THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

SYSTEMS ENGINEERING RESEARCH CENTER

What is Human-in-the-Loop, Really?

Aditya Singh, PhD Candidate & Fellow, Co-Design of Trustworthy AI Systems
Zoe Szajnfarber, Professor & Director of Strategic Initiatives

SE4AI Conference | September 28, 2023 | Washington, DC

Human-in-the-Loop (HITL)

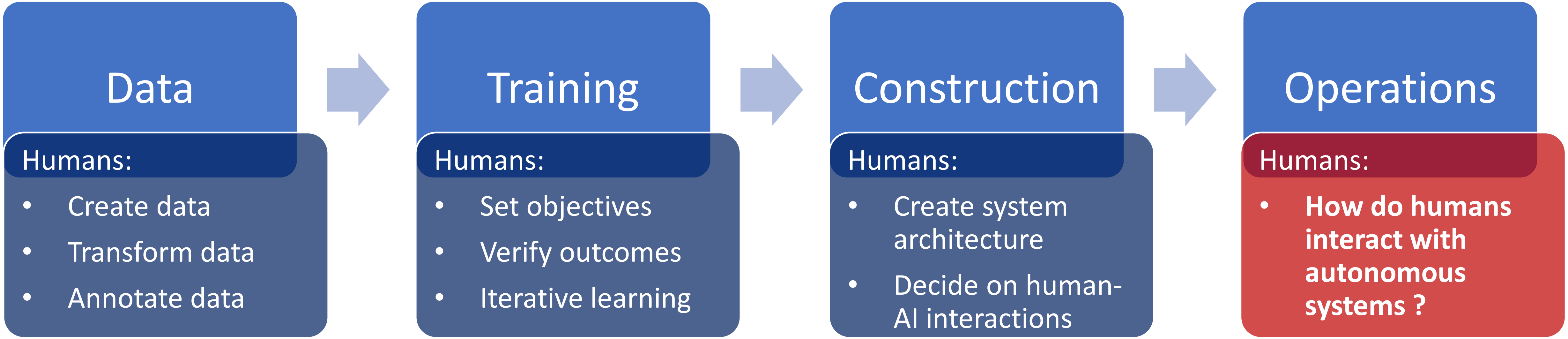
- HITL refers to a broad set of **architectures involving humans and autonomous agents interacting** to complete a task

Why use HITL?

- HITL is “grounded in the belief that human-machine teams offer superior results, **building trust by inserting human oversight into the AI life cycle**”

  (Middleton et al. 2022)

Literature Review: Humans in the AI Lifecycle



Data

Humans:

- Create data
- Transform data
- Annotate data

Training

Humans:

- Set objectives
- Verify outcomes
- Iterative learning

Construction

Humans:

- Create system architecture
- Decide on human-AI interactions

Operations

Humans:

- **How do humans interact with autonomous systems ?**

(Wu et al. 2022)

Humans are involved in the AI system development cycle in different ways

Motivation: HITL Means too Many Things

Autonomous Vehicles

(Huang et al. 2021)

Missile Defense

(Singer 2009)

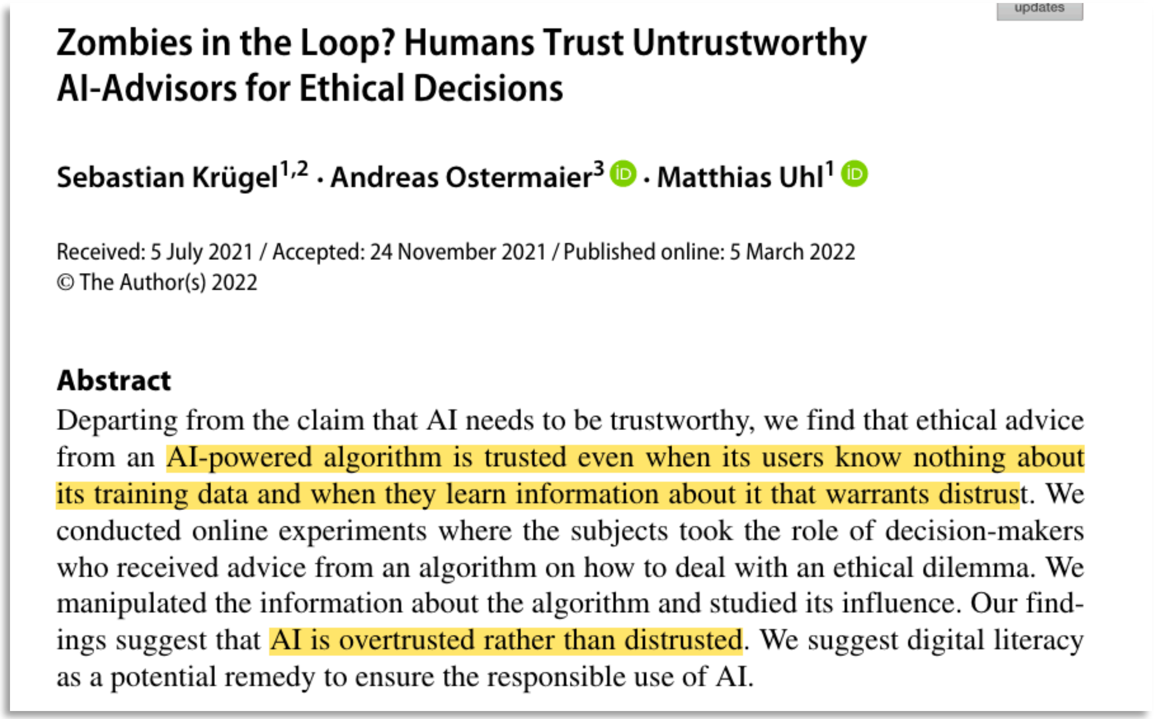
Email Filter

(Middleton et al. 2022)

HITL has been used to describe very different system architectures, creating confusion

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Feasibility of Oversight



Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions

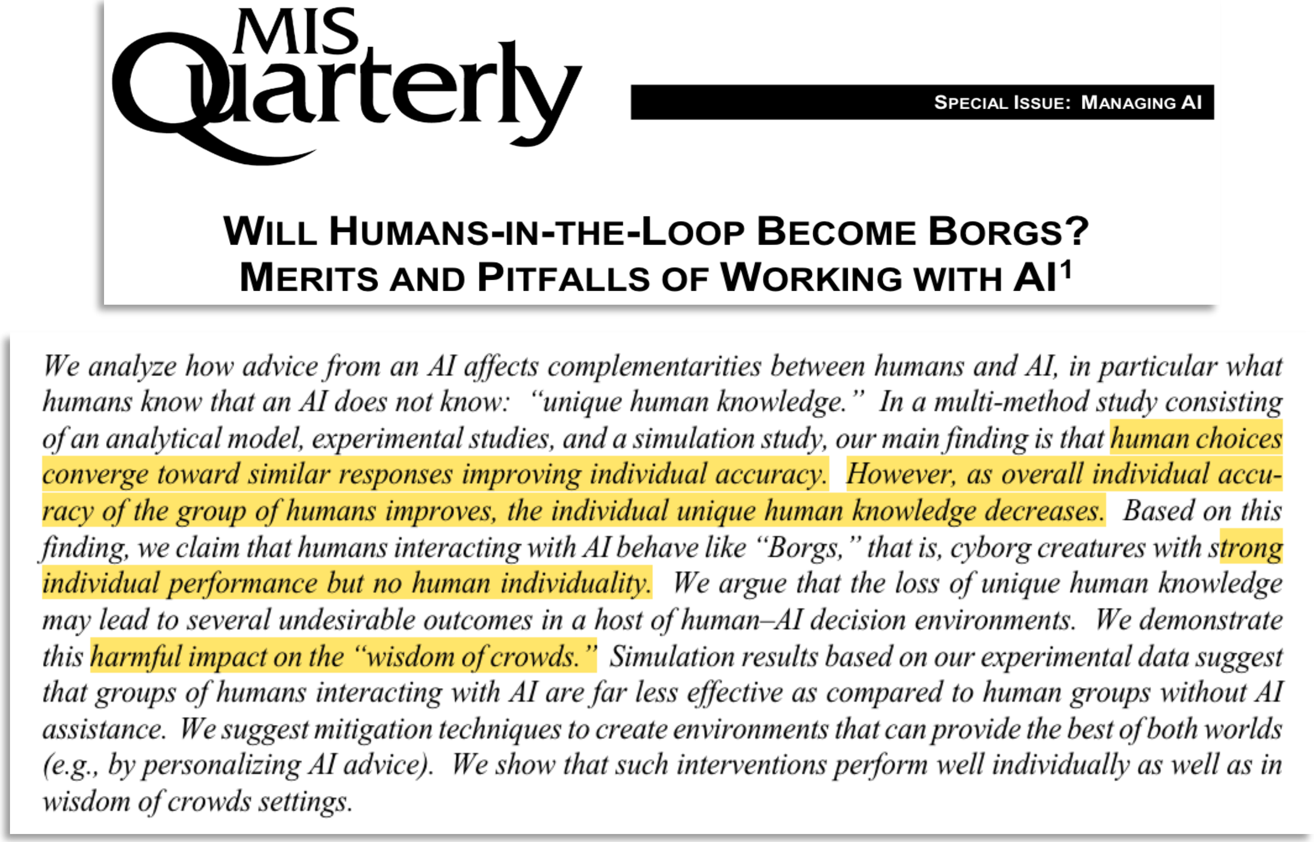
Sebastian Krügel[1,2] · Andreas Ostermaier[3] · Matthias Uhl[1]

Received: 5 July 2021 / Accepted: 24 November 2021 / Published online: 5 March 2022

© The Author(s) 2022

Abstract

Departing from the claim that AI needs to be trustworthy, we find that ethical advice from an AI-powered algorithm is trusted even when its users know nothing about its training data and when they learn information about it that warrants distrust. We conducted online experiments where the subjects took the role of decision-makers who received advice from an algorithm on how to deal with an ethical dilemma. We manipulated the information about the algorithm and studied its influence. Our findings suggest that AI is overtrusted rather than distrusted. We suggest digital literacy as a potential remedy to ensure the responsible use of AI.



MIS Quarterly

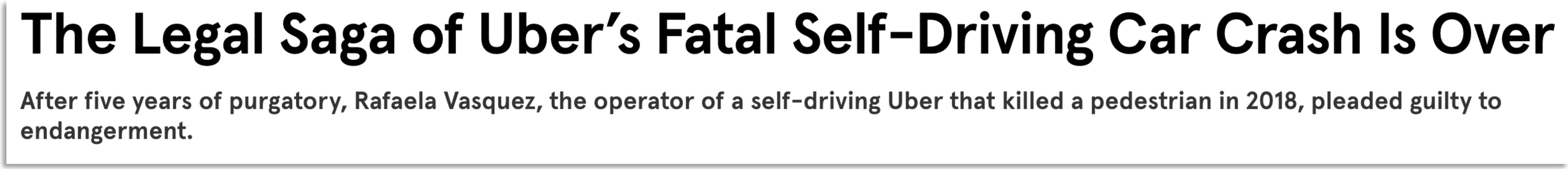
Special Issue: Managing AI

Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI[1]

We analyze how advice from an AI affects complementarities between humans and AI, in particular what humans know that an AI does not know: "unique human knowledge." In a multi-method study consisting of an analytical model, experimental studies, and a simulation study, our main finding is that human choices converge toward similar responses improving individual accuracy. However, as overall individual accuracy of the group of humans improves, the individual unique human knowledge decreases. Based on this finding, we claim that humans interacting with AI behave like "Borgs," that is, cyborg creatures with strong individual performance but no human individuality. We argue that the loss of unique human knowledge may lead to several undesirable outcomes in a host of human–AI decision environments. We demonstrate this harmful impact on the "wisdom of crowds." Simulation results based on our experimental data suggest that groups of humans interacting with AI are far less effective as compared to human groups without AI assistance. We suggest mitigation techniques to create environments that can provide the best of both worlds (e.g., by personalizing AI advice). We show that such interventions perform well individually as well as in wisdom of crowds settings.

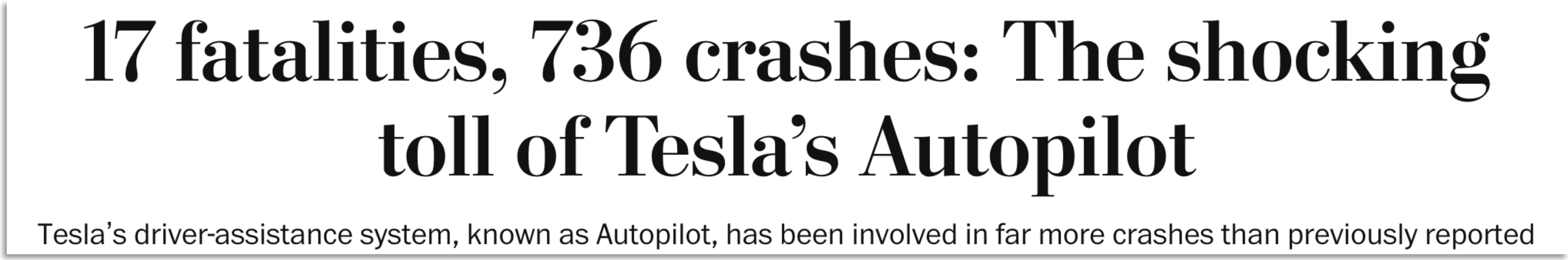
Academic literature suggests humans may be unable to perform the way we expect 'in-the-loop'

Feasibility of Oversight



The Legal Saga of Uber's Fatal Self-Driving Car Crash Is Over

After five years of purgatory, Rafaela Vasquez, the operator of a self-driving Uber that killed a pedestrian in 2018, pleaded guilty to endangerment.

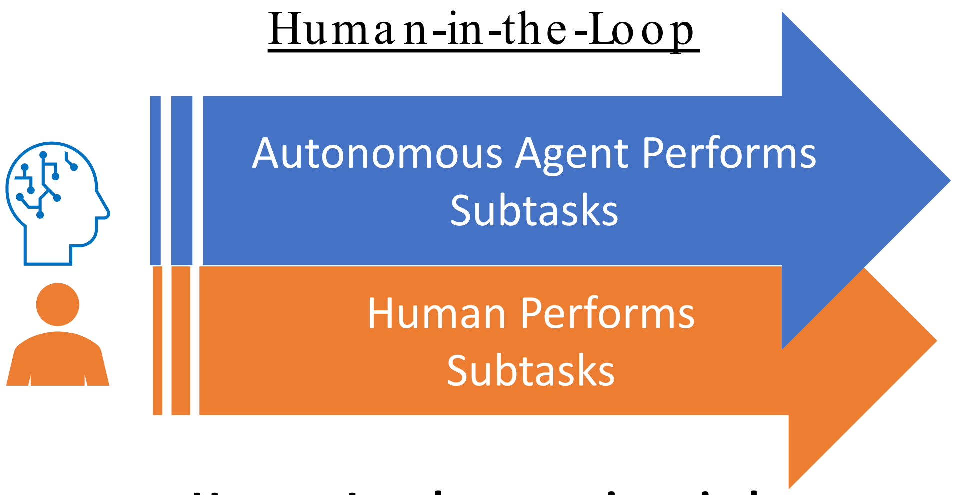


17 fatalities, 736 crashes: The shocking toll of Tesla's Autopilot

Tesla's driver-assistance system, known as Autopilot, has been involved in far more crashes than previously reported

Empirical cases similarly suggest an inability to perform in-the-loop responsibilities

Literature Review: *in* vs *on*-the-loop



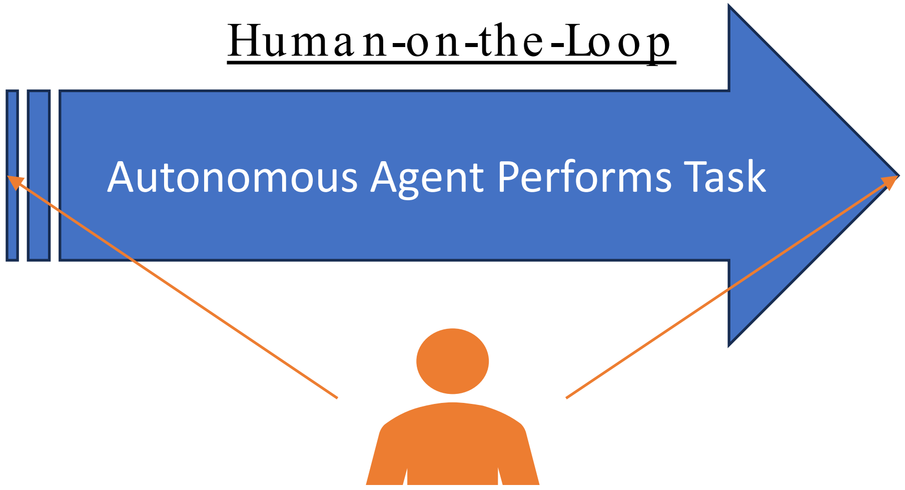
Human-in-the-Loop

Autonomous Agent Performs Subtasks

Human Performs Subtasks

Human Involvement is strictly Necessary To Complete Task

(Meng, 2023)



Human-on-the-Loop

Autonomous Agent Performs Task

Human is only a supervisor; human involvement is not strictly necessary

(Amori, 2023)

Literature has yet to meaningfully distinguish architecture beyond in and on the loop

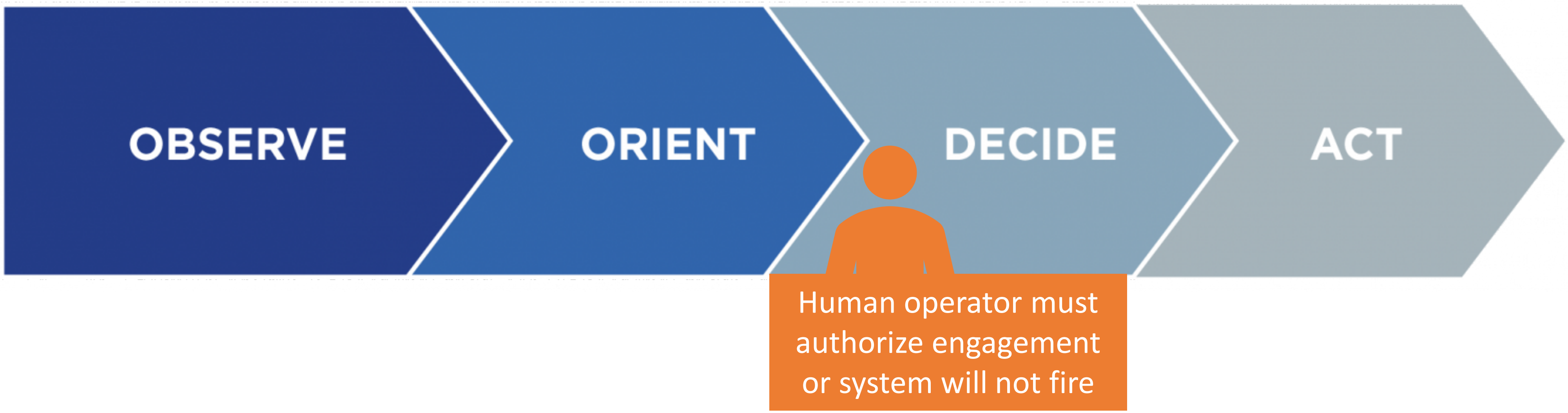
Research Goal

Characterize the space of potential architectures in which humans and autonomous agents work together

Methods

- Review empirical cases of HITL architecture across key application areas until empirical saturation
  - U.S. Missile Defense Systems
  - Autonomous Vehicles
  - Driver Assist Technologies
  - Email
  - AI Assistants
  - Border Patrol Facial Recognition
- Examine task decomposition between humans and autonomous agents and the interactions between them

Example: Patriot Semi-Automatic Mode

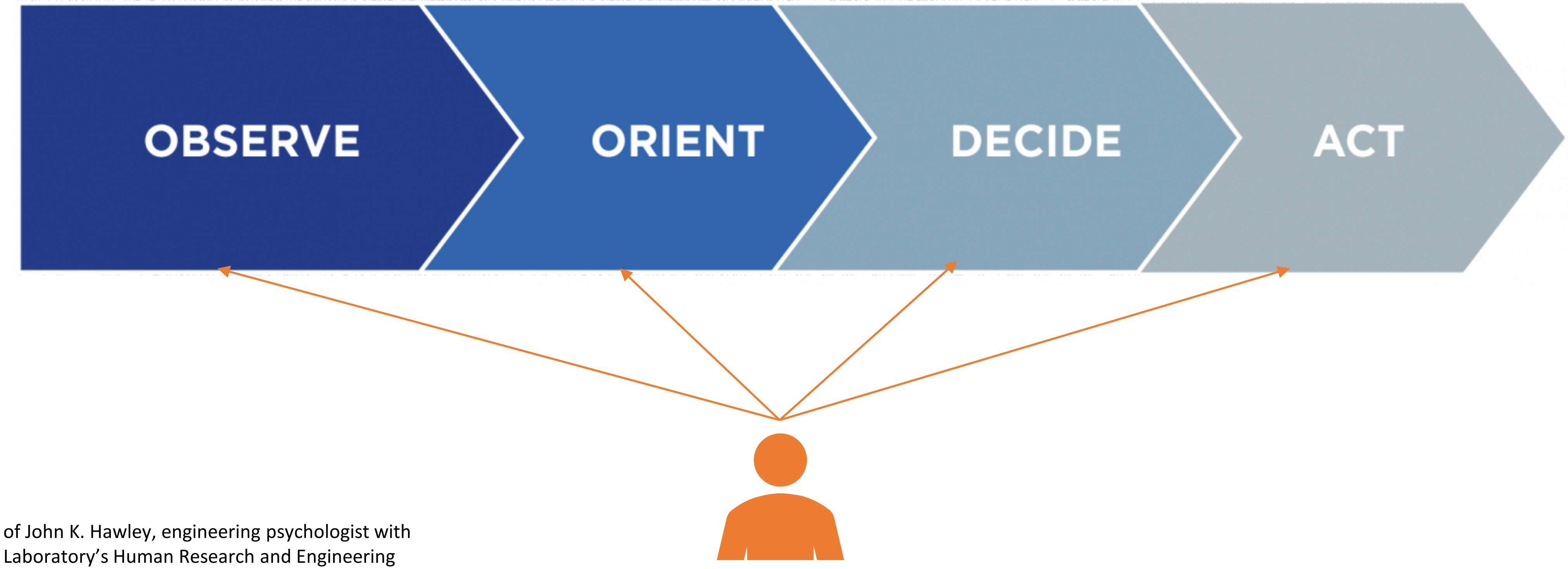


Human operator must authorize engagement or system will not fire

Adapted from the work of John K. Hawley, engineering psychologist with the U.S. Army Research Laboratory's Human Research and Engineering Directorate and principal investigator for an Army effort to examine the human role in Patriot fratricides during the Iraq War.

Human-in-the-loop; Human must approve otherwise action is not taken

Example: Patriot Automatic Mode



OBSERVE

ORIENT

DECIDE

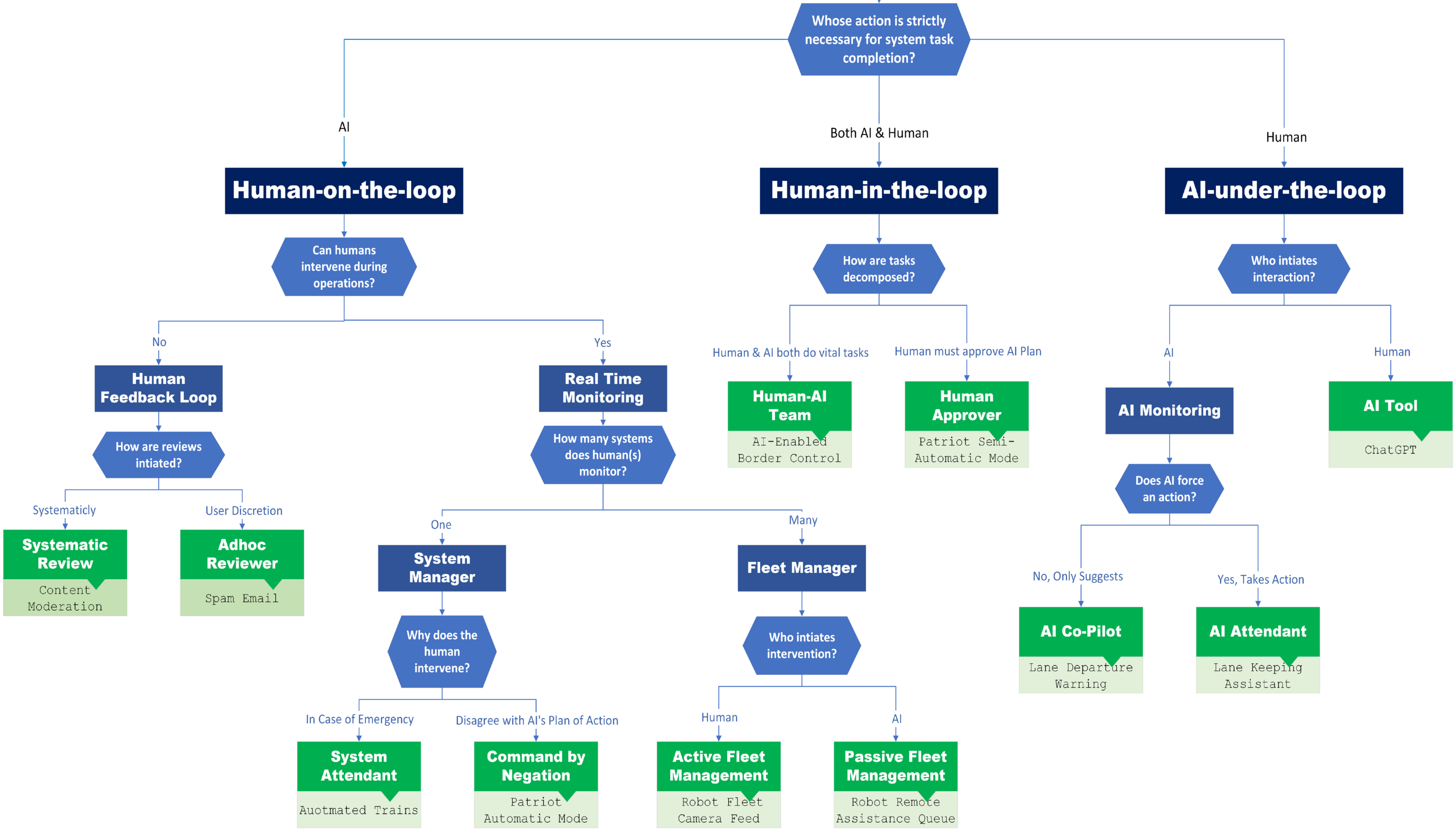
ACT

Oversight Over Cycle

Adapted from the work of John K. Hawley, engineering psychologist with the U.S. Army Research Laboratory’s Human Research and Engineering Directorate and principal investigator for an Army effort to examine the human role in Patriot fratricides during the Iraq War.

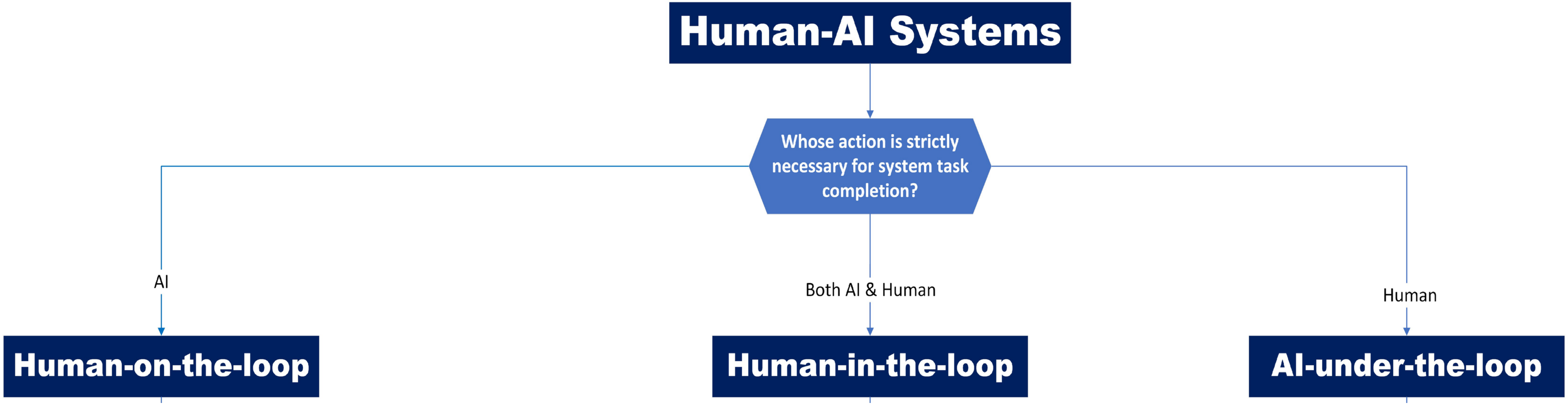
Human-on-the-loop; Human Can Override, System Automatically Proceeds Otherwise

Human-AI Systems

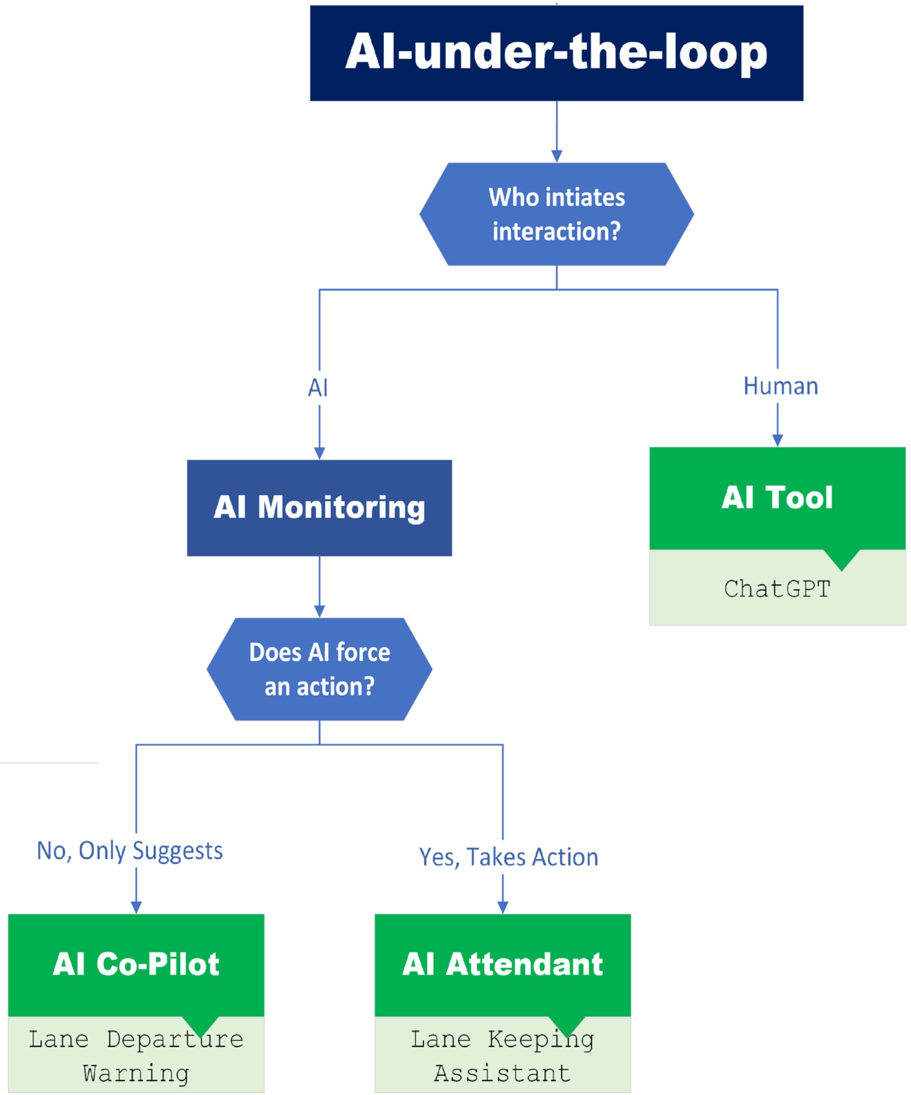


- Whose action is strictly necessary for system task completion?
  - AI → **Human-on-the-loop**
    - Can humans intervene during operations?
      - No → **Human Feedback Loop**
        - How are reviews intiated?
          - Systematicly → **Systematic Review** — Content Moderation
          - User Discretion → **Adhoc Reviewer** — Spam Email
      - Yes → **Real Time Monitoring**
        - How many systems does human(s) monitor?
          - One → **System Manager**
            - Why does the human intervene?
              - In Case of Emergency → **System Attendant** — Auotmated Trains
              - Disagree with AI's Plan of Action → **Command by Negation** — Patriot Automatic Mode
          - Many → **Fleet Manager**
            - Who intiates intervention?
              - Human → **Active Fleet Management** — Robot Fleet Camera Feed
              - AI → **Passive Fleet Management** — Robot Remote Assistance Queue
  - Both AI & Human → **Human-in-the-loop**
    - How are tasks decomposed?
      - Human & AI both do vital tasks → **Human-AI Team** — AI-Enabled Border Control
      - Human must approve AI Plan → **Human Approver** — Patriot Semi-Automatic Mode
  - Human → **AI-under-the-loop**
    - Who intiates interaction?
      - AI → **AI Monitoring**
        - Does AI force an action?
          - No, Only Suggests → **AI Co-Pilot** — Lane Departure Warning
          - Yes, Takes Action → **AI Attendant** — Lane Keeping Assistant
      - Human → **AI Tool** — ChatGPT

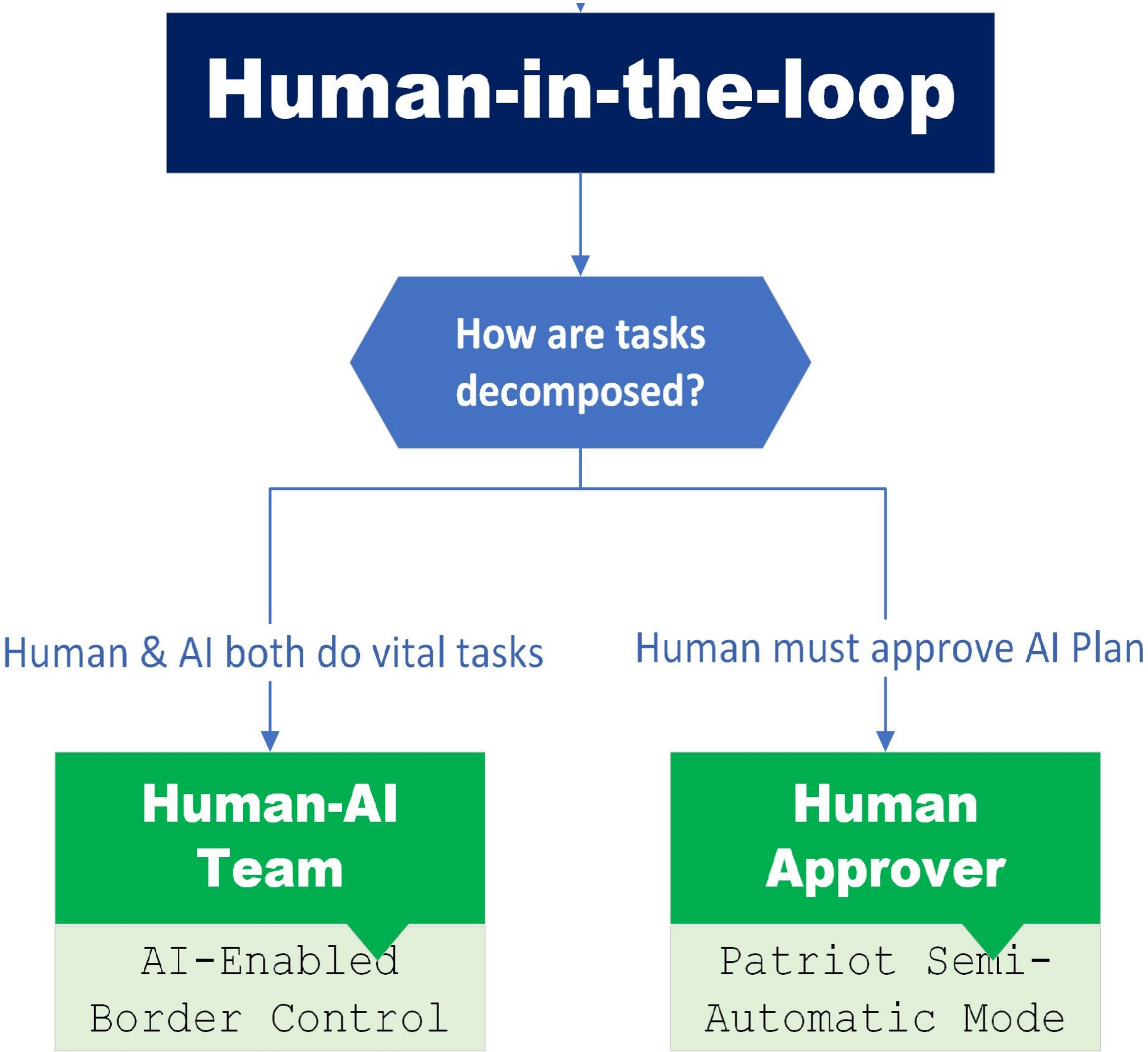
The Framework



in vs on is a matter of human involvement, but AI-under-the-loop tackles AI involvement



In these systems, we see AI playing the role humans usually play in HOTL systems



Human-in-the-loop

How are tasks decomposed?

Human & AI both do vital tasks

Human must approve AI Plan

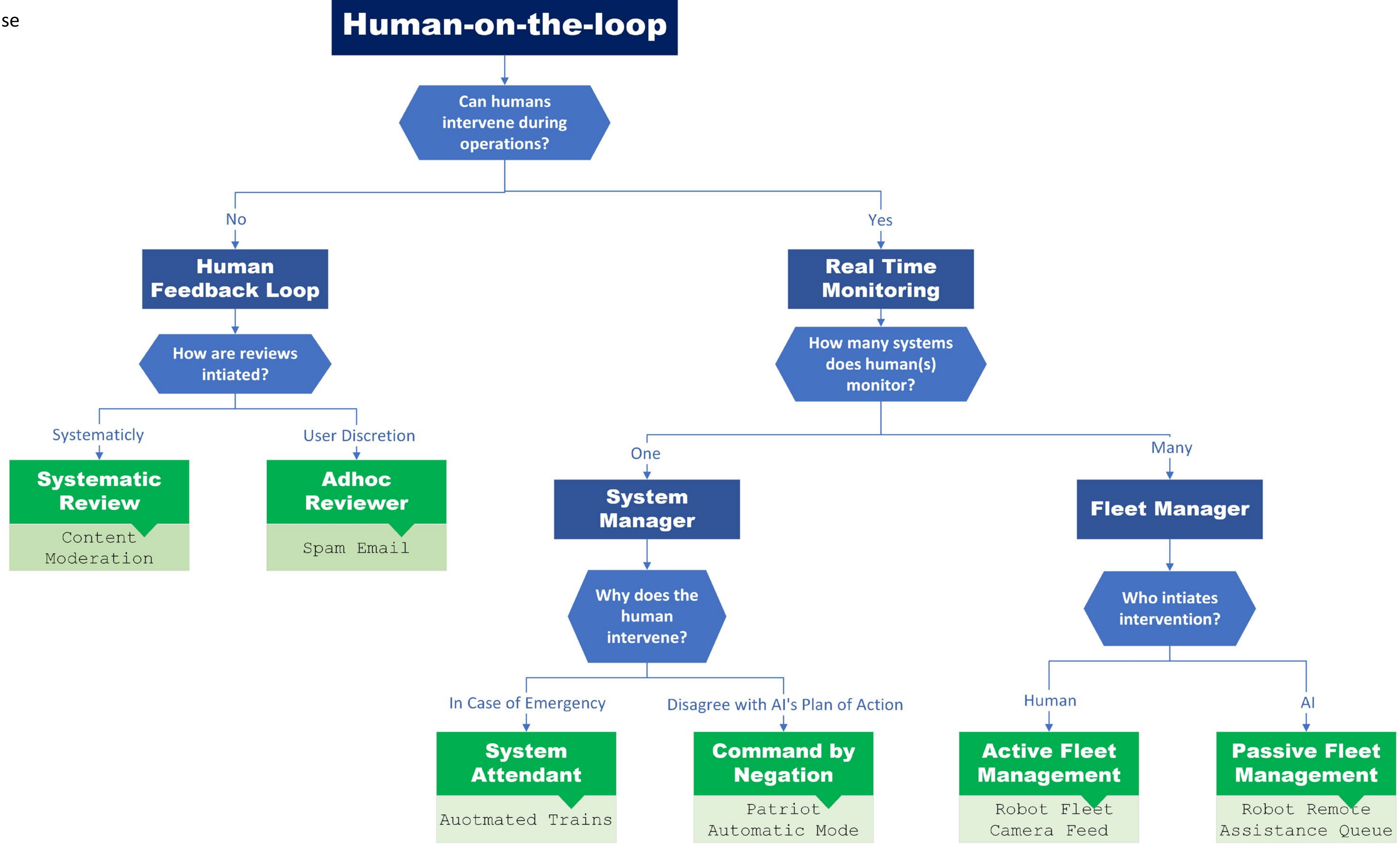
Human-AI Team

AI-Enabled Border Control

Human Approver

Patriot Semi-Automatic Mode

Broadly, HITL involves meaningful teaming or just implementing humans as approvers



Human-on-the-loop

Can humans intervene during operations?

No

Human Feedback Loop

How are reviews intiated?

Systematicly

Systematic Review

Content Moderation

User Discretion

Adhoc Reviewer

Spam Email

Yes

Real Time Monitoring

How many systems does human(s) monitor?

One

System Manager

Why does the human intervene?

In Case of Emergency

System Attendant

Auotmated Trains

Disagree with AI's Plan of Action

Command by Negation

Patriot Automatic Mode

Many

Fleet Manager

Who intiates intervention?

Human

Active Fleet Management

Robot Fleet Camera Feed

AI

Passive Fleet Management

Robot Remote Assistance Queue

There are several ways systems can be architected to utilize human oversight

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Conclusion & Future Work

- This study provides an **integrative framework for disparate efforts to characterize HAI teaming** & enables a better understanding the **expectations on the humans in the loop**
- Next steps are to analyze the strengths and weaknesses of each human-AI architecture

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

References

Middleton, Stuart, Emmanuel Letouzé, Ali Hossaini, and Adriane Chapman. 2022. “Trust, Regulation, and Human-in-the-Loop AI: Within the European Region.” *Communications of the ACM*, April, 64–68.

Wu, Xingjiao, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. “A Survey of Human-in-the-Loop for Machine Learning.” *Future Generation Computer Systems* 135:364–81. doi: 10.1016/j.future.2022.05.014.

Huang, Mengzhe, Zhong-Ping Jiang, Michael Malisoff, and Leilei Cui. 2021. “Robust Autonomous Driving with Human in the Loop.” Pp. 673–92 in *Handbook of Reinforcement Learning and Control*, *Studies in Systems, Decision and Control*, edited by K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever. Cham: Springer International Publishing.

Middleton, Stuart, Emmanuel Letouzé, Ali Hossaini, and Adriane Chapman. 2022. “Trust, Regulation, and Human-in-the-Loop AI: Within the European Region.” *Communications of the ACM*, April, 64–68.

S. Krügel, A. Ostermaier, and M. Uhl, “Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions,” *Philos. Technol.*, vol. 35, no. 1, p. 17, Mar. 2022, doi: 10.1007/s13347-022-00511-9.

A.Fügener, J. Grahl, A. Gupta, and W. Ketter, “Will Humans-in-The-Loop Become Borgs? Merits and Pitfalls of Working with AI.” Rochester, NY, Jul. 04, 2021. Accessed: Sep. 19, 2023. [Online]. Available: https://papers.ssrn.com/abstract=3879937

https://www.wired.com/story/ubers-fatal-self-driving-car-crash-saga-over-operator-avoids-prison/

https://www.washingtonpost.com/technology/2023/06/10/tesla-autopilot-crashes-elon-musk/

Meng, Xiao-Li. 2023. “Data Science and Engineering With Human in the Loop, Behind the Loop, and Above the Loop.” Harvard Data Science Review 5(2). doi: 10.1162/99608f92.68a012eb.

Amori, Michael. 2023. “Council Post: Our Place In The AI Loop: 3 Steps To Creating The Infrastructure For Responsible AI.” Forbes. Retrieved September 13, 2023(https://www.forbes.com/sites/forbestechcouncil/2023/05/03/our-place-in-the-ai-loop-3-steps-to-creating-the-infrastructure-for-responsible-ai/).

J. Hawley, “Patriot Wars,” Center for a New American Secuirty, Washington, DC, Jan. 2017. Accessed: May 01, 2023. [Online]. Available: https://www.cnas.org/publications/reports/patriot-wars