Human-Agent Collaborations: Trust in Negotiating Control

Sylvain Daronnat
University of Strathclyde
Glasgow, UK
sylvain.daronnat@strath.ac.uk

Leif Azzopardi
University of Strathclyde
Glasgow, UK
leif.azzopardi@strath.ac.uk

Martin Halvey
University of Strathclyde
Glasgow, UK
martin.halvey@strath.ac.uk

Mateusz Dubiel
University of Strathclyde
Glasgow, UK
mateusz.dubiel@strath.ac.uk

ABSTRACT

For human-agent collaborations to prosper, end-users need to trust the agent(s) they interact with. This is especially important in scenarios where the users and agents negotiate control in order to achieve objectives in real time (e.g. from helping surgeons with precision tasks to parking a semi-autonomous car or completing objectives in a video-game, etc.). Too much trust, and the user may overly rely on the agent. Insufficient trust, and the user may not adequately utilise the agent. In addition, measuring trust and trust-worthiness is difficult and presents a number of challenges. In this paper, we discuss current approaches to measuring trust, and explain how they can be inadequate in a real time setting where it is critical to know the extent to which the user currently trusts the agent. We then describe our attempts at quantifying the relationship between trust, performance and control.

Workshop proceedings *Everyday Automation Experience'19*
In conjunction with CHI'19, May 5th, 2019, Glasgow, UK
Website: http://everyday-automation.tech-experience.at

KEYWORDS
HCI; Trust; Game; Collaborative Agent; Human Factors; Performance; Control; Objective Measurements; Subjective Measurements

INTRODUCTION

Artificially intelligent agents are becoming more pervasive throughout society. As agents become more "intelligent" they will also need to become increasingly interactive and collaborative, supporting an array of tasks. For example in air traffic control, such agents are planned to help controllers with airspace management [19]. Already in other domains, such as online learning, learners are supported by conversational agents [8].

Human-Agent partnerships are influenced by a large number of components. Among these components, a key factor contributing to effective human-agent partnerships is trust. Indeed, trust has been shown to be positively tied to performance in psychology studies [1]. Lee and See [9] define Trust as "the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability. [...] An agent can either be an automated system or another person that actively interacts with the environment on behalf of the person." Changes in performance often affect perceived trust in a system [17]. It has been shown that an agent's *predictability* [12] and its *transparency* [11] aid to increase a user's trust in the system, so that the user can work more effectively with the agent. This is because predictability ensures that a user's actions will lead to consistent outcomes, and transparency will provide users with an understanding of why the agent is acting in such a manner.

However, too much trust in an agent can lead to complacency - and so the user becomes overly reliant on the agent, effectively preventing users from adequately responding to new situations in case of an agent's error, as shown in a study where participants had to manage a simulation of a partially automated life-support system [13]. On the contrary, when a user does not trust an agent, the agent becomes under-used (if used at all) and may increase workload and error risks (either because the agent is not used, or because the user needs to actively ignore the agent) [14].

Of course, the trust a user has in an agent is a complex, non-binary relationship that evolves over time and depends upon the circumstances and task(s) at hand [6]. One's initial trust in agents may largely determine whether the agent is to be accepted, and thus used, in the first place. This "default" trust has been shown to be affected by two main factors: the individual's own expertise (how good the person is at doing this task alone) and the use-case of technology (how safety critical the task is: what would be the consequences of an error?) [10]. However, this trust relationship evolves over time - and is largely dependent upon the agent's actual performance. This process is called Trust Calibration [3].

Over time it is believed that users will start to build up trust (or not) in the agent depending on how well it performs and how predictable and transparent its actions are [11]. Nonetheless, uncertainty of actions can prevent the user from knowing when to rely on the agent [5], while agent errors can adversely affect how much the agent is trusted, effectively decreasing reliance on the agent.

Monitoring trust in automation is then crucial to avoid complacency in the case of over-trust and inefficiency in the case of under-trust [4]. Consequently, if we are to develop effective collaborative agents, we need to have a better understanding of how trust between agents and humans evolves, and how it impacts overall performance [20]. Having a better understanding of this relationship and the means to assess it in real-time could lead to more robust collaborative systems capable of giving the right amount of control and responsibility to users and agents.

HOW TRUST IN AUTOMATION IS ASSESSED

Trust is generally assessed via pre and post-hoc questionnaires [12, 15, 16]. For instance, the "Complacency Potential Questionnaire" [16] was designed to evaluate user's initial trust in automation and is typically used before an experiment, while the "Trust between People and Automation"[15] questionnaire is commonly used to measure users trust in the agent after working with it. Such questionnaire are, however, subjective in nature, and only capture "trust" at a single point in time.

In terms of experimental design, most studies assessed trust relationships between users and a system according to pre-defined experimental scenarios. For instance, Correia et al. [2] used a Tangram game where users had to play with a robot in order to complete the task. Wang et al. [17] used a simulation of search and rescue robots. Another study conducted by Korber et al. in 2018, studied how users were trusting automated vehicles under different circumstances [7]. These studies underline the need to have *interactive* tasks, actively forcing users to make decisions with the possibility to trust or not the agent's inputs.

Current approaches to evaluate trust relationships in agent-human collaborative tasks present shortcomings as well. For instance, the majority of the measurements used to infer trust levels is based on task independent subjective questionnaires used in a pre or post-hoc fashion, which does not allow for assessing trust in a more objective and real-time manner. In order to address these issues, we propose a framework allowing to record, measure and analyse trust levels as they change overtime.

BEHAVIOURAL INDICATORS OF TRUST

Since trust has been shown to be positively correlated with performance, and performance depends on the negotiated interactions between the user and the agent, we then hypothesise that the user's trust in the system may be inferred from their behaviour. To this end, we have created an interactive task in the form of a video game - where various factors can be manipulated to better understand the relationship between performance, behaviour and trust. In our game, participants have to collaborate with different agents in order to aim at and destroy multiple waves of incoming missiles (see Figure 1). This context provides a simple, accessible, but challenging task within a controlled environment.

In our initial experiment, we have focused on exploring differences in perceived performance and trust in automation while participants are interacting with agents that display different behaviours and

Figure 1: A screen capture of the current implementation of the game. Here, a participant fires a projectile at an incoming missile while the cross-hair is being controlled by an agent. The game was strongly inspired by "Missile Command" originally released on Atari2600 in 1980 [18].

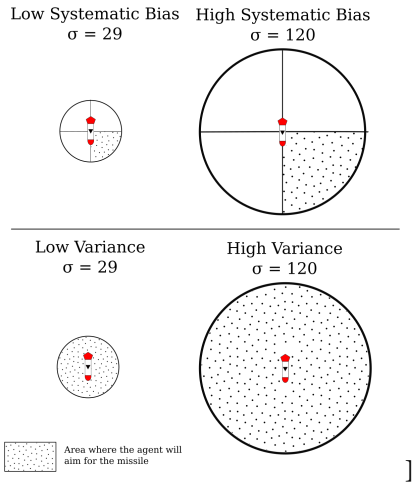

Figure 2: Examples of the different biases displayed by the agents.

levels of performance. We defined these differences according to two main characteristics: systematic biases and variance (see Figure 2 for more details). These elements introduce changes in agents accuracy and performance allowing us to see the resulting changes in perceived trust and performance. Before interacting with the agents, participants play a single player version of the game. The participants performance and trust in automation will be captured throughout each session with or without agent.

Through a logging system, we can monitor how many times a participant corrected an agent's inputs. This data allows us to infer how much a participant trusted an agent during a particular task set at a specific difficulty. Overall, these in-game data have the potential to give us more objective ways of inferring trust levels compared to the use of standardised questionnaires.

Questionnaires allow us to study how much participants are willing to trust an automated agent. As we already know how well an agent can perform, we will be able to compare the participants subjective assessments of the agent with the data related to their interaction captured using the logging system. These data will help us answer the following questions: Does the participant perception of the agent correlates with the agents performance? If so, are participants willing to rely more on the best performing agents? How much effort are the participants making with the agents they trust the most? Is it higher or lower than when they play without any agent?

FUTURE WORK

In this paper, we set out to explain why trust in automation is an essential aspects in the development of more effective collaborative systems. We then presented the state of the art methods to measure trust in automation and explained why we should introduce more objective ways of investigating trust relationships between users and agents. In our next study, we will compare subjective and objective measurements of trust in order to see how trust in automation is related to the performance and behaviour displayed by different agents. The data we will be able to gather trough this user experiment will provide us with a better understanding of trust calibration in a collaborative task and how or when an agent has too much or not enough control over a set of tasks. If allowed, we will also showcase the current implementation of our game during the workshop.

REFERENCES

[1] Jason A. Colquitt, Brent A. Scott, and Jeffery A. LePine. [n. d.]. Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. 92, 4 ([n. d.]), 909–927. https://doi.org/10.1037/0021-9010.92.4.909

[2] Filipa Correia, C Guerra, Samuel M, Francisco S. M, and Ana P. 2018. Exploring the Impact of Fault Justification in Human-Robot Trust. In *Proceedings of the 17th AAMAS Conference (AAMAS '18).* IFAAMAS, Richland, SC, 507–513. http://dl.acm.org/citation.cfm?id=3237383.3237459

[3] Ewart J. de Visser, F. Krueger, P. McKnight, S. Scheid, M. Smith, S. Chalk, and R. Parasuraman. 2012. The World is not Enough: Trust in Cognitive Agents. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56, 1 (sep

2012), 263–267. https://doi.org/10.1177/1071181312561062

[4] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. *Proceedings of the 2007 International Symposium on Collaborative Technologies and Systems, CTS* (2007), 106–114. https://doi.org/10.1109/CTS.2007.4621745

[5] K. P. Hawkins, S. Bansal, N. N. Vo, and A. F. Bobick. [n. d.]. Anticipating human actions for collaboration in the presence of task and sensor uncertainty. In *2014 IEEE (ICRA)* (2014-05). 2215–2222. https://doi.org/10.1109/ICRA.2014.6907165

[6] R. R. Hoffman, M. Johnson, J. M. Bradshaw, and A. Underbrink. [n. d.]. Trust in Automation. 28, 1 ([n. d.]), 84–88. https://doi.org/10.1109/MIS.2013.24

[7] Moritz Körber, Eva Baseler, and Klaus Bengler. 2018. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics* 66 (2018), 18–31. https://doi.org/10.1016/j.apergo.2017.07.006

[8] R Kumar and Carolyn P Rose. 2011. Architecture for Building Conversational Agents that Support Collaborative Learning. *IEEE Transactions on Learning Technologies* 4, 1 (jan 2011), 21–34. https://doi.org/10.1109/tlt.2010.41

[9] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (jan 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[10] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[11] Joseph E. Mercado, Michael A. Rupp, Jessie Y.C. Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. 2015. Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. *Human Factors* 58, 3 (2015), 401–415. https://doi.org/10.1177/0018720815621206

[12] John K. Rempel, John G. Holmes, and Mark P. Zanna. 1985. Trust in close relationships. *Journal of Personality and Social Psychology* 49, 1 (1985), 95–112. https://doi.org/10.1037/0022-3514.49.1.95

[13] Juergen S, Alain C, and David W. 2016. Experience of automation failures in training: effects on trust, automation bias, complacency and performance. *Ergonomics* 59, 6 (jun 2016), 767–780. https://doi.org/10.1080/00140139.2015.1094577

[14] Julian Sanchez, Wendy A. Rogers, Arthur D. Fisk, and Ericka Rovira. 2011. Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science* 15, 2 (sep 2011), 134–160. https://doi.org/10.1080/1463922x.2011.611269

[15] S.J. Selcon, R.M. Taylor, and E. Koritsas. 1991. Workload or Situational Awareness?: TLX vs. SART for Aerospace Systems Design Evaluation. *Proceedings of the Human Factors Society Annual Meeting* 35, 2 (sep 1991), 62–66. https://doi.org/10.1518/107118191786755706

[16] Indramani L. Singh, Robert Molloy, and Raja Parasuraman. 1993. Automation- Induced "Complacency": Development of the Complacency-Potential Rating Scale. *The International Journal of Aviation Psychology* 3, 2 (apr 1993), 111–122. https://doi.org/10.1207/s15327108ijap0302_2

[17] Ning Wang, David V Pynadath, and Susan G Hill. 2015. The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams. *Aamas* Aamas (2015), 997–1005.

[18] Wikipedia. 2018. Missile Command — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Missile%20Command&oldid=864894114. [Online; accessed 13-November-2018].

[19] Shawn R. Wolfe, P. A. Jarvis, F. Y. Enomoto, M. Sierhuis, and B. van Putten. [n. d.]. A Multi-Agent Simulation of Collaborative Air Traffic Flow Management. In *Multi-Agent Systems for Traffic and Transportation Engineering*. IGI Global, 357–381. https://doi.org/10.4018/978-1-60566-226-8.ch018

[20] Claudia Z.Acemyan and P.Kortum. 2012. The Relationship Between Trust and Usability in Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56, 1 (sep 2012), 1842–1846. https://doi.org/10.1177/1071181312561371