5-7-2019

Human-Agent Teaming - an Evolving Interaction Paradigm: An Innovative Measure of Trust

Samson Palmer

Dale Richards

Graham Shelton-Rayner

David Inch

Kurtulus Izzetoglu

HUMAN-AGENT TEAMING - AN EVOLVING INTERACTION PARADIGM: AN INNOVATIVE MEASURE OF TRUST

Samson Palmer, Dale Richards, Graham Shelton-Rayner
Coventry University, United Kingdom
David Inch
Horiba MIRA, United Kingdom
Kurtulus Izzetoglu
Drexel University, United States of America

The promise of intelligent decision support systems is presented as a harbinger for humankind. With the potential partnership between the human and autonomous system we could see a significant increase in effectiveness and safety. However, as we see both human and agent team members being integrated we must investigate ways in which we can assess not only the interaction between the two actors, but also the very nature of trust perceived by the human. In this paper we present early findings of an experiment that examines the human-autonomy interaction across different frameworks of authority; from manual to fully autonomous. Participants were asked to interact with a ground control station that supervised several unmanned systems, and presented with goals that required manually selecting or approving assets to achieve mission goals. The use of neuroimaging (in this instance Functional Near Infrared Spectroscopy - fNIRS) was used to establish neuro-correlates of trust within the prefrontal cortical region of the brain. Initial findings suggest that the dorsolateral prefrontal cortical region of the brain is heavily associated when humans are confronted with different levels of authority with human-autonomy interaction.

As we have seen in the previous papers (Part I and II), the construct of HAT, while promising in terms of assisting the human in achieving goals, is somewhat complex and not without problems. When we traditionally think of whether a HAT is performing in an efficent manner it is easy to assess the quantifiable objectives as to whether the goal has been achieved succesfully or not. However, as outlined in Part I and Part II, how the human interacts and perceives the HAT is of critical importance. A key component of this is whether the human team members (single operator supervising the system) trusts the system behaviour.

Trust has yet to be universally defined, particularly when referring to Human-Computer Trust (HCT). However, one definition suggests that HCT is "the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions and decisions of an artificially intelligent decision aid" (Madsen, 2000). All engagements and interactions, both social and with computers, require elements of trust, particularly those that require some form of cooperation. When referring to social interaction this is often called Reciprocal Trust (RT), and is critical in the development of partnerships (Krueger et al, 2007). It is therefore reasonable to assume that the development of a human-automation partnership and the formation of HCT will also require elements of RT. To this effect, researchers have often tried to adapt interpersonal trust models and measurement instruments to explain operator HCT, even creating their own scales. Examples of these include the Empirically Derived scale, the Human-Computer Trust Scale, and SHAPE (Solutions in Human Automation Partnerships in European ATM) Automation Trust Index (SATI). These have all been designed to assess operator trust of decision support systems, and although they have all benefited from systematic development and validation, they have also

attracted criticism (Lewis et al, 2018). To further confound this, much of the research surrounding HCT has utilised scales based on a third-person frame where there is a lack of personal consequence as a result of the decisions made (Miller et al, 2016).

Although the above-cited metrics of trust are based solely on subjective assessment Izzetoglu & Richards (2019) suggest that the validity of using a sole metric for such complex human behaviour should also ideally include behavioural data (such as task completion) and some form of physiological assessment. In particular Izzetoglu & Richards (Ibid.) discuss the use of acquiring data via wearable neurimaging equipment in order to assess human cortical activity. Indeed, some studies have begun to examine and characterise the neural correlates of trust and have often implicated activity within the prefrontal cortex in encoding the motives and intentions surrounding interactions with systems (Sripada et al., 2009). Additionally, some neuroimaging studies have investigated the neural correlates of cognitive processes aligned with cooperation and RT and have further shown activation of the prefrontal cortex, particularly the anterior medial prefrontal cortex (aMPFC) (Bos et al, 2009). With the advances in wearable neuroimaging techniques, the ability to monitor operators during system interaction has provided a chance to further assess the neural activity associated with trust. As we have seen in Part I and II papers of this session, human interaction with autonomous systems is complex and the role of trust plays an integral part in assessing whether the HAT is effective or not. The use of neuroimaging therefore is seen as a promising tool in providing an indication of trust within such HAT compositions.

This paper discusses a recent study whereby functional Near Infrared Spectroscopy (fNIRS) was used to assess prefrontal cortical activity during interaction with systems of varying levels of autonomy. The application of fNIRS allows us to monitor localised hemodynamic changes (mainly in oxygenation) across the prefrontal cortex (PFC) region as a result of increased cortical activity. It was hypothesised that the changes in oxygenation across different *expsoures* of decision support would be indicative of how HCT varied, whilst also how failure of autonomous systems can impact the HCT of the operator.

Exploring Prefrontal Cortical Activity During Human-Autonomy Interaction

In the current study (N=23), participants were asked to supervise eight unmanned vehicles (both ground and air - UxV) via a Ground Control Station (GCS) visual interface, as per Figure 1. Participants were tasked to utilise the different capability UxV's in order to defend each of the four bases (Bases indicated as A, B, C and D in Figure 1). Participants were asked to respond to events across different scenarios as they evolved, resulting in the participant having to either respond directly with (1) a command input (Manual mode), (2) a selection of command inputs as suggested by the system (Assisted Manual), (3) selecting a single recommended command suggested by the system with YES or NO (Assisted Auto), and finally (4) no inputs required from operator as they monitored the system respond and inform the operator of chosen action (Full Auto).

During the scenarios an event was triggered by an unknown entity entering the map and the operator would be tasked to interrogate the unknown signal. The most appropriate UxV would need to be sent to conduct a scanning task based on where assets were currently located, their sensor fit, current task, amount of fuel, speed of response, etc.

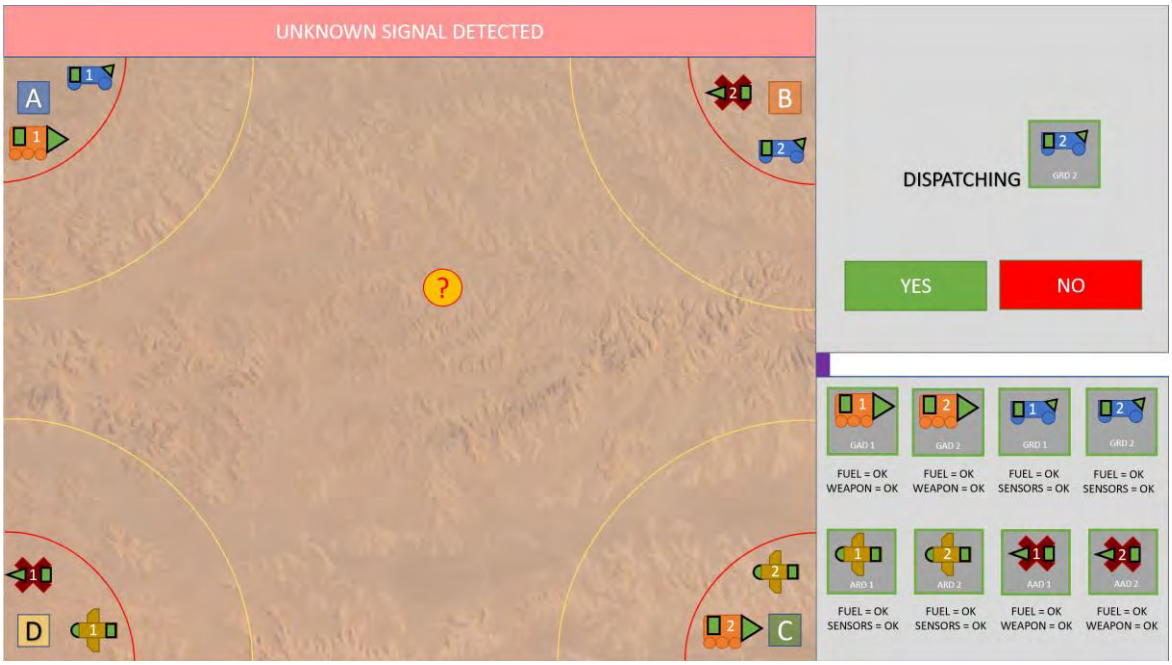

Figure 1 - GCS display showing all UxV on station in each base in Assisted Autonomy mode

Unknown signals consisted of a random mix of enemy combatants and civilians. Mission success occurred if all 4 military bases were still intact once the time ran out, and if no civilians were harmed. Mission failure occurred if an enemy combatant crossed the red barrier surrounding each base, or if a civilian was harmed. Scenario order was counterbalanced for each participant to avoid order effects of the repeated measures design of the experiment. Each scenario presented was designed with a different level of autonomy. The Assisted Manual scenario required participants to select from 2 or 3 drone options presented to them once an unknown signal appeared. However, the drone options presented were based solely on distance to target (closest drone was displayed as first option) and did not account for drone status or capabilities, therefore participants were required to check this before making a decision. The Assisted Auto mode required participants to agree to the decisions made by the system by selecting YES or NO. However, this time the decisions made were based upon all aspects of the system, and could be perceived as the best possible decision. Additionally, safeguards were in place whereby if a NO decision would result in base destruction, this option was removed. There were two Full Auto mode scenarios, where participants were required to observe the system with no interaction required. However, the integrity of one of the systems during these scenarios were manipulated in a way that could be perceived as aggressive and incorrect, whereby attack drones were always deployed first without scanning unknown signals, which ultimately ended in a civilian being harmed and thus failing the mission.

Initial Results from Experiment

The study is still ongoing at this time with current fNIRS data collected and analysed will be discussed. In order to assess prefrontal cortical activity, fNIRS neuroimaging was used. This allowed the collection of a 5-minute baseline before continuously gathering cortical activity of

participants throughout all scenarios. Following the completion of artefact removal and extraction of processed optical density data (Izzetoglu & Richards, 2019), a modified Beer-Lambert Law (MBLL) was applied for calculations of channel-specific changes in oxygenated hemoglobin (oxy-Hb) and de-oxygenated hemoglobin (deoxy-Hb). This data was collected across 16 optodes aligned with the PFC as shown in Figure 2.

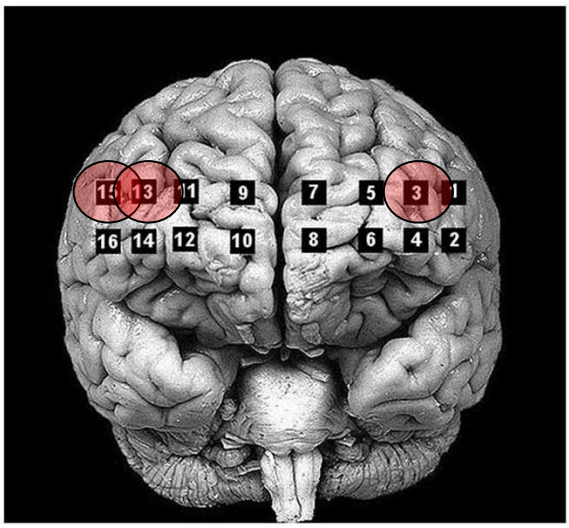

Figure 2 - Topographical image displaying optode placement across the prefrontal cortex (with associated regions that showed increased ratio of oxy-Hb and deoxy Hb circled)

Initial analysis demonstrates that optodes 3, 13, and 15 detect an increase ratio of oxy-Hb to deoxy-Hb during the Full Auto Incorrect and Assisted Manual scenarios. These scenarios were expected to place higher cognitive demand on the participant in terms of Memory (WM) and Attention, and optodes 3, 13 and 15 correlate to regions associated with both WM and Attention (Reddy et al., 2018). Figure 3 shows the mean values for each task across each optode.

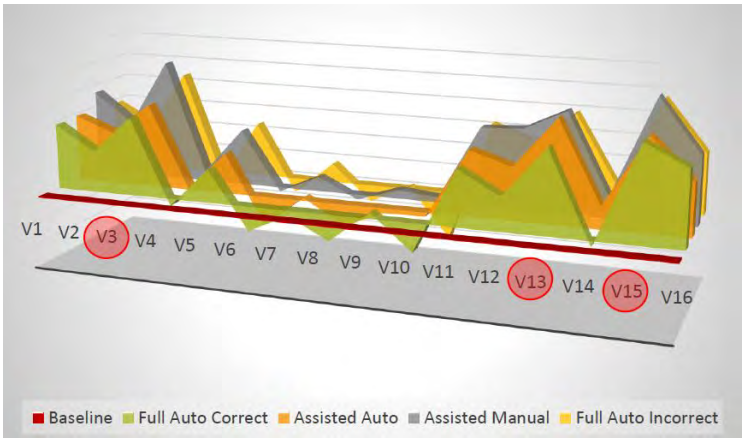

Figure 3 - Mean ratio values of oxy-Hb and deoxy Hb across all optodes (with larger variance from baseline circled)

As can be seen in Figure 3, when we compare a number of optodes between baseline fNIRS reading against the four experiment conditions (varying authority and delegation of control), we

see a significant main effect for optode 3 ($F(4,110)=4.111$, $p<.005$), optode 13 ($F(4,110)=17.504$, $p<.05$). Optode 15, while not significant, showed promise of a main effect (($F(4,110)=16.806$, $p=.13$) - but noise from this optode meant losing seven participants sets of data from the analysis.

Discussion

It is often assumed that the addition of an autonomous aid to a human operator will result in a more effective system. However, as Dzindolet et al (2003) report, this can be inaccurate: if the operator does not trust the system fully, or the operator trusts a system that does not provide correct support, it can lead to disuse or misuse. Therefore a means of assessing the level of HCT is an important tool, especially when we consider safety critical systems. This paper reports initial findings where brain-based measures via fNIRS were used to study neural mechanisms during numerous scenarios with varying levels of autonomous support. It was hypothesised that the fNIRS measures may indicate how the autonomous support system affects the human operator's trust of the system, and that the neural correlates of trust could be observed during these scenarios.

Preliminary findings indicate that optodes 3 and 13 (with possibly 15) showed activation during the Assisted Manual and Full Auto Incorrect tasks. These particular optode regions of the PFC have been linked to Working Memory (WM) and Attention (Reddy *et al* 2018), and can be aligned with Brodmann's areas 9/10 (Zilles 2010), also referred to as the dorsolateral/medial and anterior prefrontal cortex, respectively. Previous research suggests that trust, particularly reciprocal trust, can be associated with greater activity in the dorsomedial prefrontal cortex (Filkowski et al, 2015). Other research has also shown the neural correlates of uncertainty may be linked to the dorsolateral prefrontal cortex (Pushkarskaya et al, 2015) and the results of this experiment, demonstrate increased activity in these regions during scenarios that required increased operator input, and where the autonomous system could be perceived as making incorrect decisions.

It is important to understand how human operators trust intelligent systems, and thus while we can observe behavioural performance and collect self-ratings from individuals, objective physiological measures could provide insight in assessing trust. This study is part of the new exciting investigations that examine the neural correlates of trust and distrust. We propose this method of cortical assessment represents an opportunity to better understand our relationship with HAT.

References

Bos, W., Dijk, E., Westenberg, M., Rombouts, S. and Crone, E. (2009) "What Motivates Repayment? Neural Correlates Of Reciprocity In The Trust Game". *Social Cognitive And Affective Neuroscience,* 4(3), 294 - 304.

Dzindolet, M., Peterson, S., Pomranky, R., Pierce, L. and Beck, H. (2003) "The Role Of Trust In Automation Reliance". *International Journal Of Human-Computer Studies* [online] 58 (6), 697-718.

Filkowski, M., Anderson, I. and Haas, B. (2015) "Trying To Trust: Brain Activity During Interpersonal Social Attitude Change". *Cognitive, Affective, & Behavioral Neuroscience,* 16(2), 325-338.

Izzetoglu, K., & Richards, D. (2019). Human Performance Assessment: Evaluation of Wearable Sensors for Monitoring Brain Activity. In M. Vidulich & P. Tsang (Eds.), Improving Aviation Performance through Applying Engineering Psychology: Advances in Aviation Psychology (1st ed., pp. 163–180). Boca Raton, FL: CRC Press..

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., Heinecke, A. and Grafman, J. (2007) Neural Correlates Of Trust. *Proceedings Of The National Academy Of Sciences,* 104 (50), 20084-20089.

Lewis, M., Sycara, K. and Walker, P. (2018) The Role Of Trust In Human-Robot Interaction." in *Foundations Of Trusted Autonomy. Studies In Systems, Decision And Control* [online] ed. by Abbass, H., Scholz, J. and D, R. Springer, Cham, 142 - 144. available from <https://link.springer.com/chapter/10.1007/978-3-319-64816-3_8#citeas> [28 February 2019]

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K. and Ju, W. (2016) "Behavioral Measurement Of Trust In Automation". *Proceedings Of The Human Factors And Ergonomics Society Annual Meeting,* 60 (1), 1849-1853.

Pushkarskaya, H., Smithson, M., Joseph, J., Corbly, C. and Levy, I. (2015) "Neural Correlates Of Decision-Making Under Ambiguity And Conflict". *Frontiers In Behavioral Neuroscience*[online] 9. available from <https://www.frontiersin.org/articles/10.3389/fnbeh.2015.00325/full>

Reddy, P., Richards,D., & Izzetoglu, K. (2018) *Cognitive Performance Assessment of UAS Sensor Operators via Neurophysiological Measures*, in *2nd International Neuroergonomics Conference*.

Sripada, C., Angstadt, M., Banks, S., Nathan, P., Liberzon, I. & Phan, K. (2009) "Functional Neuroimaging Of Mentalizing During The Trust Game In Social Anxiety Disorder". *Neuroreport*, 20(11), 984-989.

Zilles, K. and Amunts, K. (2010) Centenary Of Brodmann's Map — Conception And Fate. *Nature Reviews Neuroscience*, 11(2), 139-145.