Trust Management for Human-AI Symbiosis

Ming Hou

Defence Research and Development Canada

***Abstract*— To facilitate the appropriate level of human trust in AI, a mechanism to continuously evaluate and calibrate human-AI trust is required. Such a Trust Management System (TMS) is integral to developing trustworthy AI systems and thus enable collaborative and effective Human-AI Teaming (HAT) in broad AI applications. This paper presents the IMPACTS (intention, measurability, performance, adaptivity, communication, transparency, security) trust model as a system level requirement framework for TMS. An associated IMPACTS homeostasis TMS is also discussed for managing trust given the dynamic and transactional nature of trust. These two models provide guidance for continuous trust monitoring and behavior adjustment to ensure calibrated trust over time.**

1. INTRODUCTION

With its increasingly autonomous information processing and decision-making capabilities in broad applications, artificial intelligence (AI) technologies hold the promise of delivering transformative changes in our society. With the collective human-machine intelligence, the human-AI teaming (HAT) is able to benefit the strengths of both while be mindful of their limitations, thus towards a collaborative symbiosis partnership. Trust then becomes an important and critical aspect of this relationship. It is trite to note that humans are more likely to rely on AI systems that they trust and reject AI systems that they do not trust. However, garnering and maintaining trust in an AI are challenging design constraints, requiring solutions that encourage acceptance by both the general public (e.g., moral issues) and the intended users in specialized domains (e.g., fitness for purpose).

If humans overtrust an AI system, they can rely too much on it, putting themselves at risk of missing certain threats or potentially propagating errors made by the system. On the other hand, if AI is undertrust, humans will not rely on the assistive technology, making it underused and potentially increasing the risk of human errors. As such, maintaining optimal level of trust is crucial to ensure the right trade-off is met. This paper documents the thinking behind the trust development framework IMPACTS (Intention, Measurability, Performance, Adaptivity, Communication, Transparency, Security) and information that has contributed to the HAT trust management research theme. The objective of the IMPACTS framework is to provide a foundation for incorporating an intelligent and adaptive system that seeks to maintain an appropriate level of trust, ideally bi-directional "trust" amongst human and AI teammates, that reflects both context and capability of autonomous AI agents within the HAT. As the first step towards this objective, a Trust Management System (TMS) for AI-enabled technologies seeks to calibrate the human trust in the system at a level appropriate to the context and capability (i.e., what we refer to as "trust-homeostasis") thereby producing more effective HATs in dynamic, complex scenarios that entail significant risks.

2. TRUST REQUIREMENTS

The IMPACTS design framework was developed to represent seven essential elements and high-level system requirements of an AI system [1][2][3]. IMPACTS is intended to assist the design of TMSs, evaluating prototypes and validating final system designs. It is based on the understanding of capability and integrity requirements of a human-AI team. Integrity is a fundamental ethos for the development of trust within a team. It requires an AI system to act with strength of character to earn and maintain the trust of the team and adhere to the highest ethical standards with reliability. It does not allow any conduct that is in any way harmful, discriminatory, illegal or inappropriate [4].

As illustrated in Figure 1, the seven dimensions of the IMPACTS design framework are:

Figure 1. IMPACTS design framework with seven essential trust requirements (modified from [1].)

- Intention: AI systems must behave in a way that is aligned with the human's intentions and ethical norms or values;
- Performance: AI must exhibit reliable, robust and predictable behaviours to maximize system performance;
- Adaptivity: AI must learn, understand and adapt to changes in the: situation, environment, system functionality, task status, human partner's mental state and performance outcomes as well as guard the human resources (e.g., attentional capacity) and time to achieve the team's common goals;

- Communication: AI must facilitate bi-directional communications through a Human-Machine Interface (HMI) to communicate the intentions, actions and decision reasoning with its human partner;
- Transparency: AI behaviours, intentions and decision reasoning must be explainable to its human partner at an appropriate time (e.g., at an optimal workload level) with the correct format and pace (e.g., intuitively understandable means and appropriate level of details) so that the human can develop an accurate mental model of AI's intentions and end states; and
- Security: AI must ensure system safety and remain protected against accidental or deliberate attacks.

The concept of IMPACTS and the related seven trust dimensions as the trustworthy HAT requirements have been validated in a large-scale military exercise in the HAT context for managing weapon engagement processes [1][2]. It has provided the basis of a TMS design framework that will be discussed in the following sections.

3. TRUST MANAGEMENT

The IMPACTS development framework is intended to be used to inform the design of the TMS. It facilitates both the concept exploration and preliminary design phases (essential requirements for effective HAT operations) as well as the prototype evaluation and final validation of the AI system. To be useful, the IMPACTS trust model must include a generic TMS design model to serve as the basis for analyzing the system requirements and identifying risks and design options.

To fulfil this requirement, an intelligent and adaptive TMS design model based on the IMPACTS trust model is developed accordingly. As illustrated in Figure 2, the TMS boundaries are shown as well as data flows to and from other Intelligent Adaptive System [5] modules including situation assessor, operator state monitor, Intelligent Adaptive Automation, Intelligent Adaptive Interface, and Adaptation Engine. The TMS trust comparator compares the adjusted user/operator trust that the operator allots to the system with the adjusted system trustworthiness that represents the amount of trust that the operator should have in the system (i.e., the deserved trust). The adjusted operator trust is calculated based on dispositional trust values computed prior to operations and adjusted based on the situational trust computed during operations and based on differences between the AI system (i.e., robots) expected and observed behavior. The adjusted system trustworthiness takes the user perceived system trustworthiness that includes contextual trust inputs from the operator state monitor system and corrects it, taking into account the robot's self-assessment of its own trustworthiness. The outputs of the trust comparator will lead to a trust repair action if the value is above or below the calibrated trust threshold values, in which a trust repair action request is sent to the Adaptation Engine with information that can inform the nature of the trust repair that is required.

Combining the outputs of the operator State Monitor and Situation Assessor, the Adaptation Engine can infer how and why operator trust was lost and, crucially, the optimal trust repair strategy that the robot needs to undertake to restore operator's trust in it. For example, a trust repair strategy might require changes to the Intelligent Adaptive Interface for the robot to apologize for being late and express regret to the planned location, explain the reason (e.g., loss of GPS signal necessitating route finding using a slower internal inertial navigation method), and suggest a change in its behavior to mitigate the same issue happening again (e.g., forewarn the operator of a deviation from its estimated time of arrival at the earliest opportunity and suggest manual control of the robot until the GPS capability is restored).

The inputs and outputs of the TMS are finite and constrained by technology (e.g., field-ability and sensitivity of trust measures) and ethical/legal considerations (e.g., trust repair strategies involving blame or denial are not ethical), and that the selection of the optimal trust repair strategy is highly deterministic based on, in part, the formal representation of the IMPACTS trust homeostasis model. Thus, it is eligible as a TMS development framework to formulate an adaptation mechanism (Adaption Engine) and facilitate intelligent adaption to human trust.

4. CONCLUSIONS

The issue of HAT trust is becoming increasingly important as our reliance on AI systems increases. Developing an understanding of how HAT technologies can work in a trustworthy manner and how to manage trust are thus critical to successful HAT. A key technological component is the development of TMSs to manage and calibrate trust between the human operator and AI systems. Hence, this paper first presents a cross-domain conceptual model of trust IMPACTS useful in informing the requirements for design and development of TMS, then described the framework of IMPACTS homeostasis TMS to dynamically monitor and manage, and maintain optimal level if trust for a collaborative human-AI symbiotic partnership.

Future work will develop and evaluate the IMPACTS homeostasis TMS prototype and elaborate various model elements. This includes the application of recommended trust measures in a series of experimental activities including simulation-based evaluations of the TMS and field trials exploring trust repair strategies for collaborative and effective HAT in future applications. The results of this research activity will help further validate the IMPACTS homeostasis model and support the future implementation of the TMS.

REFERENCES

[1] M. Hou et al., "IMPACTS: a trust model for human-autonomy teaming," Hum. Intell. Syst. Integr., pp. 1–19, 2021.

[2] M. Hou, et al. "Frontiers of Brain-Inspired Autonomous Systems: How Does Defense R&D Drive the Innovations?" IEEE Systems, Man and Cybernetics Magazine 8 (2), pp. 8–20, 2022.

[3] M. Hou, et al., "Interaction-Centered Design: An Enduring Strategy and Methodology for Complex Socio-Technical Systems". Handbook on Human-Machine Systems: State-of-the-art and Research Challenges, pp.125-139, 2023.

[4] Department of National Defence, Canada,. "Canadian Armed Forces Ethos: Trusted to Serve. The Canadian Defence Academy – Professional Concepts and Leader Development," 2022.

[5] M. Hou et al.," Intelligent Adaptive Systems: An Interaction-Centered Design Perspective", CRC Press, 2014.

Dr. Hou is a Principal Scientist and Authority in Human-Technology Interactions within the Department of National Defence (DND), Canada. He is responsible for delivering technological solutions, science-based advice, and evidence-based policy recommendations on AI and Autonomy science, technology, and innovation strategies to senior decision makers within DND and its national and international partner organizations including the United Nations. As the Canadian National Leader in human-AI/autonomy teaming, he directs the Canadian innovation ecosystems on a number of capability development programs to support Canadian major acquisition projects and large-scale live, virtual, and constructive international joint exercises. As the Co-Chair of an international Human-Autonomy Teaming Specialist Committee, he leads the development of international standards for the integration of AI-enabled autonomous systems into the civilian airspaces. Dr. Hou is the recipient of the most prestigious DND Science and Technology Excellence Award in 2020 and the President's Achievement Award of the Professional Institute of the Public Service of Canada in 2021. He is an Adjunct Professor at the University of Toronto and University of Calgary. Dr. Hou is an IEEE Fellow, Distinguished Lecturer, the Lead of Human-AI Teaming of IEEE AI Coalition, and the General Chair of the 2024 IEEE International Conference on Human-Machine Systems and International Defence Excellence and Security Symposium.

.

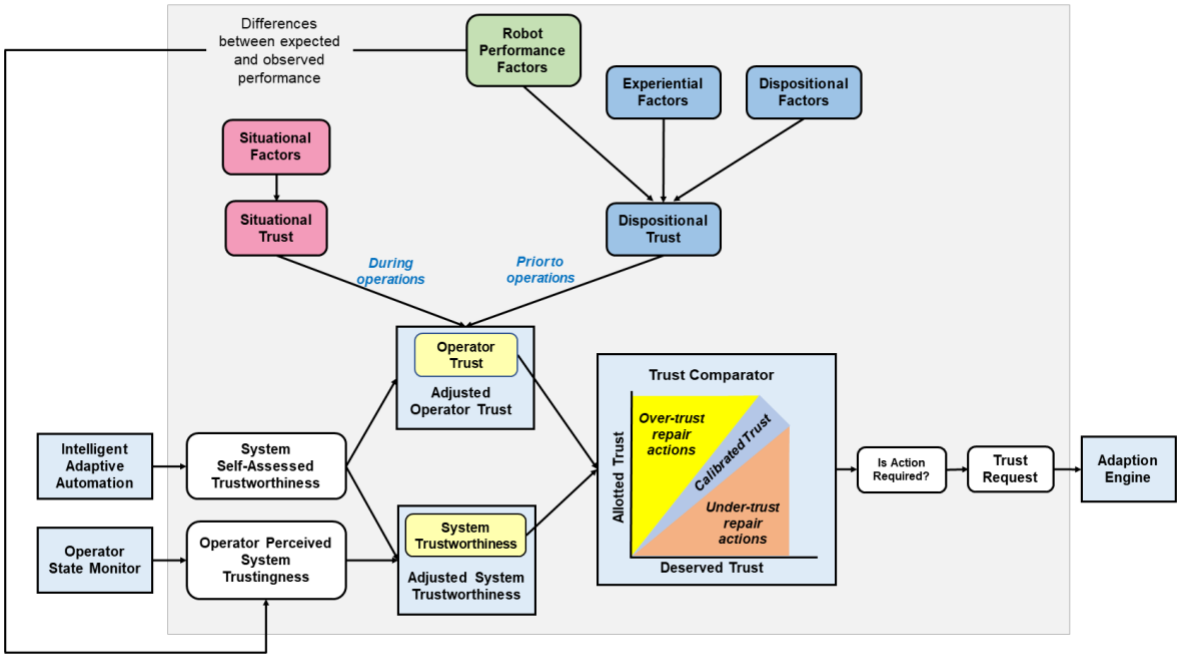

Figure 2. IMPACTS homeostasis trust management model for Collaborative Human-AI Symbiosis