AI as Collaborative Partner: Rethinking Human-AI Teaming for the Real World

Melinda Gervasio, Pedro Sequeira, Eric Yeh,
Nicholas Marion, Sarah Bakst, Helen Gent
SRI International
{firstname.lastname}@sri.com

Abstract

Much work in human-AI teaming today involves collaboration under fairly constrained settings. Humans supervise AI agents, who are relegated to following orders. The division of tasks is relatively superficial, with independent actions executed in parallel, or with simple, linear dependencies between them. Communication is rigid and turn-based, with both parties speaking in complete, unambiguous sentences. However, collaboration in dynamic, real-world environments is rarely this straightforward. Team members adopt different roles as they continually adapt to an evolving situation, using efficient, timely communication to coordinate their actions. In this paper, we explore the technology requirements for AI agents to achieve this kind of collaboration and introduce COLLEAGUE, an agent framework designed to support adaptive, mixed-initiative interaction in dynamic domains.

Introduction

Consider the following scenario: A tactical search and rescue team is looking for survivors in a collapsed building. They come across a casualty trapped under the rubble, but also observe signs of fire two hundred yards away. The team promptly executes a blend of multiple drills: two team members with cribbing and rescue equipment attend to the casualty, a third serves as lookout for the fire and other hazards, and the fourth radios the chief to provide an update on the situation. All this is accomplished with minimal but well-timed communication, enabled by standard operating procedures, practice, experience, and established norms.

High-functioning human teams performing highly interdependent, time-critical tasks often cannot rely on a traditional chain-of-command hierarchy to respond to an incident. Team members must rapidly develop a course of action that meets the commander's intent and adapt to unexpected situations and exogenous events. Sometimes they follow orders, other times they act independently or take charge, and they switch roles as needed to fulfill the mission. Now imagine a similar team, but with a mix of humans and embodied AI agents. For the AI agents to be effective in such a team, they must be able to communicate and coordinate their actions with their teammates and flexibly adapt to different roles much as the human team members do.

This type of human-AI teaming can occur in a range of settings: soldiers and autonomous drones on a mission to locate a high-value target, human and robotic technicians running laboratory protocols, chefs and robotic cooks operating a pop-up restaurant, etc. In these scenarios, the team members have different but comparable, overlapping skills, and their roles and responsibilities may vary over time. They execute interdependent tasks that must be carefully coordinated, often in cognitively demanding, high-stakes situations that mandate concise, purposeful communication.

One way to address this is through rote execution, with team members acting according to scripted, well-rehearsed policies, but this does not scale to chaotic, real-world environments. Another is through centralized control, with a team leader making all decisions, but this can create critical bottlenecks in time-critical situations. Large language models (LLMs) have helped to enable problem-solving through natural dialogue, but the collaboration settings they address are predominantly supervisory and turn-based, with tasks that are entirely independent or only loosely interrelated.

In this paper, we discuss critical capabilities for supporting successful human-AI collaboration in dynamic, real-world settings. We also present some early work on developing an agent framework with some of these capabilities.

Human-AI Teaming Requirements

Theory of Mind

A critical enabler for successful collaboration in human teams is Theory of Mind (ToM)—the ability to ascribe mental states to others to explain and predict their behavior and integrate such beliefs into one's own decision-making

(Baron-Cohen, Leslie, and Frith 1985). In addition to understanding the world dynamics and the consequences of their own actions, AI agents need to be equipped with ToM-like mechanisms to achieve effective collaboration in mixed teams. They need to be able to understand the intent behind and form expectations about the behavior of teammates while interacting with them, and adapt to perceived changes in their behavior. Aside from (physical) behavior, AI agents also need to form expectations about communications, incorporating the information they receive into their internal models of the world and their teammates, and communicating their own goals, intentions, capabilities, and knowledge to teammates to maximize common understanding and transparency.

Mixed-Initiative Interaction

Collaborative interactions are necessarily mixed-initiative: team members are at times reactive and at times proactive, able to both give and take direction and information. ToM capabilities enable AI team members to engage in such mixed-initiative interactions through the expectations they provide. Expectations about its own actions enable an AI agent to detect failures that it may want to communicate with its human teammates. For example, two team members extricating a casualty have expectations about each other's actions and responses when cribbing, so if one has difficulty putting in the next block, they will want to notify the other, as resolving the issue may require coordinated actions. Expectations about a teammate's actions inferred via ToM also enable an AI agent to speak up in the case of apparent inaction—for example, to remind the team member assigned as lookout about their assigned task if they are found watching the extrication rather than looking out for potential hazards. Expectations about teammates' information needs, together with standard communication protocols and social norms, drive an AI agent to provide status updates to relevant parties at designated points in the task (e.g., an AI lookout informing its teammates about new hazards), together with any necessary contextualized explanations (e.g., a decision to divert transport to a cardiac trauma center because the patient is experiencing a sudden major cardiac event). Attention to social norms is also important: many approaches to addressing transparency involve AI entities elaborating on their plans and explaining their every action, but humans only do this when explicitly requested to do so, or as dictated by the situation (e.g., explaining a sudden change in plans due to some local event).

Communication Beyond Words

Collaboration requires communication: team members need to know when to provide directives or information, when to silently receive instruction, when to acknowledge receipt of instructions, etc. Spontaneous speech, which becomes more prevalent the greater one's focus on a task (Berthold and Jameson 1999), is often peppered with disfluencies, such as false starts ("Okay, I'm putting in the next block… actually, lever it up more?") and repairs ("Head to waypoint Alpha—waypoint Bravo"). State-of-the-art speech recognition systems often cannot handle these ill-structured phrases, responding with a blanket "I'm not sure I understand." Further, verbal communication is just one way in which collaborative teammates exchange information. Performance of an action is itself communicative and, in shared task environments where teammates can observe each other's actions and their effects on the world, greatly reduce the need for explicit communication. Seamless integration into human teams requires AI teammates who can understand and utilize these implicit forms of communication to accurately interpret intent and formulate appropriate responses.

Human-AI Team Metrics

Task achievement—whether the team's joint tasks are achieved and the quality of their achievement (e.g., total extrication time, patient outcomes)—is an important measure of successful human-AI collaboration, but it is not sufficient. The quality of the communication and coordination also needs to be assessed. Simple measures such as the number of utterances can be problematic—in situations requiring tight coordination, rapid-fire exchanges are sometimes a sign of *good* communication. Metrics inspired by Grice's maxims (Grice 1976) potentially offer a better measure of communication quality: ensuring that the information conveyed is correct, non-superfluous, relevant, and clear.

Successful human-AI collaboration also requires *appropriate* trust. Too often, the focus is simply on increasing trust, but earlier lessons about the perils of automation (Parasuraman and Riley 1997) and recent studies on the negative effects of explainable AI (Bansal et al. 2021) caution that (seemingly) proficient AI can lead to *miscalibrated* (under- or over-) trust, resulting in the underutilization or inappropriate use of AI. Collaborative capabilities bolstered by ToM mechanisms enable socially attuned interactions aligned with human expectations and social norms. Such AI agents will be perceived as more competent and dependable, easier for humans to work with and adapt to, and more likely to foster the development of appropriate trust.

COLLEAGUE

We are currently developing COLLEAGUE (COLlaborative Language-Enabled Agents Grounded in Understanding and Explanation), an agent framework for collaboration with humans in cyberphysical domains (Figure 1). We have completed a proof-of-concept demonstration of COLLEAGUE in a collaborative kitchen environment. Here we describe some of the key collaboration capabilities in the

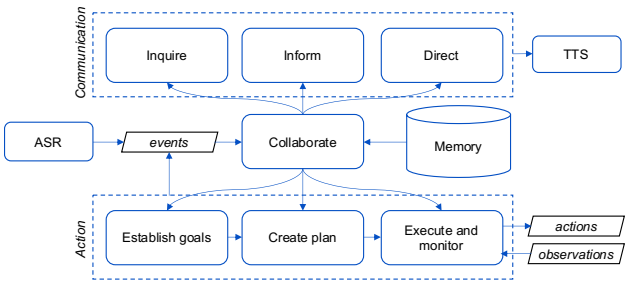

Figure 1. COLLEAGUE Agent Framework for Human-AI Collaboration. A Collaboration Manager orchestrates the flow of control and information among the action and communication components.

framework. For simplicity, we discuss collaboration between an AI agent and a human, but the framework supports collaboration between multiple humans and AI agents.

Prosody-Aware Speech Understanding

Distinguishing between different kinds of speech acts requires understanding the subtleties of prosody that indicate floor-holding, desire for help, self-corrections, etc. as well as handling the disfluencies in spontaneous speech. COLLEAGUE currently addresses two key problems in understanding natural speech. *Question identification and classification* determine whether an utterance is a question and then, whether it is a *wh-* question or a *yes/no* question. *Endpoint detection* determines when a human has finished speaking—i.e., whether a pause indicates the end of the utterance. Textual information is not always sufficient to know how to interpret a pause or to distinguish a question from a statement. We address this challenge by incorporating prosodic and other acoustic features.

In both the endpoint detection and question classification models, we represent textual information as BERT sentence embeddings (Devlin et al. 2019) and prosodic-acoustic information as a vector of frame-level features extracted using ProsodyBERT (Hu et al. 2022). These are independently fed a series of CNN layers and feed-forward layers respectively before being concatenated and processed through a final set of feed-forward layers. The final output of the endpoint detection model predicts whether a given pause is floor-holding or utterance-final. The question classification approach is a series of two models: the first predicts whether a given phrase is a question and the second, whether a question is a yes/no or wh- question. Experimental results show significant improvement with the addition of acoustic features.

LLM-Augmented Planning

Advances in large language models (LLMs) have seen their increased use in tasks requiring complex reasoning such as planning (Ahn et al. 2022, Huang et al. 2022). However, they remain susceptible to problematic behavior, particularly for plans that involve complex interdependencies between tasks (Valmeekam et al. 2023). In COLLEAGUE, we found them to be prone to generating plans that looked reasonable but were unexecutable—for example, because the plans included unknown actions or had unsatisfied preconditions. COLLEAGUE takes advantage of LLMs' proficiency in translating between natural language and formal representations (Izquierdo-Badiola et al. 2024) to translate user requests into formal planning goals—e.g., a reward function for a POMDP planner, a PDDL goal specification for a PDDL planner, or a high-level task for an HTN planner. Different domains may be more amenable to certain planning approaches; COLLEAGUE does not mandate any specific planning formalism but requires the planner to be able to generate multiagent plans, comprising the actions and expected state transitions or action effects for all agents.

Expectation-Driven Monitoring and Interaction

The multiagent plans generated by the planner provide COLLEAGUE information not only about its own actions and the human's expectations about them, but also the human's actions and their expected effects. This provides a mental model of what the human will do and when, which COLLEAGUE can use to monitor execution, and decide whether and when its own actions can be performed. Crucially, this also enables COLLEAGUE to know when to initiate communications in various situations. For example, an unexpected action by the human may trigger COLLEAGUE to clarify the human's intent, while uncertainty about what to do next may lead COLLEAGUE to ask for help. If COLLEAGUE encounters a situation that prevents it from executing its planned action or if it discovers a better way to do things, it can proactively notify the human of the change in plans along with an explanation.

Deliberate Collaboration Management

Much work on human-robot collaboration is focused on one-to-one, turn-based interaction to accomplish a joint task, which allows the use of predefined interaction workflows and simple planning approaches. However, collaboration in real-world domains is much more fluid and opportunistic: team members initiate, redirect, or abandon interactions as they adapt continuously to an evolving situation, under timing and bandwidth constraints. COLLEAGUE's Collaboration Manager (CM) manages the tasks and communications of the agent by orchestrating the flow of control through the different stages of collaboration: establishing objectives, developing a joint plan and assigning roles, executing the plan, and monitoring and adapting as needed. The CM continually receives input from the human, the action components, and the communication components, and uses

this information to decide what to do next. This encompasses a wide variety of possible actions: responding to a request for information, asking for clarification about an ambiguous request, continuing execution, alerting the user to an unexpected situation, abandoning the current task, etc. The CM uses LLMs to provide the practical, commonsense knowledge to inform its decisions on responding to different situations and generating actual responses.

Multimodal Multicriteria Memory

To support the highly context-dependent decision-making in the CM, COLLEAGUE uses a multimodal memory that supports retrieval against multiple criteria. The CM needs to be able to retrieve a relevant set of past experiences necessary to provide the information needed to interpret utterances and respond to events. Because of the diversity of these inputs, the information and relevance criteria may vary widely. For example, answering a question about the availability of all ingredients would require identifying ingredients from past dialogue about the recipe, followed by retrieval of their last known positions. Current retrieval-augmented generation approaches for LLM inference encode all memories as vectorized documents (textual summaries), supporting only a single relevance criterion. COLLEAGUE supports complex memory and inference operations through an LLM employing chain of thought (Wei et al. 2022), where the information need and retrieval criteria are identified at each step. Memories are stored by modality, but indexed by a textual description that contains additional information describing relevant queries. This allows a textual description of the current query to find the relevant modality and memory to be reconstituted, which in turn facilitates additional inferences.

Current Status

COLLEAGUE is currently implemented within an agentic architecture, which provides flexibility in the approaches used for the different components and supports the combination of AI technologies and traditional code. We have a proof-of-concept demonstration of COLLEAGUE in a collaborative kitchen task environment, built on top of the ALFRED application domain (Shridhar et al. 2020) developed on the AI2Thor simulation platform (Kolve et al. 2017) (Figure 2). Within this environment, COLLEAGUE demonstrates several of the collaborative capabilities discussed above, including understanding user directives ("Let's make a snack!"), seeking clarification by offering options ("Do you want to make a salad or a sandwich?"), recognizing questions from prosodic cues ("We have eggs!?"), generating and executing a joint plan with the user (e.g., fetching sandwich ingredients while the user prepares the bread), and adapting as needed (e.g., switching to making a sandwich

Figure 2. Human-AI Collaboration in COLLEAGUE. Making a sandwich together: COLLEAGUE getting an egg to fry while the human (off-camera) prepares the bread.

when the tomatoes turn out to be underripe). We are currently developing COLLEAGUE as a hardware robot partner that works with a human lab technician to develop and execute experimental protocols using assorted lab equipment and supplies (e.g., test tubes, reagents, centrifuges).

References

Ahn, M.; Brohan, A.; Brown, N.; et al. 2022. Do as I can, not as I say: Grounding language in robotic affordances. arXiv:2204.01691.

Bansal, G.; Wu, T.; Zhou, J.; et al. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In Proc. CHI 2021, 1–16.

Baron-Cohen, S.; Leslie, A. M.; and Frith, U. 1985. Does the autistic child have a "theory of mind"? *Cognition* 21(1): 37–46.

Berthold, A. and Jameson, A. 1999. Interpreting symptoms of cognitive load in speech input. In Proc. UM99*,* 235–244.

Devlin, J.; Chang, M. W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proc. NAACL, Vol. 1, 4171–4186.

Grice, P. 1975. Logic and conversation. In *Syntax and Semantics*. Vol. 3: Speech acts, edited by P. Cole and J. Morgan, 41–58. New York: Academic Press.

Hu, Y.; Zhang, C.; Shi, J.; et al. 2022. ProsodyBERT: Self-Supervised Prosody Representation for Style-Controllable TTS.

Huang, W.; Xia, F.; Xiao, T.; et al 2022. Inner monologue: Embodied reasoning through planning with language models. arXiv.org/abs/2207.05608.

Izquierdo-Badiola, S.; Canal, G.; Rizzo, C.; and Alenyà, G. 2024. PlanCollabNL: Leveraging large language models for adaptive plan generation in human-robot collaboration. In Proc. ICRA.

Kolve, E.; Mottaghi, R.; Han, W. et al. 2017. AI2-THOR: An interactive 3D environment for visual AI. arxiv.org/abs/1712.05474.

Parasuraman, R. and Riley, V. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39(2): 230–253.

Shridhar, M.; Thomason, J.; Gordon, D.; et al. 2020. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In Proc. CVPR., 10740–10749.