Ethical Human-AI Agent Interface Considerations

Clayton W. Couch
Michael E. Miller, PhD
Professor of Systems Engineering
Systems Engineering and Management

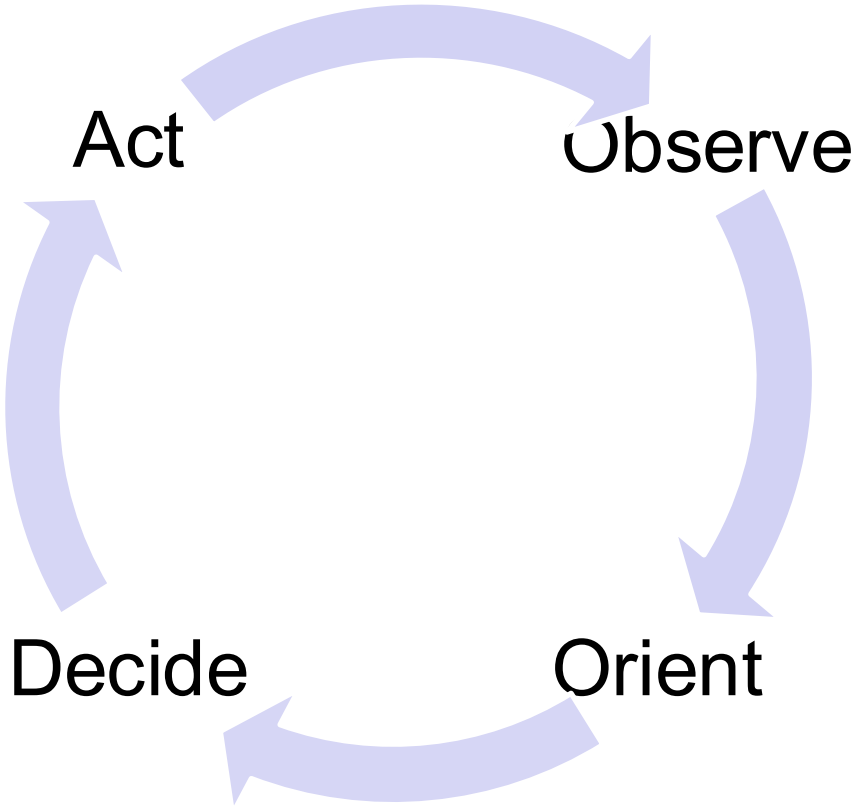
July 30, 2025

Disclosure

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the U.S. Air Force, U.S. Department of Defense, nor the U.S. Government.

AI Advantage

Goal Driven Behavior



- Humans and AI-Agents each pursue goal driven behavior
- Humans can require seconds to minutes to complete this loop
- AI Agents have the potential to close this loop much faster than humans

Use of AI as a Decision Aid

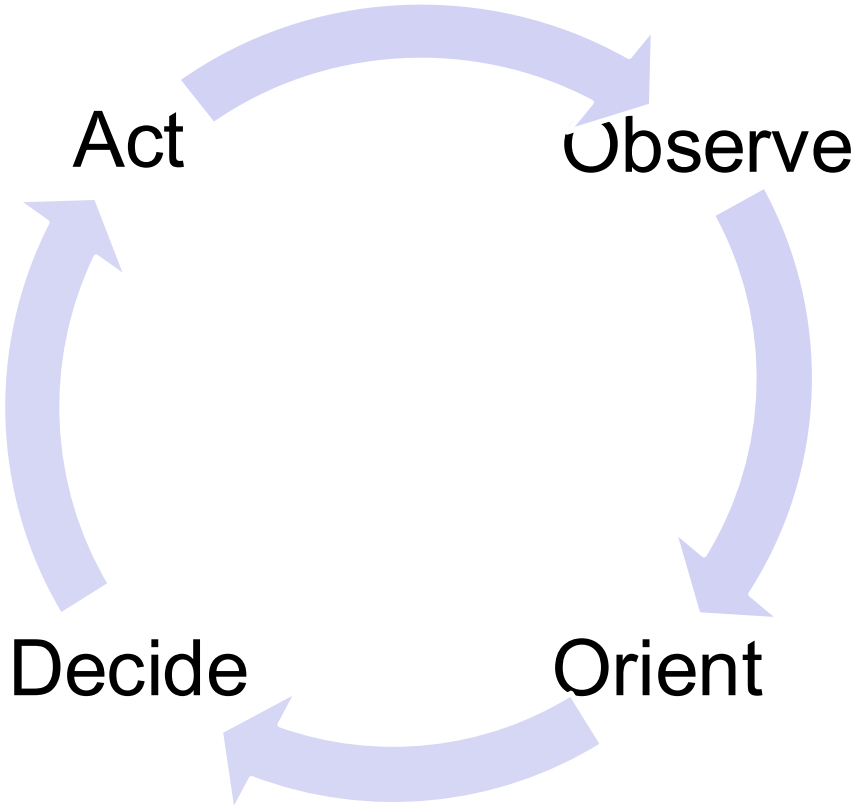
- Current legal and ethical frameworks fail to clearly define accountability for AI systems when the decision is not consistent with legal or social norms
- AI agents can prioritize task achievement while disregarding moral or ethical implications
- Often respond to a limited set of factors, contextual factors may change or reduce the importance of these factors
- AI agents may help mitigate human-factors induced decision errors, due to bias or limited processing power

Can we Overcome Issue with Teams?

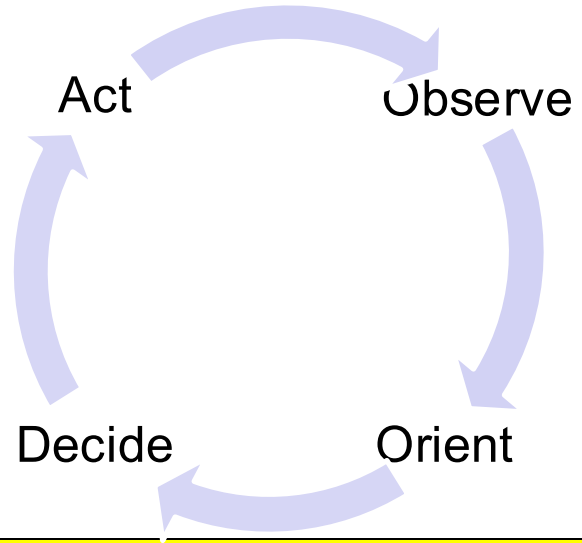
The AFIT of Today is the Air Force of Tomorrow.

Is Teaming the Solution?

Human Loop



AI Agent Loop



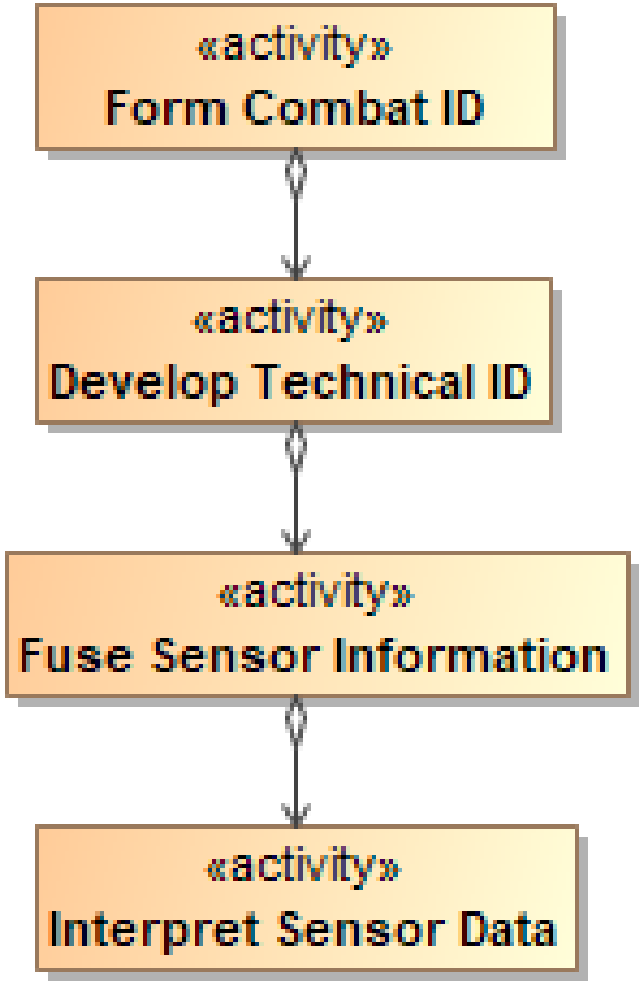
If the AI Agent Loop is faster, can the human be actively engaged in the decision?

Threats to Human Autonomy in HATs

- Inadequate Time
- Lack of Expertise
- Lack of or Inability to Gain Situation Awareness
- Lack of Mental Capacity
- Improper Trust Calibration for AI agents
- Inability to Command (e.g., missing controls)
- Bias

Decisions Do Not Occur in a Vacuum

- Multidimensional
- Exist in a Time Continuum
- Occur with Incomplete Information
- Occur with Misleading Information



Human Decision Architectures

In the Loop

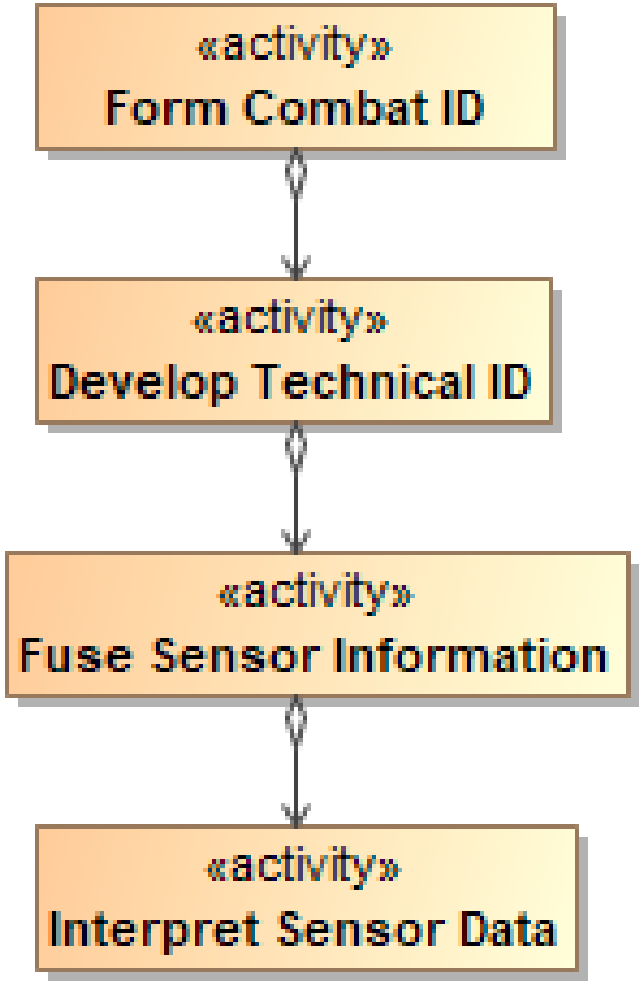
- Human participates in each phase of the perceptual or OODA loop
- Conducts work in parallel with the AI agent
- Can include assessment of whether self-derived decision or AI-derived decision is appropriate

On the Loop

- Human observes at the decision with potential to intervene in AI's action
- Human may observe some but not all environmental inputs and may partially orient, but not required
- Human plays a supervisory role, but may not autonomously form a decision or initiate an action

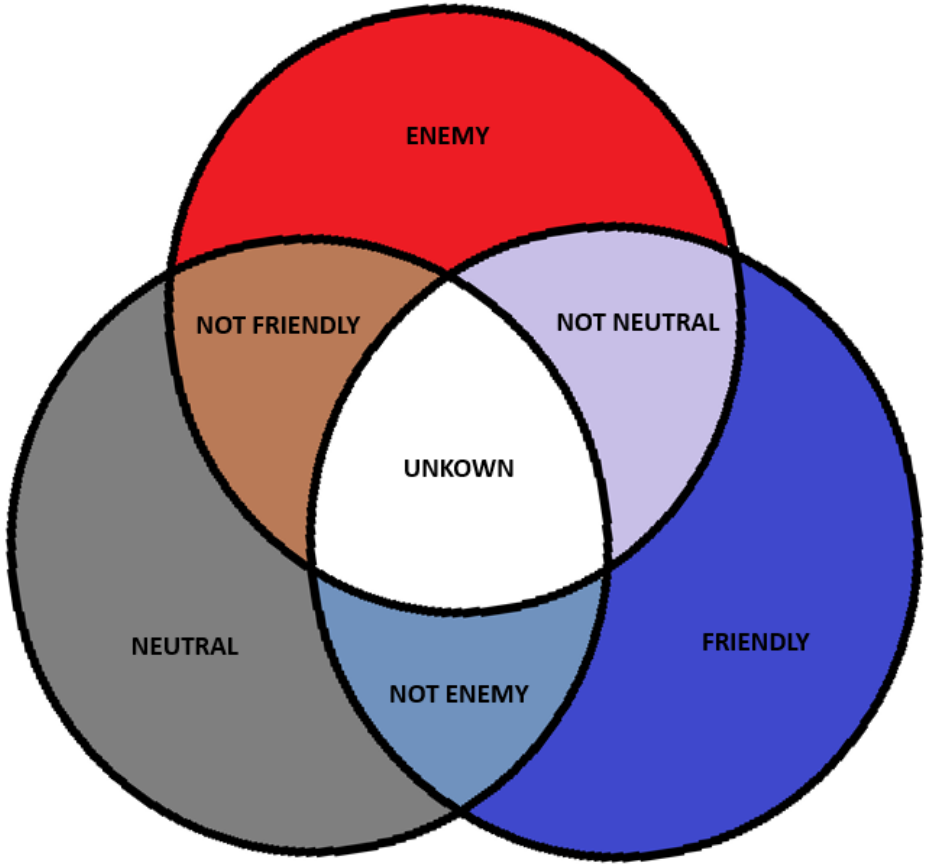
Decisions Do Not Occur in a Vacuum

- Multidimensional
- Exist in a Time Continuum
- Occur with Incomplete Information
- Occur with Misleading Information



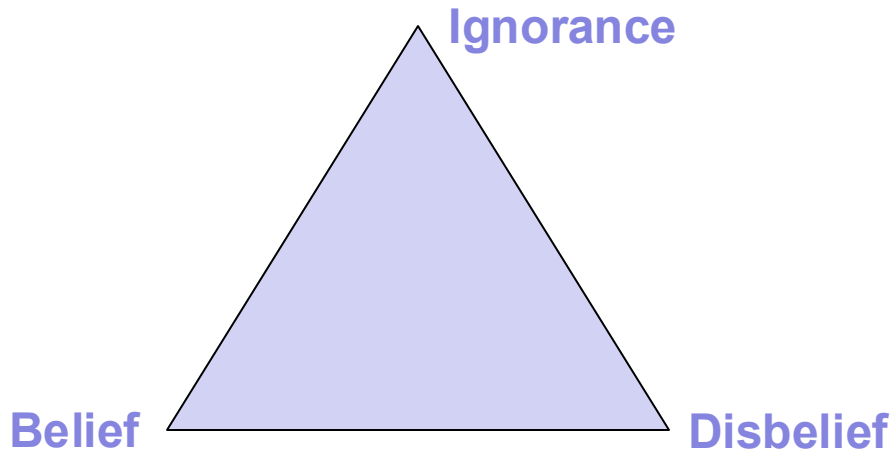
Logic for CID

Options for CID



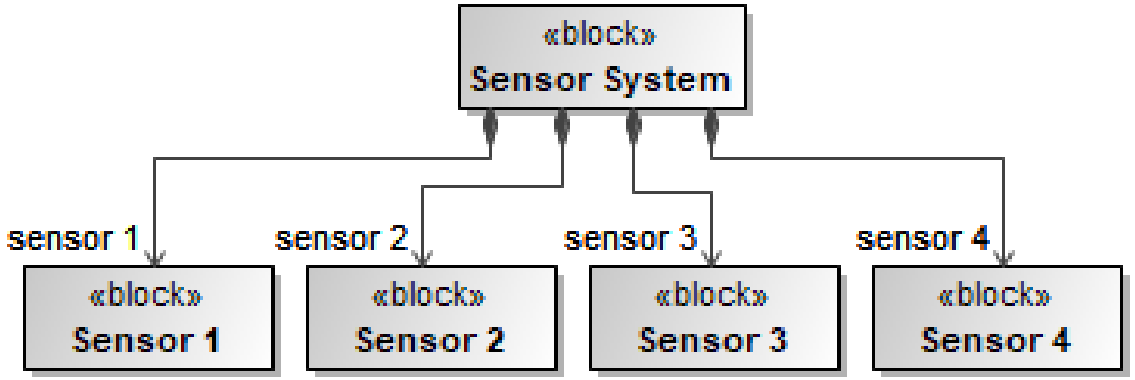
Subjective Logic

- Form a Hypothesis
  - Supported by Information
  - Alternate Supported by Information
  - Ignorance (Incomplete Information)



System Configuration

Structure

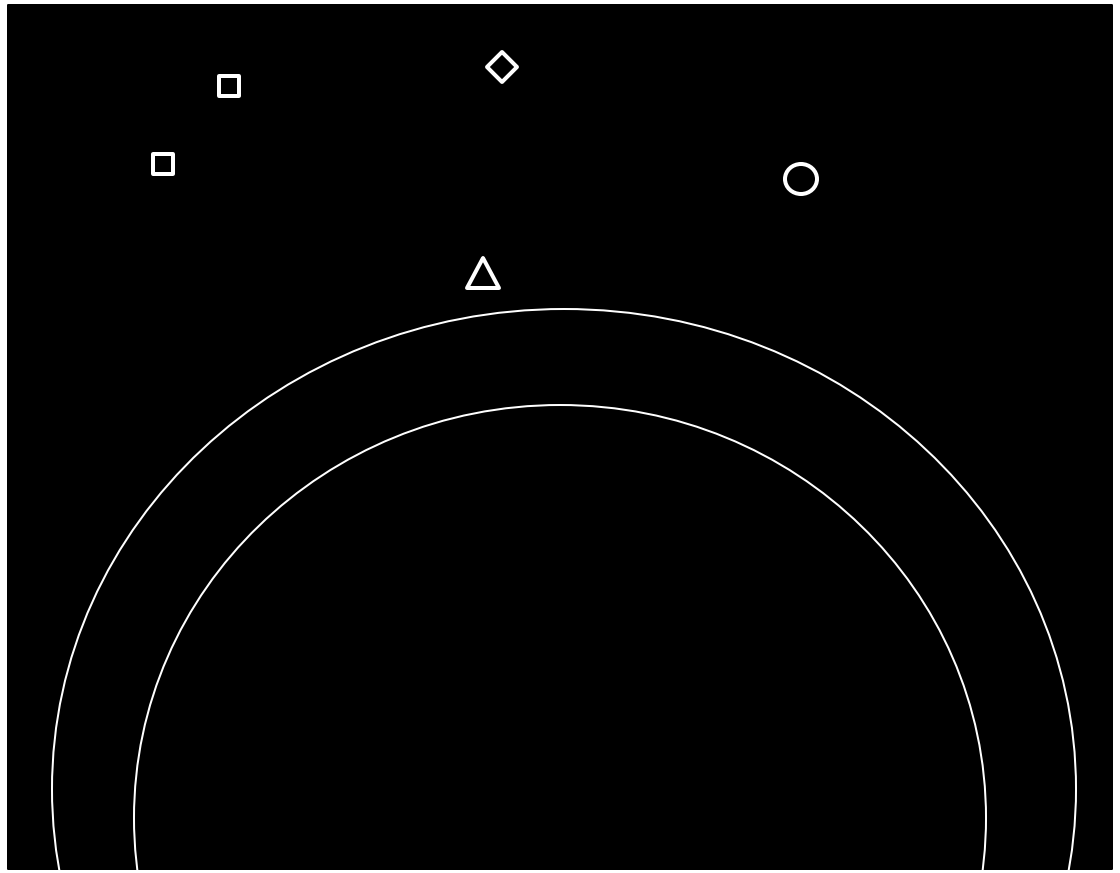


High Level Process

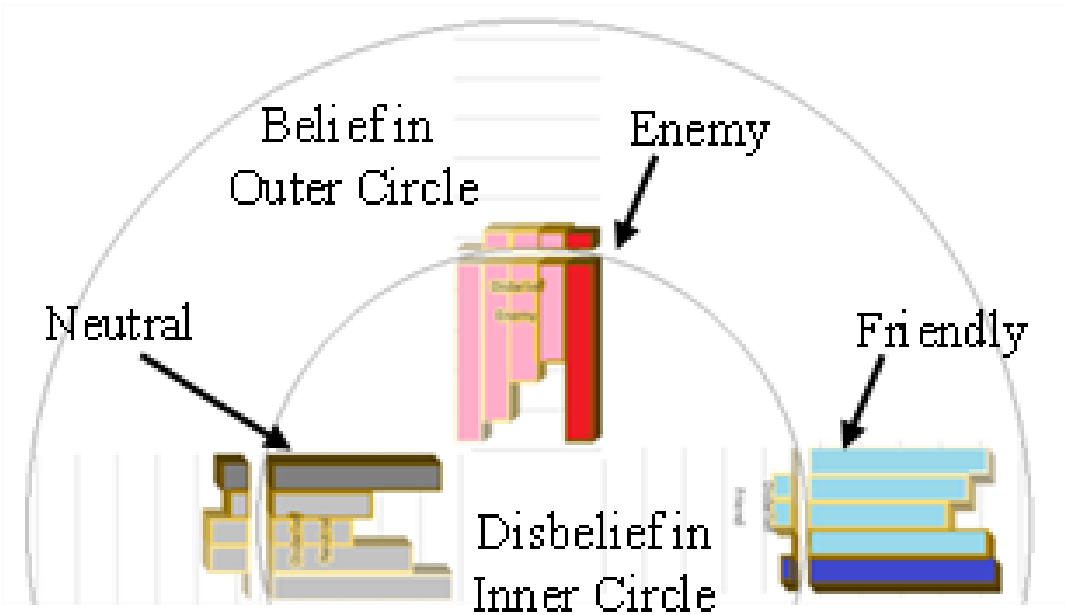
- Sense Signals
- Determine confidence of each class and confidence not each class
- Fuse entities
- Compute Overall Confidence Values

Proposed Interface

Aircraft Display



CID Support (Clear Friendly)



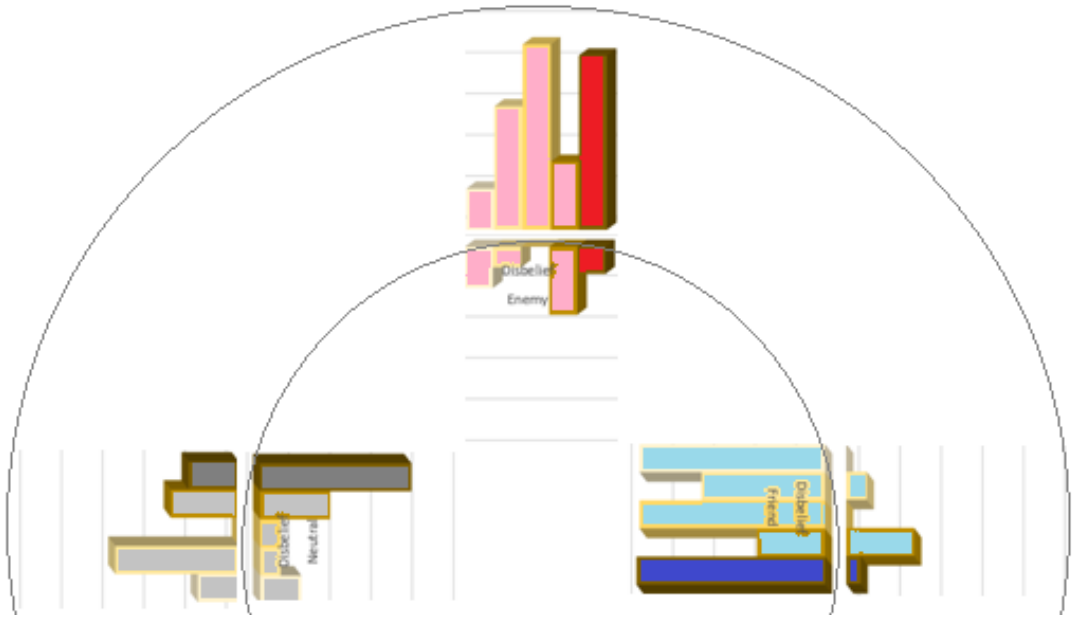
Conflict Can Be Illustrated

Conflicting Information Displayed

- Sensor 2 shows belief is neutral
- Sensor 2 and 3 also shows strong belief target is Enemy

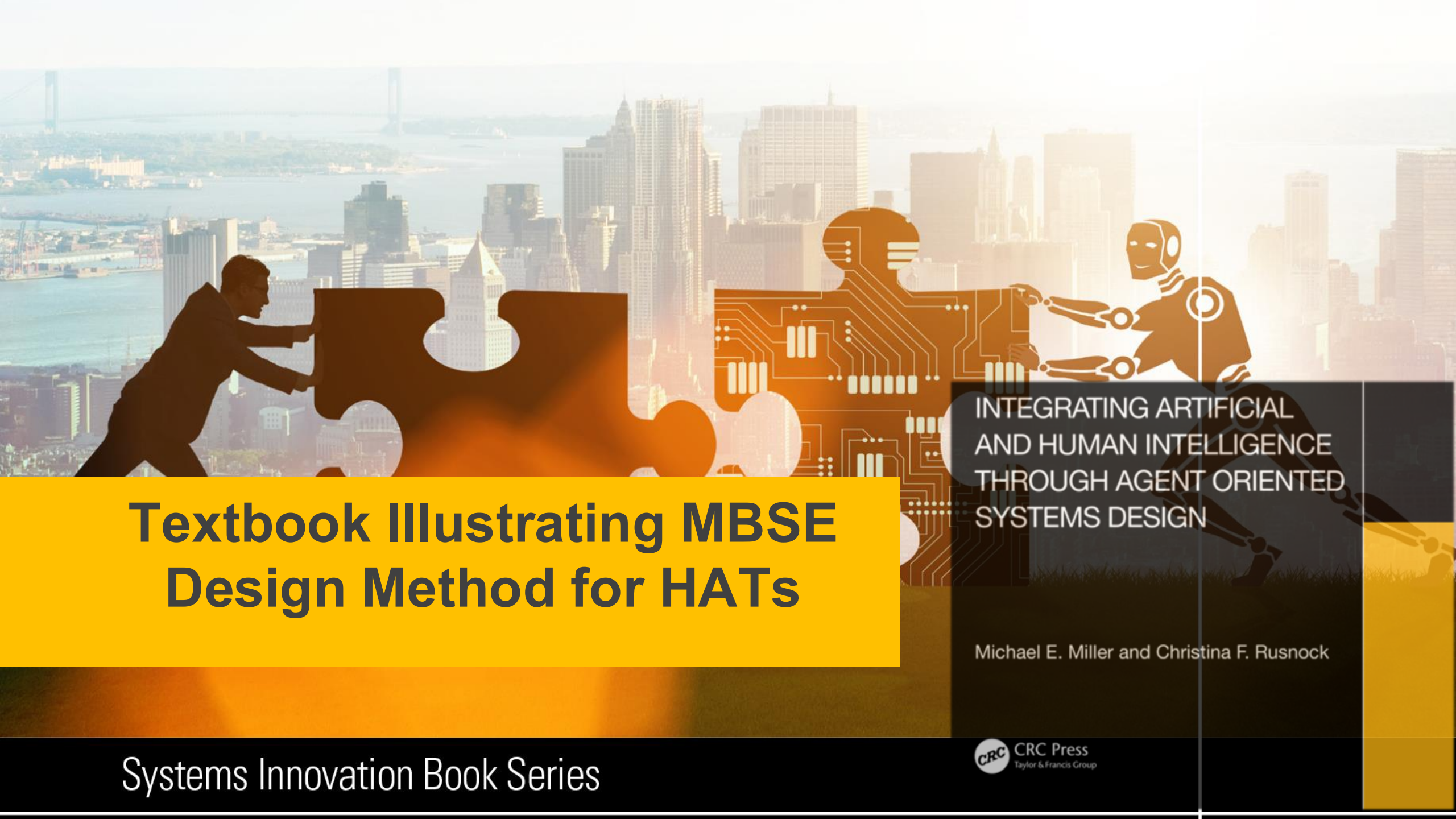
- Based on operational Information, the pilot can decide which sensors to rely upon to support decisions or work to gain additional information

Overall Belief is Enemy



Summary

- Human-AI Teaming goal:
  - Maintain Human Accountability for Safety Critical Decisions
  - Humans Can Include Ethical and Moral Considerations in Decision Loop
  - Take advantage of AI's Ability to Process Information
  - Improve the Speed of Decision-Making using AI
- For a single decision, a human can not be "in the loop" while AI speeds decisions
- Placing human on the loop for lower-level decisions while in the loop for higher-level decisions can help satisfy goals
- Example illustrates human "in the loop" for final CID decision but "on the loop" for sensors, fusion, and CID recommendation



Textbook Illustrating MBSE Design Method for HATs

References

- Air Force Doctrine Note 25-1: Artificial Intelligence (AI), 1 (2025). https://www.doctrine.af.mil/Operational-Level-Doctrine/AFDN-25-1-Artificial-Intelligence/
- Håkansson, M. (2005). An evaluation of subjective logic for trust modelling in information fusion [University of Skovde]. http://www.diva-portal.org/smash/record.jsf?pid=diva2:3403
- Johnson, J. (2022). The AI Commander Problem: Ethical, Political, and Psychological Dilemmas of Human-Machine Interactions in AI-enabled Warfare. Journal of Military Ethics, 21(3–4), 246–271. https://doi.org/10.1080/15027570.2023.2175887
- Jøsang, A. (2002). The consensus operator for combining beliefs. Artificial Intelligence, 141(1–2), 157–170. https://doi.org/https://doi.org/10.1016/S0004-3702(02)00259-X
- Miller, M. E., & Rusnock, C. F. (2024). Integrating Artificial and Human Intelligence through Agent Oriented Systems Design (1st ed.). CRC Press. https://doi.org/https://doi.org/10.1201/9781003428183