Preposition Salad: Placing Humans & AI in/on/over/along/under 'the-loop'

Aditya Singh, Zoe Szajnfarber
The George Washington University
asingh25@gwu.edu, zszajnfa@gwu.edu

Abstract

Academia, government, and industry have not clearly defined the mechanisms by which humans are expected to oversee or collaborate with AI-enabled systems. Inconsistent uses of terms like human-in-the-loop and human-AI teaming create confusion on what type of oversight and collaboration is intended. This paper presents a framework to clarify existing terminology and synthesizes disparate literature to identify existing mechanisms of human-AI interaction.

Introduction

The adoption of AI into many safety-critical industries such as transportation, defense, and medicine has created the need to ensure AI is safe and trustworthy. One of the most common prescriptions for achieving safe and trustworthy AI is to place a human-in-the-loop, which ideally provides "superior results, building trust by inserting human oversight into the AI life cycle" (Middleton et al. 2022). The promise of superior performance and decreased risk has encouraged policymakers to codify the need for a human to be 'in-the-loop' for many safety-critical systems (High-Level Expert Group on AI 2019) (China Academy of Information and Communications Technology and JD Explore Academy 2021) (Office of the Under Secretary of Defense for Policy 2023). Despite being codified as policy, it remains unclear exactly how humans should be partnered with or have oversight of AI systems. Without proper consideration of how humans and AI operate together, well-meaning policy is unlikely to achieve its intended outcomes. This paper explores the architectures of the loop--how tasks are allocated and coordinated among human and AI elements.

Framework

The literature on human AI teaming and calibrated trust is broad and distributed across disciplines including human factors, human computer interaction, and systems engineering. We provide a brief overview of the most relevant aspects related to architecture and on calibrated trust by the operator.

We drew from extensive literature from levels of automation/autonomy (Sheridan 1992) (Endsley 1987) (SAE International 2021) (Yang et al. 2017) (Proud, Hart, and Mrozinski 2003), human-computer interaction (Beer, Fisk, and Rogers 2014) (Save and Feuerberg 2012) (Draper 1995), and systems engineering (Lacher et al. 2023) (Samad 2020) (Samad 2023), as well as published case studies of human in the loop implementation (Hawley 2017) (De-Arteaga, Fogliato, and Chouldechova 2020), to create a synthesized a framework based on how the human or AI is integrated into the decision-loop. So far, research and policy has generally only defined two types of human-AI system architecture: human-in-the-loop and human-on-the-loop (Office of the Under Secretary of Defense for Policy 2023). Human-in-the-loop systems require the active involvement of both humans and AI for the system to function, while human-on-the-loop refers to an autonomous system that is monitored by a human. While these definitions provide some clarity, they mask much of the complexity and nuance of how humans and AI are expected to interact. The framework shown in Figure 1, characterizes this nuance by decomposing human-AI system architecture into 13 distinct architectures.

The framework is organized around a sequence of differentiating questions. At the highest level, Human-AI systems can be distinguished in terms of "whose" action is strictly, necessary: AI, Human or both. Starting on the left, architectures where both humans and AI must act are distinguished by *how* the human is required to act. **Human-AI team** architectures require both humans and autonomous agents to act, while in **human-in-the-loop** architectures, the human's only function in the system is to plan. Human-in-the-loop architectures can be further decomposed into **human selectors**, in which the human is selecting a plan of action for the AI system to take from a list of options or **human approvers**, in which humans approve (or reject) a singular option. For the other two cases, we distinguish between the optional party serving as a monitor or an assistant.

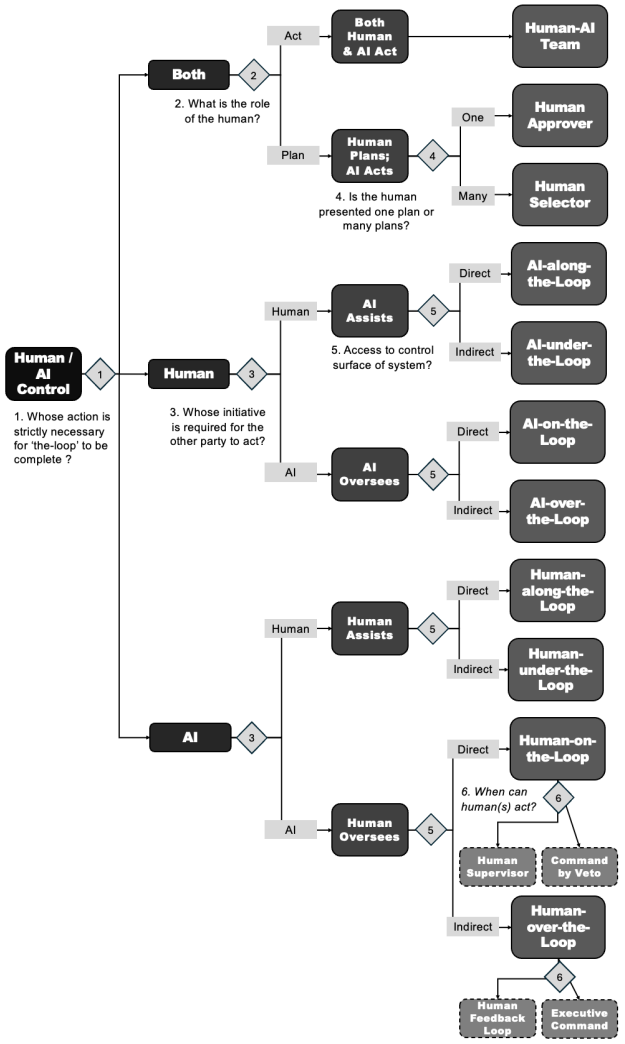

Figure 1: Human-AI Control Architecture Framework

Starting with humans serving as monitors for AI, the first distinction is the type of oversight action the party overseeing can enact. One option is indirect, in which the overseeing party can attempt to change the acting party's behavior (e.g. by providing additional information or warnings) but cannot directly change actions. Human indirect monitoring, which is sometimes called "**human-over-the-loop**" (Lacher et al. 2023), the key distinction is *when* the human asserts their influence. The options are during operation (by providing additional information the system may not have) through **executive command** or after operation (providing *post hoc* feedback that is used to update future instance of the system) as a **human feedback loop**.

The other option, direct oversight, allows the overseeing party to directly change how the system or acting party is operating and is often referred to as **human-on-the-loop** control. Direct human intervention can also be distinguished by *when* the human can assert their influence. A human can **command by veto**, preventing an AI system's planned course of action to take place before it is implemented, requiring the system to formulate a new plan, cease operation all together, or receive commands from a human. Otherwise, they can monitor the system as it is carrying out its operations and intervene when necessary, taking over control of the system temporarily, serving a true **human supervisor**.

AI can monitor humans in two ways: by directly intervening in system operation in an **AI-on-the-loop case** or by trying to influence the human's behavior in an **AI-over-the-loop case**. Humans and AI systems can both act as assistants, helping the other party directly by performing some task(s) as delegated in a **human/AI-along-the-loop** case or by providing information or guidance on a task in a **human/AI-under-the-loop** case. Assistants serve at the necessary discretion of the acting party, while monitors act when.

Implementation Examples

We provide examples of real systems that employ each architecture in Table 1. This is not meant to be an exhaustive list; rather, it is meant to illustrate how each architecture might be implemented.

Architecture	System	Implementation
Human Selector	Surgical Robots	Surgical robots generate "potential strategies for [the] surgeon to select" from, after which the "system takes over control to execute the selected plan" (Lee et al. 2024, 3).
Human Approver	Patriot Missile Defense System	"Semi-automatic mode: Human operator must authorize engagement or system will not fire" (Hawley 2017, 6)
Human-AI Team	U.S. Customs & Border Patrol Crossings	Facial recognition algorithms match travelers to their photos, helping catch imposters, while officers perform behavioral / contextual screening ("Say Hello to the New Face of Efficiency, Security and Safety," n.d.)
Human-under-the-loop	Waymo's autonomous vehicle support	Vehicle "can reach out to a human fleet response agent for additional information to contextualize its environment," but it does

		not give the agent control over the system and does not use the provided information as the sole source of information (The Waymo Team, n.d.).
Human-along-the-loop	Level 3 Driving Automation	Vehicle is driving itself but may temporarily handover control to deal with edge cases encountered on the road (SAE International 2021)
Human Feedback Loop	Gmail Spam Filter	Filters "are continuously updated with the emergence of state of the art tools, algorithms, discovery of new spam *and the feedback from Gmail users* about likely spammers" (Dada et al. 2019, 6)
Executive Command	RQ-4 Drone	"remote pilot can provide new guidance in terms of waypoints or heading but cannot directly provide inputs to the control surfaces" (Lacher et al. 2023, 2)
Human Supervisor	Uber ATG	"a human operator inside the vehicle was tasked with overseeing the system's operation, monitoring the driving environment, and if necessary, taking control of the vehicle and intervening in an emergency" (National Transportation Safety Board 2018, 8)
Command by Veto	Patriot Missile Defense System	"Automatic mode: System will fire unless human operator halts engagement" (Hawley 2017, 6)
AI-under-the-loop	En Route Air Traffic Organizer (ERATO)	ERATO is "activated on controller's initiative" and "helps the controller in identifying all the potential intruders of a given flight in the medium-short term" (Save and Feuerberg 2012, 52)
AI-along-the-loop	GitHub Co-Pilot	As the user types code or natural language comments, GitHub provides suggestions, but the user must affirm a suggestion for GitHub to act and is not obligated to do so ("About GitHub Copilot Individual," n.d.)
AI-over-the-loop	Traffic Alert & Collision Avoidance System (TCAS)	"in case of imminent risk of collision [TCAS] triggers visual and aural indications to the flight crew to perform an avoiding vertical maneuver. The execution of the maneuver itself is up to the flight crew" (Save and Feuerberg 2012, 51)
AI-on-the-loop	Airbus Auto Pilot/ Flight Director	If "Auto Pilot is engaged, the system initiates and executes a sequence of actions to fly the avoidance maneuvers until the aircraft is clear of conflict" while the "crew can monitor the sequence of actions … but cannot modify the ongoing action execution" (Save and Feuerberg 2012, 52)

Table 1: Examples of Architecture Implementations

Currently, most policy language directing humans to be included in-the-loop do not consider how exactly a human should partner with or have oversight over an AI-enabled system. This work has shown that there are several ways in which humans and AI can be partnered. Understanding the possible options helps us begin to understand the tradeoffs of each architecture to enable more precise policy that is more likely to achieve its intended goals. Future work will implement these architecture in a common setting to understand the tradeoff between the potential increase in performance from integrating AI and the potential increased risk profile of a system that incorporates AI.

Acknowledgments

This work was supported by the National Science Foundation under Grant No. 2125677.

References

Beer, J. M., A.D., Fisk and W.A. Rogers. 2014. Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction. *Journal of Human-Robot Interaction* 3 (2): 74–99. doi.org/10.5898/JHRI.3.2.Beer.

China Academy of Information and Communications Technology, and JD Explore Academy. 2021. White Paper on Trustworthy Artificial Intelligence. Beijing: Ministry of Industry and Information Technology.

Dada, E.G., J.S. Bassi, H. Chiroma, S.M. Abdulhamid, A.O. Adetunmbi, and O.E. Ajibuwa. 2019. Machine Learning for Email Spam Filtering: Review, Approaches and Open Research Problems. *Heliyon* 5 (6): e01802. doi.org/10.1016/j.heliyon.2019.e01802.

De-Arteaga, M., R. Fogliato, and A. Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–12. CHI '20. New York: Association for Computing Machinery. doi.org/10.1145/3313831.3376638.

Draper, J.V. 1995. Teleoperators for Advanced Manufacturing: Applications and Human Factors Challenges. *International Journal of Human Factors in Manufacturing* 5 (1): 53–85. doi.org/10.1002/hfm.4530050105.

Endsley, M. R. 1987. The Application of Human Factors to the Development of Expert Systems for Advanced Cockpits. In Proceedings of the Human Factors Society Annual Meeting. New York: Sage. doi.org/10.1177/154193128703101219.

GitHub. 2024. About GitHub Copilot Individual https://docs.github.com/en/copilot/copilot-individual/about-github-copilot-individual.

Hawley, J. 2017. Patriot Wars. Washington, DC: Center for a New American Security.

High-Level Expert Group on AI. 2019. Independent High-Level Expert Group on Artificial Intelligence. B-1049. Brussels: European Commission.

Lacher, A. R., L. Ren, D. R. Maroney, C. Schulenberg, and J. Daniels. 2023. Dimensional Role Analysis: The Role of Humans and Automation for Increasingly Autonomous Aviation Systems. In 2023 Integrated Communication, Navigation and Surveillance Conference *(ICNS).* Hendon: Integrated Communication, Navigation and Surveillance Conference. doi.org/10.1109/ICNS58246.2023.10124260.

Lee, A., T.S. Baker, J.B. Bederson, and B.I. Rapoport. 2024. Levels of Autonomy in FDA-Cleared Surgical Robots: A Systematic Review. *Npj Digital Medicine* 7 (1): 1–8. doi.org/10.1038/s41746-024-01102-y.

Association for Computing Machinery. 2022. Trust, Regulation, and Human-in-the-Loop AI: Within the European Region.

National Transportation Safety Board. 2018. Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona. Highway Accident Report. NTSB/HAR-19/03 PB2019-101402. Washington, D.C.: National Transportation Safety Board.

Office of the Under Secretary of Defense for Policy. 2023. DoD Directives. Autonomy in Weapon Systems. Vol. 3009.09.

Proud, R. W, J. J Hart, and R. B. Mrozinski. 2003. Methods for Determining the Level of Autonomy to Design into a Human Spaceflight Vehicle: A Function Specific Approach. Houston, Texas: National Aeronautics and Space Administration.

SAE International. 2021. Surface Vehicle Recommended Practice. J3016. SAE International.

Samad, T. 2020. Human-in-the-Loop Control: Applications and Categorization. In 3rd IFAC Workshop on Cyber-Physical & Human Systems. Beijing: International Federation of Automatic Control. doi.org/10.1016/j.ifacol.2021.04.108.

Samad, T. 2023. Human-in-the-Loop Control and Cyber–Physical–Human Systems: Applications and Categorization In *Cyber–Physical–Human Systems*, edited by A. M. Annaswamy, P. P. Khargonekar, F. Lamnabhi-Lagarrigue, and S. K. Spurgeon, 1–23. John Wiley & Sons, Ltd. doi.org/10.1002/9781119857433.ch1.

Save, Luca, and Beatrice Feuerberg. 2012. Designing Human-Automation Interaction: A New Level of Automation Taxonomy. In *Proceedings of Human Factors of Systems and Technology*. Toulouse, France: Human Factors and Ergonomics Society.

Customs and Border Protection. 2024. Say Hello to the New Face of Efficiency, Security and Safety. https://www.cbp.gov/travel/biometrics.

Sheridan, T.B. 1992. *Telerobotics, Automation, and Human Supervisory Control*. Cambridge, MA, USA: MIT Press.

Waymo. 2024. Fleet Response: Lending a Helpful Hand to Waymo's Autonomously Driven Vehicles. https://waymo.com/blog/2024/05/fleet-response.

Yang, G., J. Cambias, K. Cleary, E. Daimler, J. Drake, P.E. Dupont, N. Hata, et al. 2017. Medical Robotics—Regulatory, Ethical, and Legal Considerations for Increasing Levels of Autonomy. *Science Robotics* 2 (4): eaam8638. doi.org/10.1126/scirobotics.aam8638.