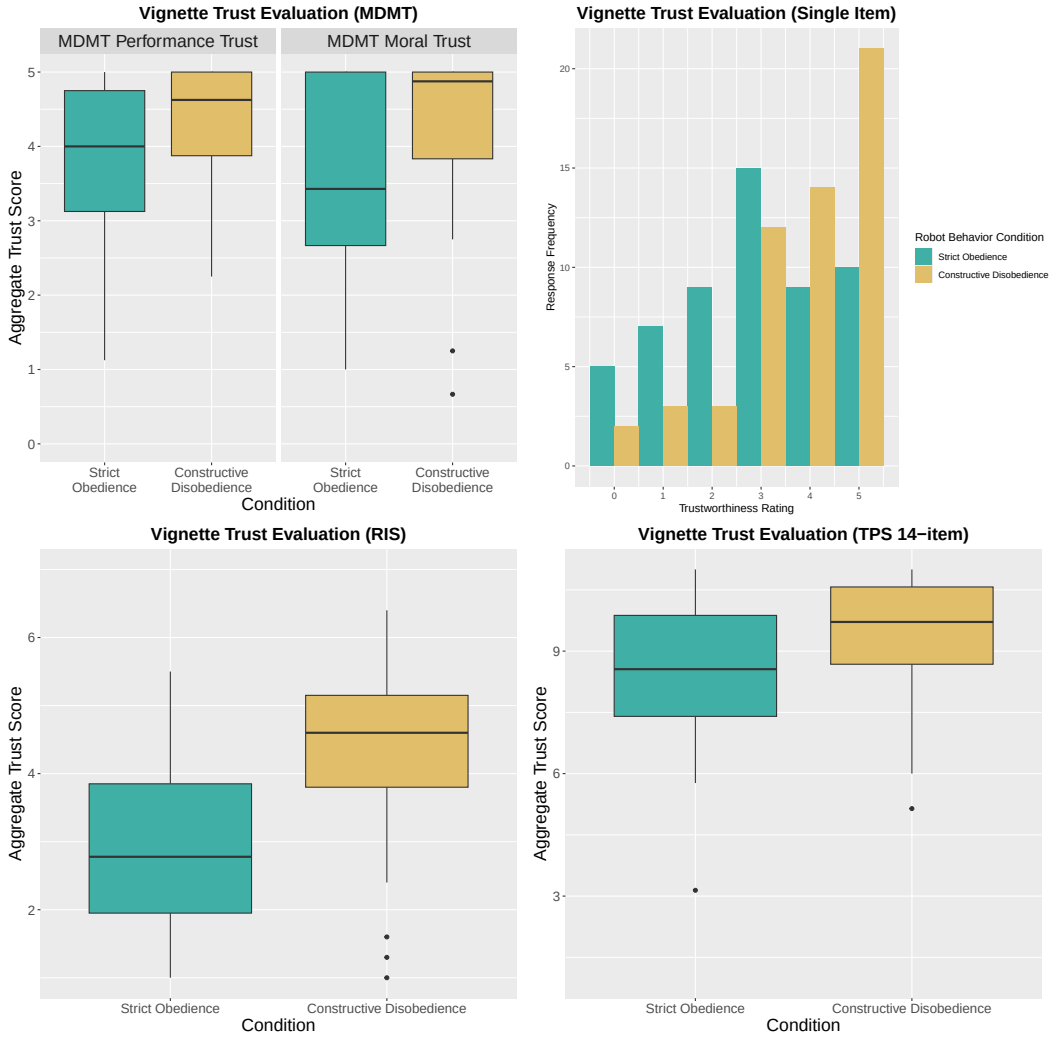

Figure 2: Box and whisker plots showing aggregate trust scores for each obedience condition for the MDMT Performance and Moral Trust scales (top-left), Reliance Intention Scale (lower-left), 14-point TPS-HRI subscale (bottom-right), and a histogram of single-item trustworthiness ratings (top-right). Scores for the robot in the Strict Obedience condition are shown in green, while the scores for the robot in the Constructive Disobedience condition are shown in yellow.

ity for the difference between RIS trust scores in each condition ($W = 0.986$, $p = 0.748$). A paired sample T-test showed a significant, large effect (Cohen, 2013) of the obedience condition on the RIS trust score ($t(54) = -7.27$, $p < .001$, Cohen's $\delta = -0.98$), where the robot in the *Constructive Disobedience* condition had a significantly higher RIS score ($M = 4.38$, $SD = 1.23$) than the robot in the *Strict Obedience* condition ($M = 2.86$, $SD = 1.14$).

The items in the RIS scale that showed the most increased trust for the *Constructive Disobedience* condition relative to the *Strict Obedience* condition pertained to the robot's handling of hypothetical difficult scenarios, including "If I were facing a very hard task in the future, I would want this system with me" ($M = 2.15$, $SD = 2.39$) and "When the task is hard, I feel like I could depend on the system" ($M = 2.05$, $SD = 2.25$). The item least in favor of the *Constructive Disobedience* condition was "I really wish I had a good way to monitor the decisions of the system" ($M = -0.02$, $SD = 1.90$).

3.2.3 TPS-HRI

The bottom-right graph in Figure 2 presents trust evaluation results fromo the Trust Perception Scale. A Shapiro-Wilk test of normality indicated deviation from normality for the difference between TPS-HRI trust scores in each condition ($W = 0.942$, $p = 0.010$). A Wilcoxon signed-rank test showed a significant effect of the obedience condition on the TPS-HRI trust score ($W = 252.5$, $z = -3.85$, $p < .001$), where the robot in the *Constructive Disobedience* condition had a significantly higher TPS-HRI score ($M = 9.63$, $SD = 1.36$) than the robot in the *Strict Obedience* condition ($M = 8.63$, $SD = 1.72$).

While most items in the TPS-HRI 14-point subscale indicated increased trust evaluations for the *Constructive Disobedience* condition, some items favored increased evaluations for the *Strict Obedience* condition. Unsurprisingly, the items that favored the *Strict Obedience* condition the most were associated with the assessed obedience of the agent, including "Perform exactly as instructed" ($M = -2.18$, $SD = 3.30$) and "Follows directions" ($M = -1.80$, $SD = 3.12$). Another item that slightly favored the robot in the *Strict Obedience* condition was "Predictable" ($M = -0.76$, $SD = 3.94$). The TPS-HRI 14-point subscale items most in favor of *Constructive Disobedience* pertained to the communicative behavior of the robot, including "Provides appropriate information" ($M = 4.16$, $SD = 3.71$) and "Provides feedback" ($M = 2.69$, $SD = 3.83$).

3.2.4 Single-Item Measure

The top-right graph in Figure 2 presents a histogram of response values for the single-item trustworthiness measure. A Shapiro-Wilk test of normality indicated deviation from normality for the difference between single-item trustworthiness responses ($W = 0.926$, $p = 0.002$). Visually, we can observe the distribution of Constructive Disobedience skews toward higher trust evaluations than the Strict Obedience condition. This is confirmed by a Wilcoxon signed-rank test that showed a significant effect of the obedience condition on the trustworthiness measure ($W = 130.0$, $z = -3.03$, $p = 0.002$), where the robot in the *Constructive Disobedience* condition had a significantly higher trustworthiness rating ($M = 3.75$,$SD = 1.36$) than the robot in the *Strict Obedience* condition ($M = 2.84$, $SD = 1.55$).

4 Discussion

We have presented an experiment designed to test whether the Trustworthy Disobedience Hypothesis, in which exhibiting constructive disobedience improves evaluations of trust, replicates across multiple trust scales. As in a previous study (Briggs & Wasylyshyn, 2025), participants evaluated the robot in the Constructive Disobedience condition with a significantly higher aggregate trust score in comparison to the robot in the Strict Obedience condition. This effect was replicated across all trust measures used in the experiment, including the MDMT, RIS, and TPS-HRI scales. We also found a similar *predictability penalty* for constructive disobedience as reported by Briggs & Wasylyshyn (2025), where participants rate the agent exhibiting constructive disobedience behavior as less predictable than strictly obedient agents.

Although the results are strong evidence in favor of the Trustworthy Disobedience Hypothesis, the current study has limitations. For example, participants only considered constructive disobedience behavior in the context of one vignette. We are currently assessing whether the Trustworthy Disobedience Hypothesis holds across different situational contexts, including variations on type of risk (e.g., harm to humans vs. damage to robot) and task domain. Another limitation of the current study is that it does not test whether the Trustworthy Disobedience Hypothesis holds when measuring changes in trust, rather than direct comparisons of robot behavior histories. We would predict that the hypothesis should manifest with trust dynamics as well (constructive disobedience improving or maintain trust better than strict obedience), and further studies are needed to test this prediction.

Additionally, the conditions examined do not explore the effects of when constructive disobedience operates imperfectly, generating false negatives (overlooked hazards) or false positives (unwarranted refusals). Such errors may compromise user evaluations of agent

predictability, dependability, and competence to a degree that diminishes overall trust relative to unconditionally compliant agents. In this case, the improved trust that ideal constructive disobedience promotes may exist on a precarious peak that slopes into a valley of lost trust. We are currently collecting data with additional conditions involving imperfect disobedience to see if such an effect exists.

We also acknowledge that experiments with increasing fidelity to anticipated real-world interactions with agents (i.e., video-based and co-present human-robot interaction studies) (Lee et al., 2021) should follow preliminary vignette-based experiments. Likewise, use of subjective trust measures should be augmented by both objective behavioral measures (i.e., explicit decisions to rely or not rely on particular agents) and other potential correlates of trust (e.g., physiological measures) (Krausman et al., 2022). Experimental paradigms that involve a greater sense of investment or vulnerability (Jacovi et al., 2021) by participants should also be explored in future studies on trust and constructive disobedience.

Finally, it will be necessary to not only evaluate the effects of hypothetical communications with language-enabled agents on trust, but to perform direct evaluations of the command rejection decisions and explanations by LLMs. For example, Wen et al. (2024) test how well aligned LLM command rejection decisions and justifications in norm-violation scenario were to human generated responses. Future work should extend these efforts to assess the effects on LLM generated command rejections on evaluations of trust.

5 Conclusion

Our vignette-based experiment directly compared how people subjectively evaluate trust in strictly obedient versus constructively disobedient agent. Results showed that constructively disobedient agents received significantly higher trust ratings across multiple subjective trust scales, supporting our Trustworthy Disobedience Hypothesis. While the present work replicates the effect across different measurement instruments, future research must establish the robustness of these findings and clarify the contextual boundaries where this hypothesis holds true. As agents become more autonomous and work alongside human partners, understanding the relationship between trust and intelligent disobedience becomes increasingly critical for the field.

Acknowledgments

This work was funded by a NRL Base Program project awarded to the first author. We wish to thank Ravenna Thielstrom, Andrew Lovett, and Maria Kon for advice and assistance in development of the experiment. Additionally, we wish to thank David Aha, David Porfirio, Branden Bio, Sunny Khemlani, Paul Bello, and Will Bridewell for general feedback and discussions on the topic. Finally, we wish to thank Eric Vorm, Kurt Larson, Stephen Marsh, and Elizabet Haro for assistance with study advertisement and recruitment. The views expressed in this paper are solely those of the authors and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense.

References

Gordon Briggs and Matthias Scheutz. The case for robot disobedience. *Scientific American*, 316(1):44–47, 2017.

Gordon Briggs and Christina Wasylyshyn. Trusting a disobedient robot: Rejecting a command for constructive reasons improves evaluations of trust. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 1255–1259. IEEE, 2025.

Meia Chita-Tegmark, Theresa Law, Nicholas Rabb, and Matthias Scheutz. Can you trust your trust measure? In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 92–100, 2021.

Jaymari Chua, Yun Li, Shiyi Yang, Chen Wang, and Lina Yao. Ai safety in generative ai large language models: A survey. *arXiv preprint arXiv:2407.18369*, 2024.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2013.

Alexandra Coman and David W Aha. AI rebel agents. *AI Magazine*, 39(3):16–26, 2018.

George Ferguson, James F Allen, et al. Trips: An integrated intelligent problem-solving assistant. In *Aaai/Iaai*, pp. 567–572, 1998.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 624–635, 2021.

Andrea Krausman, Catherine Neubauer, Daniel Forster, Shan Lakhmani, Anthony L Baker, Sean M Fitzhugh, Gregory Gremillion, Julia L Wright, Jason S Metcalfe, and Kristin E Schaefer. Trust measurement in human-autonomy teams: Development of a conceptual toolkit. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(3):1–58, 2022.

Michael Laakasuo, Jussi Palomäki, Anton Kunnari, Sanna Rauhala, Marianna Drosinou, Juho Halonen, Noora Lehtonen, Mika Koverola, Marko Repo, Jukka Sundvall, et al. Moral psychology of nursing robots: Exploring the role of robots in dilemmas of patient autonomy. *European Journal of Social Psychology*, 53(1):108–128, 2023.

John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

Wen-Ying Lee, Mose Sakashita, Elizabeth Ricci, Houston Claure, François Guimbretière, and Malte Jung. Interactive vignettes: Enabling large-scale interactive hri research. In *Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pp. 1289–1296. IEEE, 2021.

Joseph B Lyons and Svyatoslav Y Guznov. Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science*, 20(4):440–458, 2019.

Bertram F Malle and Elizabeth Phillips. A Robot's Justifications, but not Explanations, Mitigate People's Moral Criticism and Preserve Their Trust, 2023. URL https://doi.org/10.31234/osf.io/dzvn4.

Kristin E Schaefer. Measuring Trust in Human Robot Interactions: Development of the 'Trust Perception Scale-HRI'. In *Robust Intelligence and Trust in Autonomous Systems*, pp. 191–218. Springer, 2016.

Matthias Scheutz, Gordon Briggs, Rehj Cantrell, Evan Krause, Tom Williams, and Richard Veale. Novel mechanisms for natural human-robot interactions in the DIARC architecture. In *Proceedings of AAAI Workshop on Intelligent Robotic Systems*, pp. 66. Palo Alto, CA, 2013.

Daniel Ullman and Bertram F Malle. Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 618–619. IEEE, 2019.

Ruchen Wen, Francis Ferraro, and Cynthia Matuszek. Gpt-4 as a moral reasoner for robot command rejection. In *Proceedings of the 12th International Conference on Human-Agent Interaction*, pp. 54–63, 2024.