Mediating Explainer for Human Autonomy Teaming

Siri Padmanabhan Poti[1,*], Christopher J. Stanton[1]

[1]The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Westmead Innovation Quarter, Building U, Level 4, 160 Hawkesbury Road, Westmead, NSW 2145, (Australia)

Abstract

This paper examines the environment of mission-critical HAT operations, and conceptually models the human intent, AAI agency, and societal context. The conceptual model employs agency theory to describe the relationship between human principals and Autonomous Artificial Intelligence (AAI) agents in HAT. Further, an application of stakeholder theory prompts the inclusion of societal stakeholders' roles in mission-critical HAT operations. The model reveals the opportunity for incorporating an intermediary mechanism of a non-human Mediating Explainer (MeX). MeX offers a novel means of resolving the asymmetries of information and decision-making power in HAT relationships.

Keywords

XAI Human Autonomy Teaming, Social Legitimacy, Agency Theory, Stakeholder Theory,

1. Introduction

Autonomous Artificial Intelligence (AAI) systems can perform complex tasks and can be employed to deal with ambiguities that require human-like abilities [1, 2]. AAI systems are considered 'rational agents' that 'operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue goals' in a manner that can bring about 'the best outcome' or the 'best expected outcome' in non-deterministic situations [3]. AAI is often required to operate in unstructured environments that are partially known [4], presenting challenges [1, 5], requiring planning and adaptive decisions, while also raising the possibility of unexpected or unintended outcomes to the organization or society at large [1]. Thus, in mission-critical operations, a human collaboration, cooperation or control of AAI systems, referred to as Human Autonomy Teaming (HAT), is implemented [5, 6].

1.1. Aims and Method

This paper examines the motivations for employing HAT in mission-critical operations in terms of the human intent, AAI agency, and the societal context. It presents a 'conceptual model' [7] that 'descriptive[ly]' and 'normative[ly]' [8, 9] examines 'the focal construct' [7] of explanations

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Corresponding author.

siri.padmanabhan@westernsydney.edu.au (S. Padmanabhan Poti); c.stanton@westernsydney.edu.au (C. J. Stanton)

https://www.westernsydney.edu.au/marcs (S. Padmanabhan Poti); https://www.westernsydney.edu.au/marcs (C. J. Stanton)

0000-0002-0877-7063 (S. Padmanabhan Poti); 0000-0001-7814-6120 (C. J. Stanton)

in mission-critical HAT operations. The conceptual model explores the multi-level delegation of risks and hierarchical devolution of decision-making power to non-human autonomous AI agents to fulfill human intent in a societal context. It employs agency theory to describe the persuasion of economy and efficiency in HAT environments. The model, applying stakeholder theory, broadens the discourse of explainability in the HAT environment [9], enabling its social legitimacy [10]. It proposes a central role for a non-human Mediating Explainer (MeX) in resolving the asymmetries of information and decision-making power, overcoming limitations of bounded rationality within the HAT environment.

2. Conceptual Model of the HAT environment

Employing agency theory and stakeholder theory [9, 11, 12, 13], the intent-agency-context in the HAT environment is conceptually modelled here. Principal-agent relationships in an organizational context may be well understood with the 'descriptive' aspects of agency theory [8, 9]. The 'economics paradigm' [13] available in agency theory, can describe the reason for HAT in mission-critical operations. Additionally, to avoid a possible 'partial view of the world' from agency theory alone, the stakeholder theory serves as a 'complementary' theory for a broader societal context [14], providing 'normative' and 'descriptive' aspects [9] to the HAT environment (Fig. 1.). Stakeholder theory extends the HAT environment to the collective context of 'principled moral reasoning' [9] and expectations from society at large.

2.1. Describing the HAT environment with Agency Theory

The need to leverage complementary skills of 'risk-averse' [14] human principals and AAI agents meant to engage with high risk and uncertainty, can be considered the main motivation for HAT [6, 15]. For example, in safety-critical undersea and littoral 'mine counter-measure' (MCM)' operations [16], where missteps may lead to dire consequences [17], HAT is deployed to achieve MCM goals [18]. Applying the context of agents and principal-agent relationships to HAT, human individuals in their capacity as the principals may be seen to engage AAI as their agent to perform a task or service 'on their behalf' and 'delegate' some of their 'decision-making authority' [19] to the AAI system. The 'agency' role of the AAI system in risky or complex HAT operations can be considered an 'acting for' relationship, where the presence of a 'formidable physical, social, temporal or experiential barrier separates principal and agent' [13]. As found in other agency theory related studies across multiple disciplines and contexts, in the HAT environment too, there is 'outcome uncertainty' that 'trigger[s] the risk implications of the theory' [14]. Differential risk exposure is the reason the relationship between the human-principal and AAI-agents in HAT exists. The essence of the HAT relationship can be mostly described from examining information asymmetry and asymmetry of decision-making power in HAT [14].

2.1.1. Information Asymmetry in HAT

By viewing 'information as [a] commodity' in the relationship between human principals and AAI agents [9], the explanations, interpretations, and assurance statements of AAI systems in

HAT can be understood as the information with which their intentionality and behavior can be explained, interpreted, or assured to humans [20]. In human-human teams, mutual communication is seen as prerequisite to establishing 'team flow' while working on tasks requiring collaboration, which in turn manifest as 'mutual trust' [15]. However, in HAT operations, there is an innate inability to communicate due to the absence of shared mental models [21, 22] among human principals, human agents, AAI systems and stakeholders. Compounding this, 'information asymmetries' could be amplified due to human principals and the AAI agent being misaligned in their values from inapt utility incentives built into AAI systems [23].

According to the SAFE-AI model [24], it is vital for the human to have access to information-rich communication from an AAI system for situation awareness levels of 'perception', 'comprehension' and 'projection'. These let the human principal or human agent teammate perceive the AAI system's current activities and decisions, comprehend the causality, assess the implications of decisions and activities, and project the next steps [24]. There might be no discernible benefit from explainability of autonomous agents when likelihood and impact of risks from their operation are negligible, or when the risks are universally known and accepted [25]. However, in mission-critical operations, explanation of intentionality during task execution is crucial for human trust in AAI systems [25]. When information communicated from AAI agency back to the human principal or other fellow human agents is not 'correct', 'relevant' or is unsuited to the human's understanding, the asymmetry can impact success of operations [24].

There are also the variety of information asymmetries that may stem from the issues of 'self-interest' [14], lack of capability and motivation and naïve or willful non-compliance of 'human[s]-in-the-loop' [26]. On one hand, the human principal's trust in the AAI system may be influenced by the communication of the AAI system's intent, especially as the AAI autonomy to act on that intent increases [4]. On the other hand, regardless of the information provided by the agents, the 'bounded rationality' typical of humans [14]would constrain the human principals, fellow human agent and other stakeholders in HAT from understanding the entirety of the logic and rationale presented directly from multiple AAI agents in real-time.

2.1.2. Asymmetry of Decision-making Power in HAT

In mission-critical HAT operations, situations involve 'goal conflict', 'risk aversion', 'cooperative effort', and constraints from 'bounded rationality' [14]. The HAT environment is also thought to have centralized decision-making by the human principal, where delegation of decision-making power is 'hierarchical' [27]. When economy and 'efficiency' mainly constitute the 'criterion for effectiveness' [14], a 'centralized' decision-making authority is delegated through 'hierarchies' [27] for an allocation of resources and outputs [14, 28]. The HAT environment may be seen as having 'unprogrammed or team-oriented jobs in which evaluation of behaviors is difficult' [14]. This prompts the ongoing 'agency problem' of the cost incurred by the principal 'to verify what the agent is actually doing' [14]. The asymmetry of decision-making power would shift towards AAI as the 'Level of Autonomy (LOA)' varies from 'manual control' to 'High agent autonomy' over ten levels [2]. The power of decision-making would be based on the LOA, and the extent of human delegation of decision-making to autonomous systems in HAT based on capability of systems, risks involved, and trust in systems [2]. Also, in terms of the modes of human collaboration in HAT, human principals may be seen to share decision-making power with AAI

systems through the various modes of HAT such as 'human in the loop' (HITL), 'human on the loop' (HOTL), 'human starts the loop' (HSTL), and 'human ends-the-loop' (HETL) [5, 18].

Despite the inclusion of humans principals to the loop in various forms, several challenges [1] to societal values may be expected from decision-making in HAT, and these may affect the 'economic and moral nature of relationships and decision-making in, and around, organizations' [11, 29]. Also, human collaborations with AAI agents can be expected to be 'susceptible to social dynamics' and 'group dynamics' [30]. To include the several stakeholders that mission-critical HAT operations are accountable to [29], stakeholder theory is employed here to 'normatively' and 'descriptively' [9] broaden the decision-making structure towards a 'polyarch[y]' [27].

2.2. Broadening the HAT Environment with Stakeholder Theory

The social legitimacy of any mission-critical HAT operation by an organization or principal may require to be seen in a larger moral light and one of broader societal trust than just limited to the utilitarian, economic or efficiency aspect of that operation [9]. Stakeholder theory can be thought of as a broader approach to organizational behavior, wherein the available descriptions from agency theory can be 'subsumed' [9]. An analysis of the stakeholders, their influence, interest and expectations in the HAT environment provides for the optimal engagement of stakeholders to gain their trust [9, 12]. Drawing attention to stakeholders draws out the 'implicit social contract' [9] among the principals, agents and society that are involved in the HAT environment. In mission-critical operations involving HAT, the stakeholders range from managers of AAI system-based operations in organizations, end users or consumers of AAI systems, organizations that develop AAI, and all those affected directly or indirectly by AAI system-based operations or decisions [29].

Stakeholder inclusion expands the seemingly limited scope of a HAT operation from being an interaction between human principal and AAI agent to a milieu of 'value creation' and trust among stakeholders [11, 12] in society. The normative aspects of stakeholder theory embed the utilitarian HAT operations in the broader perspective of organization, community, and society. This also balances out the 'hierarchic' nature of the 'decision-making structure' with a 'polyarchic' context of stakeholders and society [27]. The balance enables decision-making that can navigate the 'Type I errors' of 'rejecting a superior alternative' and 'Type II errors' of 'accepting an inferior alternative' [27]. The decision-making shifts from being driven by economy and efficiency alone, to being embedded in the intent-agency-context milieu.

Given the diversity of human principals, human agents, and societal stakeholders in the HAT environment, their varying influence levels and interest levels may represent a spectrum of 'mental models' or 'knowledge structures' [21, 22] in use to understand the mission-critical HAT operation. There may also be multiple 'shared mental models' [21] in different sub-teams within the HAT environment depending on their experience and expertise. The communication or information exchanged would need to be tailored to the purpose, such as for evaluation, justification, management or learning [29]. The conceptual model presented thus far, now presents the opportunity for MeX to resolve the asymmetries in the intent-agency-context environment of mission-critical HAT operations (Fig.1).

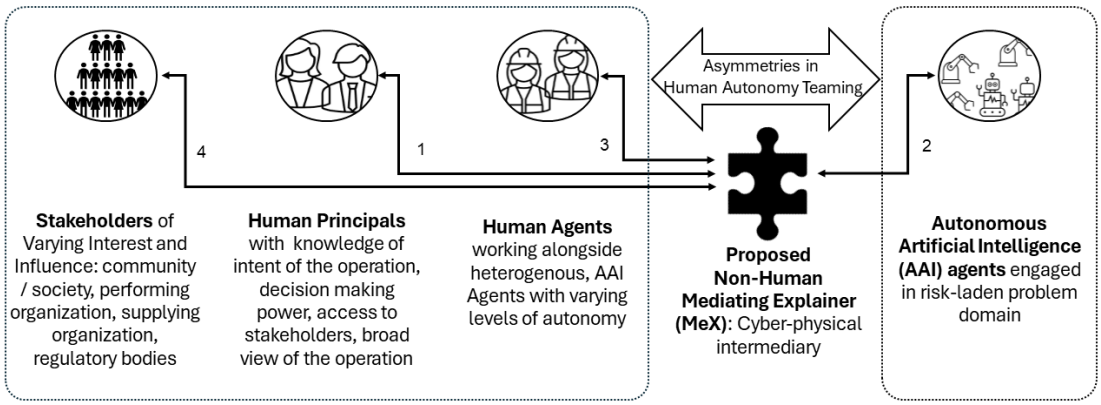

Figure 1: Non-human Mediating Explainer (MeX) in the HAT environment.

3. Opportunity for a Mediating Explainer (MeX)

From related work in literature pertaining to complex socio-political settings, 'intermediation' is known to help in resolving imbalances typical of principal-agent relationships, such as in the 'distribution of knowledge' and 'discretionary power' [31]. MeX in HAT (Fig.1.) may be likened to a 'middleware' layer in 'Service Oriented Architecture (SOA)' for 'cyber-physical systems' [32, 33]. It can also be seen as a broad platform that could host an 'artificial SMM' [21] or 'Decision Support Tool (DST)' in the HAT environment [34]. It would be able to reach out to the affordances provided within AAI systems to gather information [35]. In multi-level principal-agent HAT relationships [36], MeX can take on the non-judgmental and non-fearing nature of non-human agents [6, 15], facilitating a safe environment for putting forward recommendations and explanations to the principal, organization, society, and other stakeholders. This may otherwise not be available from fear of judgement or punishment. The 'recommendations', 'reputations' or 'referrals' [37] from MeX, as a third party, can be 'influencing factors' in resolving imbalances [37]. MeX would effectively adapt complex information available from the AAI systems to suit the workload and information needs of the human principal, while also making available information to the organization and stakeholders. MeX would be able to resolve the asymmetries among the stakeholders in the HAT environment through situation updates, anticipating information needs, monitoring for errors, and offering assistance, thereby allowing the HAT to realize their full potential without impeding each other.

MeX seeks to abstract the bounded rationality that renders humans unable to mentally model the entirety of the AAI agent's logic and rationale. The abstraction reduces the effect of the bounded rationality and self-interest of human principals and stakeholders on the AAI agent's decision making and performance, preventing failures seen in similar mission-critical situations [26]. Although intermediaries in principal-agent relationships are well established [31], an intermediary of the kind proposed (Fig.1) is novel in HAT. The interaction between the human principals and the multiple AAI systems of varying functions and capabilities (Fig.1 labels 1, 2) is mediated by the non-human intermediary. Additionally, any interaction between human agents, such as fellow soldiers or fellow workers, and AAI (Fig.1 labels 3, 2), can also be mediated by

MeX. It also can mediate between the AAI and the supplying organization to run maintenance tasks. It can complement HAT interactions by providing access to stakeholders (Fig.1 labels 4, 2) for audits and status outputs that are trustworthy, unbiased, non-fearing and on-demand, even as the AAI systems go about their tasks.

4. Discussion and Conclusions

Examining HAT relationships in the new paradigm of delegated decision-making by multiple, highly autonomous systems, would warrant a paradigm shift to a macrocosmic view of the overall HAT environment, that goes beyond human principals attending to individual AAI agents. For example, in multi-agent, mission critical operations involving highly autonomous agents, the intermediary, itself an AI agent, would be able to collate and customize explanations, assurance statements and interpretations without being limited by a human's bounded rationality. Also, the relationships between human principals and human agents may be founded on personal trust, either dyadic or organizational [38], and may be seen to vary depending on social dynamics, group dynamics and other human factors [30]. In contrast, the shifted paradigm would require shifting from personal trust in individual AAI systems in the HAT, to an impersonal trust or institutional trust reposed in the specific HAT environment or operation as a whole, within the societal context. MeX would be able to provide a platform for non-fearing and non-judgmental [6, 15] and holistic explanations, interpretations and assurance statements.

At lower levels of LOA, MeX could raise concerns about adding another layer of non-human systems, as well as bringing up further questions of trust in that intermediary itself. The opportunities for MeX can gather importance only when decision-making in HAT shifts towards AAI agents with higher LOA and increased heterogeneity of agency. Further research in the form of case studies, in domains such as MCM, is needed to bring out the desirable features and the implementation specifics of MeX. Empirical testing of simulations of the non-human intermediary will be required to evaluate how holistic explanations, assurances and interpretations from MeX could help in fostering institutional trust. It also needs to be examined if incorporating features of similar cyber-physical intermediaries such as the 'artificial SMM' [21] or DST [34] in MeX could bring about significant impact in HAT performance and social legitimacy.

This conceptual paper contributes a novel means of examining the asymmetries of risk-exposure, information, and capabilities in human-autonomy team relationships in the social context. It incorporates the needs of principals, agents, organizations, society, and other stakeholders through the device of a non-human intermediary. The key contributions of this paper are i) outlining the intent, agency, and context in the HAT environment, ii) broadly identifying issues as human principal-AAI agent asymmetries, iii) extending the scope of HAT environment to include societal stakeholders, and iv) presenting opportunities for a Mediating Explainer (MeX) that would be less constrained by fear, self-interest and bounded-rationality.

References

[1] W. Xu, From automation to autonomy and autonomous vehicles: Challenges and opportunities for human-computer interaction, interactions 28 (2021). doi:10.1145/3434580.

[2] T. O'Neill, N. McNeese, A. Barron, B. Schelble, Human–autonomy teaming: A review and analysis of the empirical literature, Human Factors 64 (2022) 904–938. doi:10.1177/0018720820960865.
[3] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, eBook, Global Edition, Pearson Education, 2021.
[4] J. B. Lyons, S. A. Jessup, T. Q. Vo, The role of decision authority and stated social intent as predictors of trust in autonomous robots, Topics in cognitive science (2022). doi:10.1111/tops.12601.
[5] S. Nahavandi, Trusted autonomy between humans and robots: Toward human-on-the-loop in robotics and autonomous systems, MSMC 3 (2017) 10–17. doi:10.1109/MSMC.2016.2623867.
[6] J. B. Lyons, K. Sycara, M. Lewis, A. Capiola, Human–autonomy teaming: Definitions, debates, and directions, Frontiers in Psychology 12 (2021).
[7] E. Jaakkola, Designing conceptual articles: four approaches, AMS Review 10 (2020) 18–26. doi:10.1007/s13162-020-00161-0.
[8] S. P. Shapiro, The social control of impersonal trust, The American journal of sociology 93 (1987) 623–658. doi:10.1086/228791.
[9] N. A. Shankman, Reframing the debate between agency and stakeholder theories of the firm, Journal of Business Ethics 19 (1999) 319–334.
[10] S. Aureli, Social Legitimacy, Springer International Publishing, Cham, 2020, pp. 1–5. doi:10.1007/978-3-030-02006-4_678-1.
[11] F. Bridoux, J. Stoelhorst, Stakeholder theory, strategy, and organization: Past, present, and future, Strategic Organization 20 (2022) 797–809. doi:10.1177/14761270221127628.
[12] R. Kivits, S. Sawang, Stakeholder Theory, Springer International Publishing, 2021, pp. 1–8. doi:10.1007/978-3-030-70428-5_1.
[13] S. P. Shapiro, Agency theory, Annual Review of Sociology 31 (2005) 263–284. doi:10.1146/annurev.soc.31.041304.122159.
[14] K. M. Eisenhardt, Agency theory: An assessment and review, The Academy of Management Review 14 (1989) 57. doi:10.2307/258191.
[15] J. J. Van Den Hout, O. C. Davis, M. C. Weggeman, The conceptualization of team flow, The Journal of Psychology 152 (2018) 388–423. doi:10.1080/00223980.2018.1449729.
[16] L. Willett, Australian Navy Trials Autonomous Maritime Systems, Technical Report 02529793, Media Transasia India Ltd, 2019. URL: https://www.armadainternational.com/download/7005/ARM_1906_07.pdf.
[17] D. W. Boyles, Navy/marine corps team takes a new look at mcm, Marine Corps gazette 80 (1996) 32.
[18] R. Vine, E. Kohn, Concept for robotic and autonomous systems, 2020.
[19] M. C. Jensen, W. H. Meckling, Theory of the firm: Managerial behavior, agency costs and ownership structure, Journal of financial economics 3 (1976) 305–360. doi:10.1016/0304-405X(76)90026-X.
[20] V. Alonso, P. de la Puente, System transparency in shared autonomy: A mini review, Frontiers in Neurorobotics (2018). doi:10.3389/fnbot.2018.00083.
[21] R. W. Andrews, J. M. Lilly, D. Srivastava, K. M. Feigh, The role of shared mental models in human-ai teams: a theoretical review, Theoretical Issues in Ergonomics Science 24 (2023)

129–175. doi:10.1080/1463922x.2022.2061080.
[22] N. Staggers, A. F. Norcio, Mental models: concepts for human-computer interaction research, International Journal of Man-Machine Studies 38 (1993) 587–605. doi:10.1006/imms.1993.1028.
[23] D. Hadfield-Menell, The Principal-Agent Alignment Problem in Artificial Intelligence, Ph.D. thesis, 2021. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-207.html.
[24] L. Sanneman, J. A. Shah, The situation awareness framework for explainable ai (safe-ai) and human factors considerations for xai systems, International Journal of Human-Computer Interaction 38 (2022) 1772–1788. doi:10.1080/10447318.2022.2081282.
[25] A. Rosenfeld, A. Richardson, Explainability in human–agent systems, Autonomous Agents and Multi-Agent Systems 33 (2019) 673–705. doi:10.1007/s10458-019-09408-y.
[26] L. F. Cranor, A framework for reasoning about the human in the loop, 2008.
[27] M. Christensen, T. Knudsen, Design of decision-making organizations, Management Science 56 (2010) 71–89. doi:10.1287/mnsc.1090.1096.
[28] M. Harris, A. Raviv, Some results on incentive contracts with applications to education and employment, health insurance, and law enforcement, The American Economic Review 68 (1978) 20–30.
[29] C. Meske, E. Bunde, J. Schneider, M. Gersch, Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities, Information Systems Management 39 (2022) 53–63. doi:10.1080/10580530.2020.1849465.
[30] C. Deligianis, C. J. Stanton, C. McGarty, C. J. Stevens, The impact of intergroup bias on trust and approach behaviour towards a humanoid robot, J. Hum.-Robot Interact. 6 (2017) 4–20. doi:10.5898/JHRI.6.3.Deligianis.
[31] B. Van Der Meulen, New roles and strategies of a research council: intermediation of the principal-agent relationship, Science and Public Policy 30 (2003) 323–336. doi:10.3152/147154303781780344.
[32] T. Müller, S. Kamm, A. Löcklin, D. White, M. Mellinger, N. Jazdi, M. Weyrich, Architecture and knowledge modelling for self-organized reconfiguration management of cyber-physical production systems, International Journal of Computer Integrated Manufacturing 36 (2023) 1842–1863. doi:10.1080/0951192x.2022.2121425.
[33] N. Mohamed, J. Al-Jaroodi, S. Lazarova-Molnar, I. Jawhar, Middleware to Support Cyber-Physical Systems, 2016. doi:10.1109/PCCC.2016.7820605.
[34] T. Miller, Explainable ai is dead, long live explainable ai!: Hypothesis-driven decision support using evaluative ai, Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (2023).
[35] S. Padmanabhan Poti, C. J. Stanton, Enabling affordances for ai governance, Journal of Responsible Technology (2024). doi:10.1016/j.jrt.2024.100086.
[36] T. Bernhold, N. Wiesweg, Principal-Agent Theory: Perspectives and practices for effective workplace solutions, 2021, pp. pp. 117–128. doi:10.1201/9781003128786-10.
[37] L. Viljanen, Towards an ontology of trust, 2005. doi:10.1007/11537878_18.
[38] S. P. Shapiro, Standing in another's shoes: How agents make life-and-death decisions for their principals, Academy of Management Perspectives 30 (2016) 404–427. doi:10.5465/amp.2013.0158.