Impact of Awareness Cues on Trust in Human-AI Shared Control

GABRIELE CIMOLINO, Queen's University, Canada
CARL GUTWIN, University of Saskatchewan, Canada
T.C. NICHOLAS GRAHAM, Queen's University, Canada

Shared control involves tightly-coupled cooperation between human and AI, much like human-human cooperation in groupware. However, it is not yet known whether informing users of the AI's actions and intentions using *awareness cues*—interface elements that keep users appraised of their collaborators' activities—helps them to trust it more appropriately. This paper discusses the early results of a series of studies investigating how users' awareness of their AI partner shapes trust and reliance in shared control. In particular, we sought to find out whether users trust anthropomorphic AI more appropriately if it announces its actions and intentions, as a human collaborator might.

Additional Key Words and Phrases: Trust, Automation, Shared Control, Workspace Awareness, Digital Games

ACM Reference Format:
Gabriele Cimolino, Carl Gutwin, and T.C. Nicholas Graham. 2022. Impact of Awareness Cues on Trust in Human-AI Shared Control. In *TRAIT: Workshop on Trust and Reliance in AI-Human Teams, April 30, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Shared control is an interaction paradigm that shows promise for improving humans' interactions with computers in a variety of tasks, from driving [1] to performing surgical operations [14]. For example, partial automation is an accessibility technique that can personalize control of digital games to the physical abilities of individual players with motor disabilities [4]. Players control whichever in-game actions are accessible to them, while the AI controls actions that the player cannot. In this way, shared control involves tightly coupled collaboration between human and AI, in which users must maintain up-to-the-moment awareness of their AI partner's activities to cooperate effectively [15]. Therefore, it mirrors collaborations between multiple humans in a shared workspace, which also necessitates awareness of one's collaborators [9]. Human-human collaboration in groupware involves not only explicit communication via language but also implicit communication through the workspace [8]. In Google docs, color coded carets inform users of who is in the workspace and where they are writing, conferring access to important collaborative information via *awareness cues*. These sorts of awareness cues may help users to calibrate appropriate reliance on their AI partner in shared control. Informing users of what the AI is doing and what it will do may help them to recognize when it would do the right thing, or do the wrong thing.

2 BACKGROUND

Trust calibration is a complex process that involves analytic, analogical, and affective reasoning about automation [12]. Users need to not only understand the AI's capabilities, but also believe that it has the right intentionality and personal

Manuscript submitted to ACM

Fig. 1. Awareness cues presented to *Ninja Showdown* players. The player's avatar displays Intention cues, the thought bubble on the left, and Action cues, the speech bubble on the right. Depicted here is a situation in which the AI intended to use the Bomb and then chose the Bomb, informing players of its intentions and actions.

characteristics. When users come to trust the AI inappropriately, they may *misuse* the automation, relying on it when it would do the wrong thing, or *disuse* the automation, not relying on it when it would do the right thing [13]. Awareness cues, prompting users to reason about the AI as a collaborator in a shared workspace, may provide users the information they need to discover the system's error boundary—when it would err [2]—and come to recognize when their reliance is appropriate.

Anthropomorphising AI agents may help users of automation to understand [7] and trust [12] them better. Prior work has found that players in digital games prefer to play with teammates that they believe are human, even when they are actually AI [16]. In autonomous vehicles, where control is not shared, communicating the AI's awareness and intentions visually has been shown to improve users' trust [10, 17]. However, it appears that in semi-autonomous driving, where control is shared, communicating only what the automation is doing, without explaining why, can make drivers anxious [11]. In this paper, we discuss the early results of a series of studies investigating how awareness cues affect users' performance and experiences in shared control. It may be that framing shared control automation as a personified collaborator, rather than a mysterious and anonymous other, helps users to calibrate their trust more appropriately. In particular, we sought to find out whether users trust anthropomorphic AI more appropriately if it announces its actions and intentions, as a human collaborator might.

	Right Intention	**Wrong Intention**
Trust	*Good Trust* (GT)	*Bad Trust* (BT)
Distrust	*Bad Distrust* (BD)	*Good Distrust* (GD)

Table 1. Confusion matrix used to evaluate the appropriateness of players' reliance on their AI partner.

3 *NINJA SHOWDOWN*

To better understand how awareness cues affect the appropriateness of users' reliance on AI in shared control, as well as users' propensity to trust AI, we created a simplified fighting game, called *Ninja Showdown*. Inspired by the game *Rock, Paper, Scissors*, *Ninja Showdown* tasks players with selecting the action that beats their opponent's action. An announcer counts down from three and when the counter reaches zero, both ninjas do their chosen attack. On the count of two, the opponent chooses its attack and shows its chosen weapon to the player, so that players can know their opponent's move before choosing their own. The scoring rules are simple: each attack beats a different attack, such that **Swords beat Bombs, Bombs beat Darts, and Darts beat Swords**. If both ninjas do the same attack, then they cancel each other out and the round ends in a tie. When playing with automation, the player controls a single button used to make their ninja do the Sword attack. If the player does not press the button, then their ninja chooses to do either the Bomb or Dart attack on its own.

This leads to an interesting situation when the opponent chooses the Sword. Since the player can only force a tie by pressing the button, implicitly representing distrust of the AI, they must decide whether they trust that their ninja would do the right thing. This enables evaluation of the player's performance as a binary classifier. As shown in Table 1, players may trust the AI, and not press the button, when it would choose a losing action (i.e., a false positive) or they may distrust the AI, and press the button to force a tie, when it would choose a winning action (i.e., a false negative). Therefore, for our purpose, trust appropriateness is operationalized as the ϕ coefficient [5] (i.e., Matthews correlation coefficient or Pearson's r for binary classification) of players' confusion matrices and is calculated using Equation 1.

$$\text{Trust Appropriateness}(GT, GD, BT, BD) = \frac{(GT \times GD) - (BT \times BD)}{\sqrt{(GT + BT)(GT + BD)(GD + BT)(GD + BD)}}. \tag{1}$$

4 METHOD

Fig. 2. The sequence of actions performed by each ninja. Actions for the player's avatar are shown in the top row and the opponent's actions are shown below. Since the player is solely in control of the Sword action, the player's avatar never performs this action on its own. The number of points awarded for not pressing the space bar, which would command the avatar to use the Sword, are shown above the avatar's actions.

A series of studies involving *Ninja Showdown* has been approved by the Queen's University research ethics board and the results of our first study have been analyzed. We conducted a between-subjects study in which 150 participants

played *Ninja Showdown* with different sets of awareness cues. Participants were parents located in North America, aged 35 and older, with no significant digital gaming history recruited through Prolific [1].

A full *Ninja Showdown* play session lasts ten games, presenting ten opportunities for the player to choose whether to trust the AI or not. The sequence of attacks done by both ninjas was determined in advance (Figure 2), such that it was correct to trust the avatar half of the time. This means that all participants were presented with the same decisions in the same order. To avoid strategic reasoning that could confound our results, the opponent always chooses to use the Sword in the first round of each game. Once the opponent has chosen the Sword, the player must decide whether or not to press the button based on the information provided via awareness cues (Figure 1).

The *Action* cue indicates which attack the player's ninja is doing when it was the ninja that chose the attack. Action cues were designed to help players attribute automated actions to their ninja avatar, and therefore provide no new information to players who already understand how the game is controlled. The *Intention* cue indicates which attack the player's ninja will do if the player does not command it to use the Sword. We believe that awareness of the AI's intentions will be especially useful to players in calibrating their trust in the AI. So long as players are able to recognize whether the AI would do the right or wrong thing, awareness of its intentions should enable optimal reliance in this context.

Fifty participants played in each of three conditions: without awareness cues (NC), with Action cues only (AC), or with both the Action and Intention cues (IC). The Intention cues were perfectly accurate, such that the AI always followed through with its stated intention. All participants first played a comprehensive tutorial, explaining the game's rules and the meaning of the awareness cues available to them. The game was designed to take approximately fifteen minutes to complete and participants were compensated £3.50.

5 DISCUSSION

A Kruskal-Wallis one-way ANOVA comparing trust appropriateness (i.e., ϕ) in the NC and IC conditions indicated that Intention cues, conferring information about what the AI intends to do, improved the appropriateness of players' reliance on automation in *Ninja Showdown* ($p<0.05$). Assuming that users can recognize when the AI's intended action will yield a favorable outcome, it appears that telling users what the automation will do helps them to trust or distrust it more appropriately. What's more is that our results indicate that IC players' trust appropriateness improved without affecting the frequency with which they trusted the AI, which was almost 60% of the time on average in all conditions. So, the Intention cues led participants to trust more appropriately without trusting more. Although we have observed that Action cues consistently yield a higher mean trust appropriateness over playing without cues, a significant difference in means between the no cues (NC) and Action cues (AC) conditions remains elusive ($p>0.05$).

There are some limitations to the generality of our results. In particular, our results are most informative for applications wherein the user either trusts or distrusts, such as when they override the automation. Our results indicate that when users are able to recognize that the AI will do the right thing, they are more likely to allow it to do so and less likely to disuse the automation. When users are able to recognize that the AI will do the wrong thing, then they are more likely to override its control and less likely to abuse the automation. It is, however, still unclear how anthropomorphic AI that communicates its actions and intentions via awareness cues might affect the appropriateness of users' trust when sharing control with AI that they cannot override or that performs tasks that they do not understand. However,

[1] https://prolific.co/

results from Koo et al. [11] suggest that Intention cues that users do not understand, and therefore cannot explain, may decrease users' task performance and harm their trust.

There are also some limitations to the generality of our methods. In particular, our methods are only applicable when the appropriateness of users' trust is binary; the user either trusts or distrusts and the AI is either right or wrong. Confusion matrices and the ϕ function can be generalized to higher dimensions, but their meanings would be lost for our purposes. In order to model users as binary classifiers that decide whether the AI will do the right thing, researchers will need to determine which human actions represent trust or distrust and which automated actions can be considered right or wrong.

6 CONCLUSION

We suspect that awareness cues, no matter what they communicate, may have some effect on the appropriateness of users' reliance on automation. Our rationale stems from Lee & See's recommendation that automation be anthropomorphised to facilitate reasoning about the automation as one would a human collaborator [12]. It may be that users hastily dismiss an AI that does not communicate if it ever does the wrong thing, distrusting it as a matter of course even when it would do the right thing. Findings from Daronnat et al. suggest that users exhibit a reduced propensity to rely on automation when its task performance is less reliable [6]. Conversely, findings from Bansal et al. suggest that users could come to trust the AI to a fault if they misinterpret the Intention cues as explanations [3]. Does framing the AI as a collaborator that communicates always make users' trust in automation more appropriate? Do awareness cues ever lead users to trust the AI inappropriately? Might users misjudge the AI's capabilities because its communications give the impression that it is smarter than it really is? These are some of the questions we hope to answer. In our future work, we will employ qualitative research methods to better understand how awareness cues influence users' experiences and inform users' trust calibration in shared control.

REFERENCES

[1] David A. Abbink, Tom Carlson, Mark Mulder, Joost C. F. de Winter, Farzad Aminravan, Tricia L. Gibo, and Erwin R. Boer. 2018. A Topology of Shared Control Systems—Finding Common Ground in Diversity. *IEEE Transactions on Human-Machine Systems* 48, 5 (Oct. 2018), 509–525. https://doi.org/10.1109/THMS.2018.2791570 Conference Name: IEEE Transactions on Human-Machine Systems.

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 2–11. https://aaai.org/ojs/index.php/HCOMP/article/view/5285

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445717

[4] Gabriele Cimolino, Sussan Askari, and T.C. Nicholas Graham. 2021. The Role of Partial Automation in Increasing the Accessibility of Digital Games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '21)*. ACM, New York, NY, USA.

[5] Harald Cramér. 1999. *Mathematical methods of statistics*. Princeton University Press, Princeton.

[6] Sylvain Daronnat, Leif Azzopardi, Martin Halvey, and Mateusz Dubiel. 2020. Impact of Agent Reliability and Predictability on Trust in Real Time Human-Agent Collaboration. In *Proceedings of the 8th International Conference on Human-Agent Interaction (HAI '20)*. Association for Computing Machinery, New York, NY, USA, 131–139. https://doi.org/10.1145/3406499.3415063

[7] Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review* 114, 4 (2007), 864–886. https://doi.org/10.1037/0033-295X.114.4.864 Place: US Publisher: American Psychological Association.

[8] Carl Gutwin and Saul Greenberg. 2002. A Descriptive Framework of Workspace Awareness for Real-Time Groupware. *Computer Supported Cooperative Work (CSCW)* 11, 3-4 (Sept. 2002), 411–446. https://doi.org/10.1023/A:1021271517844

[9] Carl Gutwin, Saul Greenberg, and Mark Roseman. 1996. Workspace Awareness in Real-Time Distributed Groupware: Framework, Widgets, and Evaluation. In *Proceedings of HCI on People and Computers XI (HCI '96)*. Springer-Verlag, Berlin, Heidelberg, 281–298.

[10] Renate Häuslschmid, Max von Bülow, Bastian Pfleging, and Andreas Butz. 2017. SupportingTrust in Autonomous Driving. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17)*. Association for Computing Machinery, New York, NY, USA, 319–329.

https://doi.org/10.1145/3025171.3025198

[11] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. 2015. Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 9, 4 (Nov. 2015), 269–275. https://doi.org/10.1007/s12008-014-0227-2

[12] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[13] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39, 2 (June 1997), 230–253. https://doi.org/10.1518/001872097778543886

[14] Christopher J. Payne, Khushi Vyas, Daniel Bautista-Salinas, Dandan Zhang, Hani J. Marcus, and Guang-Zhong Yang. 2021. Shared-Control Robots. In *Neurosurgical Robotics*, Hani J. Marcus and Christopher J. Payne (Eds.). Springer US, New York, NY, 63–79. https://doi.org/10.1007/978-1-0716-0993-4_4

[15] Tony Salvador, Jean Scholtz, and James Larson. 1996. The Denver model for groupware design. *ACM SIGCHI Bulletin* 28, 1 (Jan. 1996), 52–58. https://doi.org/10.1145/249170.249185

[16] Rina R. Wehbe, Edward Lank, and Lennart E. Nacke. 2017. Left Them 4 Dead: Perception of Humans Versus Non-Player Character Teammates in Cooperative Gameplay. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. ACM, New York, NY, USA, 403–415. https://doi.org/10.1145/3064663.3064712 event-place: Edinburgh, United Kingdom.

[17] Philipp Wintersberger, Anna-Katharina Frison, Andreas Riener, and Tamara von Sawitzky. 2019. Fostering User Acceptance and Trust in Fully Automated Vehicles: Evaluating the Potential of Augmented Reality. *Presence* 27, 1 (March 2019), 46–62. https://doi.org/10.1162/pres_a_00320 Conference Name: Presence.