

NATIONAL  
ACADEMIES

Sciences  
Engineering  
Medicine

# Preserving Human Agency and Accountability in Human-AI Teaming

*Michael Muller, BOHSI member (for identification only: IBM Research)*

My views are my own, and do not necessarily represent  
views of IBM or of the National Academies.

NOVEMBER 2022 UPDATE



# Problems (in 2025) with Human-AI Interactions

- AI in need of human control – even for simple tasks
  - We initiate a kind of “batch processing” and receive a summarized outcome
  - We are beginning to have transparency into the *steps* of a process – but *less of* the intermediary (data-states of a process) - to generate that outcome
  - We don’t know what may be missing – the “haunting” of intermediary steps in generative AI
- AI hallucinations
  - AI sometimes generates information of questionable veracity
  - Hallucinations are a direct consequence of the probabilistic nature of transformer architectures
    - “Stochastic parrots”
  - We can mitigate hallucinations. However, when AIs generate, their likelihood of hallucination is always  $> 0$
  - If we suppress hallucinations, do we also suppress creativity?
- Let’s look at some AI-based alternatives

NOVEMBER 2022 UPDATE

# Experiment Series 1: H-AI Teaming for Co-Creativity



# 01

# Human-AI Teaming for Co-Creativity

- Analogy-based design (one strategy of many)
  - “How is a database like a toolbox?  
What data tools are in the data toolbox?”
  - “What is a good metaphor for a human using a computer?”
    - 1<sup>st</sup> answer: “A conductor leading an orchestra”
    - 2<sup>nd</sup> answer: “A pilot flying an airplane”
    - **Reframing** as “creativity move”

# Human-AI Teaming for Co-Creativity

- Analogy-based design (one strategy of many)
  - “How is a database like a toolbox?  
What data tools are in the data toolbox?”
  - “What is a good metaphor for a human using a computer?”
    - 1<sup>st</sup> answer: “A conductor leading an orchestra”
    - 2<sup>nd</sup> answer: “A pilot flying an airplane”
    - **Reframing** as “creativity move”
- Brainstorming with an AI agent
  - Divergent thinking
  - Convergent thinking

**A** **Human:** The business problem is: We need a plan to get our employees to try out a new AI service. The service will help them with receiving support from their human resources department. **How can we generate enthusiasm for using the AI service for HR?**

**B** **AI:** ... Here are **a few brainstorming ideas** to generate enthusiasm for using the AI service for HR:

1. Create an engaging demo [ . . . ]
2. Communicate the value proposition [ . . . ]
3. Offer incentives [ . . . ]
4. Provide training and support [ . . . ]
5. Foster a culture of experimentation [ . . . ]
6. Share success stories: [ . . . ]

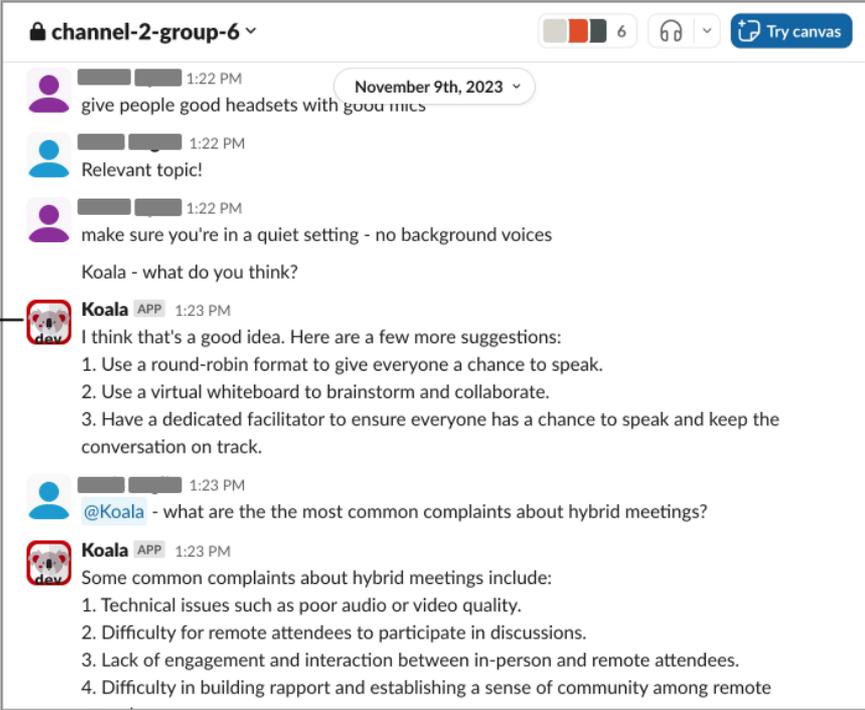
**C** **Human:** I think we want to **avoid the time and cost of training**. Please **suggest alternatives** to your idea #4 about training and support.

**D** **AI:** ... Instead of traditional training, **you could consider** the following alternatives:

1. Self-guided tutorials [ . . . ]
2. Interactive onboarding [ . . . ]
3. Peer-to-peer support [ . . . ]
4. Online forums or communities [ . . . ]

# Multiple Humans Brainstorming with an AI agent

- Slack channel + AI agent (Koala)
  - “How can we improve online meetings?”
  - Brainstorming in a familiar setting (Slack)
    - Divergent thinking: Humans ask AI for suggestions or input
    - Convergent thinking: Humans choose 3 “final” ideas to report to an imaginary client
  - Testbed for ReactiveAI vs. ProactiveAI



The screenshot shows a Slack channel named "channel-2-group-6" with 6 members. The date is November 9th, 2023. The conversation includes:

- A user asks: "give people good headsets with good mics"
- A user says: "Relevant topic!"
- A user says: "make sure you're in a quiet setting - no background voices"
- A user asks: "Koala - what do you think?"
- Koala** (APP) responds: "I think that's a good idea. Here are a few more suggestions:
  1. Use a round-robin format to give everyone a chance to speak.
  2. Use a virtual whiteboard to brainstorm and collaborate.
  3. Have a dedicated facilitator to ensure everyone has a chance to speak and keep the conversation on track."
- A user asks: "@Koala - what are the the most common complaints about hybrid meetings?"
- Koala** (APP) responds: "Some common complaints about hybrid meetings include:
  1. Technical issues such as poor audio or video quality.
  2. Difficulty for remote attendees to participate in discussions.
  3. Lack of engagement and interaction between in-person and remote attendees.
  4. Difficulty in building rapport and establishing a sense of community among remote"



# Experiment Series 2: H-AI Exploration of different moral codes



02

# AI agents with Different Moral Codes Advise a Human

- Create 2 pseudo-agents with different moral codes
  - **Human Employee** + Individualist vs. Collectivist  
**Manager-agents**
  - **Human Graduate Student** + Speed-of-thesis vs. Breadth of **Thesis-Professor-agents**
  - **Microentrepreneur** + Business-success vs. Neighborhood-success **Banker-agents**

# AI agents with Different Moral Codes Advise a Human

- Create 2 pseudo-agents with different moral codes
  - **Human Employee** + Individualist vs. Collectivist  
**Manager-agents**
  - **Human Graduate Student** + Speed-of-thesis vs. Breadth of **Thesis-Professor-agents**
  - **Microentrepreneur** + Business-success vs. Neighborhood-success **Banker-agents**
- Create 4 pseudo-agents with different beliefs about customer values
  - Car sellers at an auto dealership
    - Performance seller-agent
    - Safety seller-agent
    - Cost seller-agent
    - Luxury seller-agent

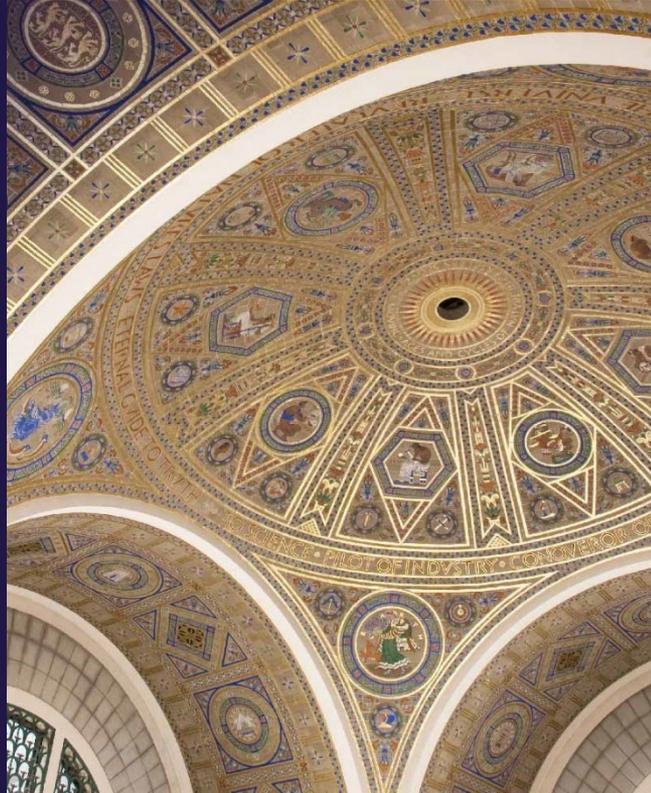
# AI agents with Different Moral Codes Advise a Human

- Create 2 pseudo-agents with different moral codes
  - **Human Employee** + Individualist vs. Collectivist **Manager-agents**
  - **Human Graduate Student** + Speed-of-thesis vs. Breadth of **Thesis Professor-agents**
  - **Microentrepreneur** + Business-success vs. Neighborhood-success **Banker-agents**
- Create 4 pseudo-agents with different beliefs about customer values
  - Car sellers at an auto dealership
    - Performance seller-agent
    - Safety seller-agent
    - Cost seller-agent
    - Luxury seller-agent

Initial observations: Each pseudo-agent...

- Adopts its own distinct moral position
- Articulates its own and its partners' positions
- Negotiates with others to reach consensus
- Retains its original moral position during and after that consensus

# Lessons Learned



# Conclusion

## Future Trajectories of Human-AI Collaboration & Teaming

- What level of autonomy should AI systems have?
- How should we divide tasks between humans and AI?
- What does it mean to trust and collaborate with AI in a team setting?
- How can we design AI teammates that enhance human cognition and performance?

# Conclusion

## Future Trajectories of Human-AI Collaboration & Teaming

- What level of autonomy should AI systems have?
- How should we divide tasks between humans and AI?
- What does it mean to trust and collaborate with AI in a team setting?
- How can we design AI teammates that enhance human cognition and performance?

## Proposed Answers in 2025

- Under **human control** – and **human review** with calibrated proactivity
- **Human assigns tasks to AI assistant** on a “sliding scale” to balance workloads + organizational needs
- Collaboration and trust via **human-verification of AI outcomes** (chain-of-thought + **direct inspection**)
- AI teammates are (or become) assistants that support humans in **human endeavors**

# Conclusion

## Future Trajectories of Human-AI Collaboration & Teaming

- What level of autonomy should AI systems have?
- How should we divide tasks between humans and AI?
- What does it mean to trust and collaborate with AI in a team setting?
- How can we design AI teammates that enhance human cognition and performance?

## Proposed Answers in 2025

- Under **human control** – and **human review** with calibrated proactivity
- **Human assigns tasks to AI assistant** on a “sliding scale” to balance workloads + organizational needs
- Collaboration and trust via **human-verification of AI outcomes** (chain-of-thought + **direct inspection**)
- AI teammates are (or become) assistants that support humans in **human endeavors**

## → New Questions

- How much human control and review are needed?
- How much proactivity is useful in each of many diverse tasks?
- How to balance control and co-creativity?