Collaborations Between People and AI Systems (CPAIS)

Human - AI Collaboration Framework and Case Studies

September, 2019

PARTNERSHIP ON AI

Human - AI Collaboration Framework and Case Studies

Introduction

Developing best practices on collaborations between people and AI systems – including issues of transparency and trust, responsibility for specific decisions, and appropriate levels of autonomy – depends on a nuanced understanding of the nature of those collaborations. For example, a human interacting one-on-one with a I on a consumer website requires different features and responses than one, or perhaps even a group of individuals, sitting in an autonomous vehicle.

With these issues in mind, the Partnership on AI, with members of its Collaborations Between People and AI Systems Expert Group, has developed a Human-AI Collaboration Framework containing 36 questions that identify some characteristics that differentiate examples of human-AI collaborations. We have also prepared a collection of seven Case Studies that illustrate the framework and its applications.

This project explores the relevant features one should consider when thinking about human-AI collaboration, and how these features present themselves in real-world examples. By drawing attention to the nuances – including the distinct implications and potential social impacts – of specific AI technologies, the Framework can serve as a helpful nudge toward responsible product/tool design, policy development, or even research processes on or around AI systems that interact with humans.

As a software engineer from a leading technology company suggested, this Framework would be useful to them because it would enable focused attention on the impact of their AI system design, beyond the typical parameters of how quickly it goes to market or how it performs technically.

> *"By thinking through this list, I will have a better sense of where I am responsible to make the tool more useful, safe, and beneficial for the people using it. The public can also be better assured that I took these parameters into consideration when working on the design of a system that they may trust and then embed in their everyday life."*
> Software Engineer, PAI Research Participant

The Framework is not intended to prescribe what constitutes appropriate answers to the questions it poses. Rather, the questions are intended to be provocations that advance understandings of human-AI collaboration, contribute to the AI community's decision-making around responsible AI development and deployment, and ultimately affect technology practice.

As such, no two people may answer the questions uniformly for the same case example (and you may find yourself disagreeing with the case writer answers). Even individual case study writers may have trouble coming to a single answer for some of the questions, choosing instead to offer their commentary and to hedge their answers. This type of fruitful dialogue, whether internal or amidst teams developing AI tools and policies, is exactly the type of thinking that the Framework is intended to provoke.

The seven Case Studies included in this report provide a glimpse into the variety of potential collaborations between people and AI systems, and includes specific examples of the Framework in practice. The cases and completed Framework questionnaires are as follows:

1. Virtual Assistants and Users *(Claire Leibowicz, Partnership on AI)*
2. Mental Health Chatbots and Users *(Yoonsuck Choe, Samsung)*
3. Intelligent Tutoring Systems and Learners *(Amber Story, American Psychological Association)*
4. Assistive Computing and Motor Neuron Disease Patients *(Lama Nachman, Intel)*
5. AI Drawing Tools and Artists *(Philipp Michel, University of Tokyo)*
6. Magnetic Resonance Imaging and Doctors *(Bendert Zevenbergen, Princeton Center for Information Technology Policy)*
7. Autonomous Vehicles and Passengers *(In Kwon Choi, Samsung)*

About the Project and its Participants

This project came out of the first in-person meeting of PAI's Collaborations Between People and AI Systems (CPAIS) Expert Group, which took place in November, 2018. The CPAIS expert group consists of about 30 representatives from across technology, academia, and civil society. Within these sectors, group members represent varied disciplinary training and roles (e.g., policy, research, product). While the members of the CPAIS Expert Group all work in the realm of human-AI collaboration, each used disparate heuristics when developing the parameters for this Framework, and assessing potential best practices for how machines and humans should collaborate. Our discussions illuminated the range of collaborations between people and AI systems, and specifically, how better understanding this spectrum can improve how the AI community develops tools and policies affecting human-AI interaction.

Human - AI Collaboration Framework

I. Nature of Collaboration

Stage of development or deployment

1. Is the AI fixed once deployed or evolving over time via model updates/continual interaction?

2. To what extent is there ongoing collaboration between the AI's developer(s) and the AI system? [No collaboration, limited collaboration, moderate collaboration, active collaboration]

3. Is the AI system currently used by people other than the AI's original developers?

Goals

4. Are the goals of the human-AI collaboration clear or unclear?

5. What is the nature of the collaboration's goals? [Physical, knowledge/intellectual, emotional, and/or motivational in nature]

6. Is empathy a precondition for the human-AI interaction to function as intended?

7. Are the human and the AI system's goals aligned?

Interaction Pattern

8. Is the collaboration repeated over time or is it a one-time engagement? If over time, at what time-scale?

9. Is the interaction concurrent – with both human and AI contributing in parallel – or does it depend on taking turns?

Degree of agency

10. Does the AI or human agent contribute more to the system's decision-making? Action-taking?

11. How much agency does the human have? The AI system? [None, limited, moderate, high, full]

II. Nature of Situation

Location and context

12. Are other people or other AI systems involved as third-parties? This can apply to either 1-1 collaborations or multiple people & AI (e.g., in the instance of an AI teacher interacting with a classroom of human students, those human students would not be included as third-parties).

13. Are the human and AI agents co-located physically or virtually?

Awareness

14. Is the human likely aware that they are interacting with an AI system?

15. Does the human need to consent before interacting with the AI system?

Consequences

16. How significant are the consequences should the AI fail to perform as designed/expected? What are those consequences? [Low, moderate, high]

17. How significant are the benefits of the AI to the users should it perform as designed/ expected? What are those benefits? [Low, moderate, high]

18. What are the potential consequences and benefits of the outcome of the collaboration?

19. What might be the broader impacts of the human-AI collaboration?

20. To what extent do typical users consider privacy and security when interacting with the AI agent? [Low, Moderate, High]

Assessment

21. Who is the main party or individual assessing the nature and effectiveness of the human-AI collaboration?

22. Are assessments of the human-AI collaboration’s outcome subjective or objective?

Level of Trust

23. Are both the human and the AI agent trusting and trustworthy? AI trustworthiness can be defined broadly, driven by task competence, safety, authority, and authenticity, amongst other features (e.g., we know an AI comes from the same affiliation it claims to be from).

III. AI System Characteristics

Interactivity

24. What is the mode of the interaction between the two agents? [Via screen, voice, wearables, virtual reality, or something else]

25. Could the nature of the data that the AI system operates over impact its interactivity? E.g., a wearable AI may interact with the user via voice, but may make inferences over voice as well as sensor data.

Adaptability

26. Is the AI system passively providing information or proactively anticipating the next steps of the interaction?

Performance

27. How predictable is the AI system? [Low, moderate, high]

28. Does the system often produce false-positives? False-negatives?

Explainability

29. Can the AI system communicate its confidence levels to a human?

30. How does the AI system communicate its decision-making process and inputs to that decision-making process to the human?

Personification

31. How human-like is the AI system? [Not very, moderately, or highly human-like]

32. How easily anthropomorphized is the AI system?

IV. Human Characteristics

Age

33. Is the person(s) collaborating with the AI system a child (under 18), an adult (18 - 65), or a senior (over 65)? Some mixture of the above, when >1 person collaborating with the AI system?

Differently-abled

34. Does the person collaborating with the AI have special needs or accommodations?

Culture

35. Are there cultural consistencies/norms for those collaborating with the AI system?

36. What level of previous technology interaction has the user(s) of the system had? [Low, moderate, high]

CASE 1: Virtual Assistants and Users
Claire Leibowicz, Program Lead (PAI)

Context/Scenario

Employing a personal assistant used to only be possible for those who could afford to hire someone. Now, a variety of virtual assistant technologies have come to market to help with scheduling, tasks, and other facets of daily life, ostensibly democratizing the time-saving and frictionless experience that human assistants provide. Amazon's Alexa, Apple's Siri, Facebook's M, Google's Assistant, IBM's Watson Assistant, Microsoft's Cortana and others, all offer users various capabilities such as reminders, information gathering, Q&A, shopping lists, music playback, and more in response to verbal requests from users. These tools can also communicate with other cloud technologies to manage elements like the temperature and lighting in one's home. They can even tell jokes (albeit, pre-programmed jokes).

AI System

Automated personal assistants often use natural language processing (NLP), speech recognition, and other machine learning techniques in order to interpret voice or text inputs from users. The more ubiquitous, innovative devices are voice-activated and are the result of advances in speech recognition that can interpret auditory requests - often in different languages and from many different voices. The AI systems can then pattern match in order to retrieve information from online sources like Wikipedia or weather services. Such systems may interact with users and "learn" over time, better understanding context cues offered and returning more accurate and fluid responses to queries.

Human-AI Collaboration

Most modern virtual assistants are native to physical devices that look akin to speakers, and while they are tacitly "listening" all the time, the virtual assistants typically only engage when users offer a specific verbal request: "Hey Siri," or "Okay Google," or "Alexa" for example. Just as a human might not respond until you greet them or solicit conversation, so too does the virtual assistant remain dormant until called to action. Given their omnipresence, many virtual assistants have prompted privacy concerns, as these devices are typically placed in homes and other private spheres, constantly using speech data that could be stored by the companies which developed the devices.

Different brands' virtual assistants have distinct voices and personalities, intended to help build trust between the user and machine and to promote interactions with an ease and seamlessness that might come from talking to someone helpful in daily life.

Human users tend to anthropomorphize AI-powered virtual assistants, and the fact that most AI voice assistants have female voices by default raises concerns about how gender biases are coded into technology products. A recent UNESCO report, for example, examines how AI voice assistants projected as young women perpetuate harmful gender biases, especially in light of persistent digital gender divides. As these technologies become more usable and frictionless, complex tradeoffs between ease/convenience and privacy/bias will need to be taken into consideration.

CPAI HUMAN-AI COLLABORATION FRAMEWORK
CASE 1: Virtual Assistants and Users

I. NATURE OF COLLABORATION

Stage of development or deployment

1. Is the AI fixed once deployed or evolving over time via model updates/continual interaction?
Evolving over time

2. To what extent is there ongoing collaboration between the AI's developer(s) and the AI system? [No collaboration, limited collaboration, moderate collaboration, active collaboration]
Likely active collaboration, as software updates are made frequently

3. Is the AI system currently used by people other than the AI's original developers? Yes

Goals

4. Are the goals of the human-AI collaboration clear or unclear?
Ostensibly clear, but there are many goals that users may weigh differently (e.g., information gathering, task management, etc.)

5. What is the nature of the collaboration's goals? [Physical, knowledge/intellectual, emotional, and/or motivational in nature] Knowledge/intellectual primarily, but could be physical (changes temperature or lighting in one's house, for example), or motivational if the user asks for a pump-up speech in advance of a scheduled task

6. Is empathy a precondition for the human-AI interaction to function as intended?
Not a precondition, but some sense of connection and emotional attentiveness is important for users to like and trust the device

7. Are the human and the AI system's goals aligned? Yes. The AI system is hoping to help the user, and the user hopes to get help from the device.

Interaction pattern

8. Is the collaboration repeated over time or is it a one-time engagement? If over time, at what time-scale? Repeated over time, but likely with different queries each time

9. Is the interaction concurrent – with both human and AI contributing in parallel – or does it depend on taking turns? Taking turns, call and response

Degree of agency

10. Does the AI or human agent contribute more to the system's decision-making? Action-taking? or action-taking? Both contribute. The human decides what to ask, but the machine has some agency in figuring out how to interpret the query and what amount of data to respond with/amount of context to provide

11. How much agency does the human have? The AI system? [None, limited, moderate, high, full]
Human has full agency (can shut the machine off or get rid of it) AI system has moderate agency (defining its agency based on its ability to "choose" which questions to answer and how)

II. NATURE OF SITUATION

Location and context

12. Are other people or other AI systems involved as third-parties? This can apply to either 1-1 collaborations or multiple people and an AI (e.g., in the instance of an AI teacher interacting with a classroom of human students, those human students would not be included as third-parties). Could be several people talking to the virtual assistant at once

13. Are the human and AI agents co-located physically or virtually?
If responding to speech, like the examples listed in the case, the virtual assistant is housed in a device that is co-located physically with the human

Awareness

14. Is the human likely aware that they are interacting with an AI system?
Yes, the human is aware (though children might not realize it is an inanimate object)

15. Does the human need to consent before interacting with the AI system? Yes

Consequences

16. How significant are the consequences should the AI fail to perform as designed/expected? [Low, moderate, high] Moderate, depending on the situation. If someone for example asks a virtual assistant to tell them when their flight is or an important appointment is, they may miss an event with grave consequences

17. How significant are the benefits of the AI to the users should it perform as designed/expected? [Low, moderate, high] Moderate (likely depends on who you ask). Disclaimer, I do not use a virtual assistant. Might free-up time and enable convenience for users

18. What are the potential consequences and benefits of the outcome of the collaboration?
Certain types of sensitive information could lead to grave consequences depending on the response the virtual agent triages and ultimately offers. For example there was a controvery around users asking Siri where the nearest abortion clinic is located and it failing to provide such information to users, which could offer grave consequences for women in need.

19. What might be the broader impacts of the human-AI collaboration? Greater ubiquity of virtual assistants across a variety of domains that triage and filter information in ways that may be different than humans do, with less contextual awareness but with greater access to information.

20. To what extent do typical users consider privacy and security when interacting with the AI agent? [Low, Moderate, High] Amongst users, low. If you were someone who considered privacy risk moderately or high, you would likely be reluctant to use the device

Assessment

21. Who is the main party or individual assessing the nature and effectiveness of the human-AI collaboration? User

22. Are assessments of the human-AI collaboration's outcome subjective or objective?
Could be both. Objective, if the device fails to retrieve accurate factual responses to one's inquiry. Subjective, if a user does not like how the device communicates such information, or does not trust it or its personality.

Level of trust

23. Are both the human and the AI agent trusting and trustworthy? AI trustworthiness can be defined broadly, driven by task competence, safety, authority, and authenticity, amongst other features (e.g., we know an AI comes from the same affiliation it claims to be from). Yes, if the virtual assistant successfully answers the user's queries

III. AI-SYSTEM CHARACTERISTICS

Interactivity

24. What is the mode of the interaction between the two agents? [Via screen, voice, wearables, virtual reality, or something else] Most often, via voice and speaker-like devices (or one's phone)

25. Could the nature of the data that the AI system operates over impact its interactivity? E.g., a wearable AI may interact with the user via voice, but may make inferences over voice as well as sensor data. Yes

Adaptability

26. Is the AI system passively providing information or proactively anticipating the next steps of the interaction? Passively providing information

Performance

27. How predictable is the AI system? [Low, moderate, high] Moderate

28. Does the system often produce false-positives? False-negatives? Yes, to both (gives inaccurate information after interpreting a query properly; does not interpret a query properly)

Explainability

29. Can the AI system communicate its confidence levels to a human? Not really, but sometimes it can preface its answer by saying "can you add more detail?" or something like that to clarify or get more information

30. How does the AI system communicate its decision-making process and inputs to that decision-making process to the human? It can sometimes imply that it is uncertain or looking up an answer, in order to buy-time or communicate with users in a more human-like fashion

Personification

31. How human-like is the AI system? [Not very, moderately, or highly human-like]
Moderately, based on speech-based communication (not in terms of physical embodiment)

32. How easily anthropomorphized is the AI system? Easily

IV. HUMAN CHARACTERISTICS

Age

33. Is the person(s) collaborating with the AI system a child (under 18), an adult (18 - 65), or a senior (over 65)? Some mixture of the above, when >1 person collaborating with the AI system?
Could be all. Children may be less likely to understand that the voice in the machine is not a real human on the other end of the line

Differently-abled

34. Does the person collaborating with the AI have special needs or accommodations? Possibly

Culture

35. Are there cultural consistencies/norms for those collaborating with the AI system?

Yes. Firstly, the devices are now able to respond to speech in many different languages. Certain linguistic nuances and ways of collaborating may vary depending on the cultural context in which the users and the device are situated

36. What level of previous technology interaction has the user(s) of the system had? [Low, moderate, high] Likely high if using a personal assistant

CASE 2: Mental Health Chatbots and Users *Yoonsuck Choe, Corporate Vice President, Samsung Research (Samsung)*

Context/Scenario

Myriad AI-driven technologies serve as personalized mental health services, whether as companions for those suffering from diagnosed mental health issues or as tools for those seeking mental health resources without a clinical diagnosis. Many such systems are AI-driven chatbots that remember and adjust responses based upon their interactions with the user. The goal of mental health chatbots is seemingly apparent: to assess and improve the mental health of its users. These tools work effectively when they connect with the user, and such a connection is fostered as the user provides increasingly rich data as input to the system over time. The ubiquity of smartphones combined with advancements in natural language processing allow for such applications to be increasingly effective and popular. Advanced AI techniques enable this type of accessible, personalized emotional support in the palm of a user's hands.

AI System

Woebot is one such example of an AI-driven chatbot for mental health, created by a clinical psychologist and integrated with Facebook Messenger to replicate conversations a patient might have with an in-person therapist. It asks questions about the user's mood, feelings, and thoughts, and even reflects that it is listening while using techniques based upon cognitive behavioral therapy (CBT); it behaves this way while emphasizing that it is not a replacement for human connection and conveying a limit to its services.

Human-AI Collaboration

Building trust and empathy between human and therapist is key to an in-person mental health relationship and also vital to the effectiveness and prolonged use of chatbot mental health services. AI chatbots' effectiveness for mental health purposes is influenced by trust and an empathic connection between human and AI. A users ability to connect and respond to the level of artificial empathy projected by the system influences the use and effectiveness of such a system. In order to use the system effectively, users should trust such AI chatbots in the same way that they would trust a human therapist sworn to respect patient confidentiality. Mental health chatbot reviews often describe them as seeming artificial or scripted in their responses. While improvements to natural language processing will increase adaptability and the quality of responses, inducing empathy between chatbot and human might be a different challenge than inducing empathy between a human patient and human therapist. Promisingly, Morris, Kouddous, Kshirsagar, & Schueller (2018) have conducted work on how "conversational agents can express empathy in nuanced ways that account for the unique circumstances of the user," a technique that can be used to build robust digital mental health chatbots. It will be important to understand how users assess such systems and what induces empathy and trust in the collaboration, characteristics vital to a successful human-human mental health intervention. Ultimately this may allow society to reap the benefits of cost-effective, accessible, and ubiquitous mental health chatbots.

CPAI HUMAN-AI COLLABORATION FRAMEWORK
CASE 2: Mental Health Chatbots and Users

I. NATURE OF COLLABORATION

Stage of development or deployment

1. Is the AI fixed once deployed or evolving over time via model updates/continual interaction?
Evolving over time via model updates/continual interaction (since the chatbot needs to continually update its responses based on the specific user interacting with the system)

2. To what extent is there ongoing collaboration between the AI's developer(s) and the AI system? [No collaboration, limited collaboration, moderate collaboration, active collaboration]
Active collaboration (since developer may have to be aware of unexpected questions from users, and also actively update the system based on the latest scientific findings)

3. Is the AI system currently used by people other than the AI's original developers?
Yes (there are several mental health chatbot services offered).

Goals

4. Are the goals of the human-AI collaboration clear or unclear?
Yes (To assess and improve mental health conditions)

5. What is the nature of the collaboration's goals? [Physical, knowledge/intellectual, emotional, and/or motivational in nature] All of the following: knowledge/intellectual, emotional, and/or motivational in nature

6. Is empathy a precondition for the human-AI interaction to function as intended?
Somewhat, depending on the expected level of engagement.

7. Are the human and the AI system's goals aligned? Yes

Interaction pattern

8. Is the collaboration repeated over time or is it a one-time engagement? If over time, at what time-scale? Repeated over time; potentially life-long

9. Is the interaction concurrent – with both human and AI contributing in parallel – or does it depend on taking turns? Taking turns

Degree of agency

10. Does the AI or human agent contribute more to the system's decision-making? Action-taking? Both. The human agent offers information to the system that then makes decisions and acts

11. How much agency does the human have? The AI system? [None, limited, moderate, high, full]
Human: Full (but isn't this always the case?)
AI: Unknown, as it depends, since no AI so far can be said to possess human-level agency

II. NATURE OF SITUATION

Location and context

12. Are other people or other AI systems involved as third-parties? This can apply to either 1-1 collaborations or multiple people and an AI (e.g., in the instance of an AI teacher interacting with a classroom of human students, those human students would not be included as third-parties). Depends on the situation: No (This kind of consultation is strictly private in most cases.) / Yes. (Sometimes, human therapist may participate in the dialog)

13. Are the human and AI agents co-located physically or virtually?
Virtually (AI is a web-service or a mobile app, not a physical robot).

Awareness

14. Is the human likely aware that they are interacting with an AI system? Yes

15. Does the human need to consent before interacting with the AI system?
Yes (Any medical advice needs strong consent, perhaps even by law)

Consequences

16. How significant are the consequences should the AI fail to perform as designed/expected? What are those consequences? [Low, moderate, high] Moderate or High, depending on what the user expected from the chatbot. The consequences are severe if the expectation was high and it didn't live up to these expectations, possibly worsening the mental health condition.

17. How significant are the benefits of the AI to the users should it perform as designed/expected? What are those benefits? [Low, moderate, high]
High. Benefits include greatly improved mental health and user happiness.

18. What are the potential consequences and benefits of the outcome of the collaboration?
From the user's point of view, improved mental health. From the AI's point of view (developers, etc.), the chance to improve the system's performance to benefit other users or future users.

19. What might be the broader impacts of the human-AI collaboration? Generally happier, safer society.

20. To what extent do typical users consider privacy and security when interacting with the AI agent? [Low, Moderate, High] High

Assessment

21. Who is the main party or individual assessing the nature and effectiveness of the human-AI collaboration? Professional (psychotherapist).

22. Are assessments of the human-AI collaboration's outcome subjective or objective?
Partly subjective (only user can answer whether they are happier). Partly objective (behavioral observation can serve as indirect assessment).

Level of trust

23. Are both the human and the AI agent trusting and trustworthy? AI trustworthiness can be defined broadly, driven by task competence, safety, authority, and authenticity, amongst other features (e.g., we

know an AI comes from the same affiliation it claims to be from). Yes. Both must trust each other for the best outcome

III. AI SYSTEM CHARACTERISTICS

Interactivity

24. What is the mode of the interaction between the two agents? [Via screen, voice, wearables, virtual reality, or something else] Via screen (and maybe in the future, by voice)

25. Could the nature of the data that the AI system operates over impact its interactivity? E.g., a wearable AI may interact with the user via voice, but may make inferences over voice as well as sensor data. Yes. For text only chat, word choice, etc. would be the most important. For voice chat, emotional tone/prosody of the voice may also be very important. The chatbot may also show calming images/scenes to the user

Adaptability

26. Is the AI system passively providing information or proactively anticipating the next steps of the interaction? Proactive.

Performance

27. How predictable is the AI system to the user? [Low, moderate, high]
Moderate. If too predictable or too unpredictable, user can quickly lose interest

28. Does the system often produce false-positives? False-negatives? N/A

Explainability

29. Can the AI system communicate its confidence levels to a human?
Perhaps. The user can ask "are you sure?"

30. How does the AI system communicate its decision-making process and inputs to that decision-making process to the human? The user may ask why the chatbot thinks so, or why it is providing such advice. The chatbot may then have to explain why.

Personification

31. How human-like is the AI system? [Not very, moderately, or highly human-like]
Highly human-like (in terms of the dialogue)

32. How easily anthropomorphized is the AI system? Quite easily, of course depending on the user. Recall the story when ELIZA was first deployed decades ago. Although it was a simple pattern-matching based chatbot, it was highly-engaging to naive users and they thought it was almost human.

IV. HUMAN CHARACTERISTICS

Age

33. Is the person(s) collaborating with the AI system a child (under 18), an adult (18 - 65), or a senior (over 65)? Some mixture of the above, when >1 person collaborating with the AI system? All ages

Differently-abled

34. Does the person collaborating with the AI have special needs or accommodations? Depends

Culture

35. Are there cultural consistencies/norms for those collaborating with the AI system?
Within a single culture, mostly yes, but across different cultures, the norms may differ
Within a single culture, depending on the level of education, etc. the norms may differ a bit

36. What level of previous technology interaction has the user(s) of the system had?[Low, moderate, high] Varies, since mental health issues affect all walks of life

CASE 3:
Intelligent Tutoring Systems and Learners *Amber Story, Associate Executive Director for Scientific Affairs, (American Psychological Association)*

Context/Scenario

One-on-one tutoring is an effective way to promote learning and performance. More and more, the tutor in question is a non-human, intelligent tutoring system (ITS), sometimes taking the form of a social robot or virtual agent powered by artificial intelligence. Today's ITS are both educational and entertaining, captivating student interest and enhancing engagement to maximize learning. Intelligent tutoring systems can be used to promote learning across multiple domains, including language, mathematics, science, and even social and procedural skills. The potential for ITS to teach a variety of types of knowledge – including procedural knowledge, factual knowledge, problem solving skills, and even problem posing skills – means that even within this seemingly contained use case, there may be different forms of interaction necessary for the various sub-tasks that constitute learning.

AI System

There are numerous ITS available, including ALEKS, Cognitive Tutor, and AutoTutor to name just a few. Most systems have three components in common; they have an expert model that represents the content of the curriculum, a student model based on the assessment of an individual student's knowledge and skill level, and a pedagogical model that represents the method of instruction. The instruction process is a continual loop of the presentation of materials and assessments so that the student model is dynamically updated to represent the current state of knowledge.

Human-AI Collaboration

Intelligent tutoring systems are capable of presenting problems or scenarios, monitoring inputs from the student and adapting their behavior accordingly, and providing feedback to the student on how they are performing. More advanced systems can adjust their actions in response to a student's speech (e.g., pitch, tempo), and facial expressions - features that supposedly signal changes in students' level of attention, frustration, and engagement. For example, an "empathetic" system might mimic some of the students' expressions or gestures that suggest boredom and then suggest a change of problem or story to one that the student will find more engaging. The collaboration between a student and an ITS can be effective in producing interactive and lasting learning. Research suggests that collaborating with an ITS can be as effective for mastering skills and knowledge as working one-on-one with a human tutor (VanLehn, 2011), and it is typically far more effective than learning in a traditional classroom setting (Bloom, 1984, for reference to the traditional classroom setting).

CPAI HUMAN-AI COLLABORATION FRAMEWORK
CASE 3: Intelligent Tutoring Systems and Learners

I. NATURE OF COLLABORATION

Stage of development or deployment

1. Is the AI fixed once deployed or evolving over time via model updates/continual interaction?
Could be either fixed or evolving

2. To what extent is there ongoing collaboration between the AI's developer(s) and the AI system? [No collaboration, limited collaboration, moderate collaboration, active collaboration]
It may vary, but it would make sense if the developers could push upgrades

3. Is the AI system currently used by people other than the AI's original developers? Yes

Goals

4. Are the goals of the human-AI collaboration clear or unclear?
Goals are clear, as these are all instructional with educational/learning goals

5. What is the nature of the collaboration's goals? [Physical, knowledge/intellectual, emotional, and/or motivational in nature] Knowledge/intellectual primarily but could also be social, emotional and or motivational in nature

6. Is empathy a precondition for the human-AI interaction to function as intended? Not a precondition

7. Are the human and the AI system's goals aligned? I'm not sure what the AI's goals would be.

Interaction pattern

8. Is the collaboration repeated over time or is it a one-time engagement? If over time, at what time-scale? During the instructional period, there would be repeated interactions, but the collaboration may be a one-time or across time interaction

9. Is the interaction concurrent – with both human and AI contributing in parallel – or does it depend on taking turns? Collaboration is fundamentally turn-taking, with response and feedback

Degree of agency

10. Does the AI or human agent contribute more to the system's decision-making? Action-taking? or action-taking? I am not sure, as they both contribute.

11. How much agency does the human have? The AI system? [None, limited, moderate, high, full]
The human has high agency, the AI can vary between no agency to high

II. NATURE OF SITUATION

Location and context

12. Are other people or other AI systems involved as third-parties? This can apply to either 1-1 collaborations or multiple people and an AI (e.g., in the instance of an AI teacher interacting with a classroom of human students, those human students would not be included as third-parties).
There might be a teacher supervising the ITS - student collaboration.

13. Are the human and AI agents co-located physically or virtually? It could be either physical (such as when there is a robotic ITS), or virtual (such as when the ITS is an avatar on a screen).

Awareness

14. Is the human likely aware that they are interacting with an AI system? Yes, the human is aware.

15. Does the human need to consent before interacting with the AI system?
I would think in most cases, the human consents out of the fact that they are aware they are interacting with the system and continue the collaboration.

Consequences

16. How significant are the consequences should the AI fail to perform as designed/expected? [Low, moderate, high] Failure to learn may have differing levels of consequences, but unlikely to be highly consequential, more Low to Moderate.

17. How significant are the benefits of the AI to the users should it perform as designed/expected? [Low, moderate, high] Depending on what is being taught, the benefits could be quite high.

18. What are the potential consequences and benefits of the outcome of the collaboration?
Benefit would be individualized tutoring, which is effective and can be scaled.

19. What might be the broader impacts of the human-AI collaboration?
More effective and efficient learning platforms.

20. To what extent do typical users consider privacy and security when interacting with the AI agent? [Low, Moderate, High] Low to Moderate

Assessment

21. Who is the main party or individual assessing the nature and effectiveness of the human-AI collaboration? Student, teacher, parent

22. Are assessments of the human-AI collaboration's outcome subjective or objective?
Depending on what is being taught, it might be either. However, for most instructional domains, the outcomes can be assessed objectively

Level of trust

23. Are both the human and the AI agent trusting and trustworthy? AI trustworthiness can be defined broadly, driven by task competence, safety, authority, and authenticity, amongst other features (e.g., we know an AI comes from the same affiliation it claims to be from). Human likely to be trusting, but may not be trustworthy if it tries to trick the AI system. AI likely to be trusting and trustworthy

III. AI SYSTEM CHARACTERISTICS

Interactivity

24. What is the mode of the interaction between the two agents? [Via screen, voice, wearables, virtual reality, or something else] Via screen/embodied system through voice, text, and touch

25. Could the nature of the data that the AI system operates over impact its interactivity? E.g., a wearable AI may interact with the user via voice, but may make inferences over voice as well as sensor data. Yes

Adaptability

26. Is the AI system passively providing information or proactively anticipating the next steps of the interaction? Depending on how advanced the system is, it could be proactively anticipating

Performance

27. How predictable is the AI system? [Low, moderate, high] ITS are usually high performing, predictable systems

28. Does the system often produce false-positives? False-negatives? I imagine that this varies by system and commercial producer

Explainability

29. Can the AI system communicate its confidence levels to a human?
I would not think that it would need to explain its feedback

30. How does the AI system communicate its decision-making process and inputs to that decision-making process to the human? I don't think it does

Personification

31. How human-like is the AI system? [Not very, moderately, or highly human-like]
Could vary, but usually portrayed as a person or animal-like creature

32. How easily anthropomorphized is the AI system? Easily

IV. HUMAN CHARACTERISTICS

Age

33. Is the person(s) collaborating with the AI system a child (under 18), an adult (18 - 65), or a senior (over 65)? Some mixture of the above, when >1 person collaborating with the AI system?
Age range could vary substantially, though often used with children.

Differently-abled

34. Does the person collaborating with the AI have special needs or accommodations?
This is possible

Culture

35. Are there cultural consistencies/norms for those collaborating with the AI system?
As learning involves social cognition, cultural norms would need to be considered and incorporated into the collaboration.

36. What level of previous technology interaction has the user(s) of the system had? [Low, moderate, high] Varies, but likely moderate.

CASE 4: Assistive Computing and Motor Neuron Disease Patients

Lama Nachman, Intel Fellow, Director of Anticipatory Computing Lab (Intel)

Context

Augmentative and Alternative Communication (AAC) systems enable people with motor neuron disease (MND) to communicate and connect with the world. There are more than three million people worldwide living with MND, which occurs when specialist nerve cells in the brain and spinal cord – motor neurons – stop working properly. Amyotrophic lateral sclerosis (ALS) is one form of MND and impacts people's ability to perform basic functions like gripping, walking, breathing, swallowing, and speaking.

AAC systems typically rely on gaze tracking to input text and perform different tasks, while other modalities are used as people lose their ability to control their gaze or keep their eyelids open. Users can leverage these modalities to trigger a computer interface enabling them to access different computer applications. One main application helps people with MND communicate with others as they gradually lose their ability to speak. Users type what they want to say, and then trigger a text to speech system (TTS) that speaks their words aloud. In early stages of the disease, when users are still able to speak clearly, they typically rely on speech recognition to interact with the system and supply commands or dictate text.

AI System

AI has a large role to play in AAC systems. Deep learning has improved TTS systems dramatically over the last few years by enabling people to personalize the voice from these systems and make it sound very close to a person's actual voice. Additionally, people can bank their voice while they can still speak normally and have a TTS system that mimics their voice, enabling them to retain their voice long after they lose their ability to speak.

Since every interaction with the machine is very cumbersome and costly, predictive text plays a huge role in improving communication efficiency; word prediction in the interface can dramatically reduce the number of characters users need to type. In fact, in Stephen Hawking's AAC system aided by word prediction, he had to type less than 10% of all characters. The accuracy of speech recognition systems has also improved dramatically, especially in realistic settings with far field (distant) speech, due to deep learning and improvements in microphone arrays. All of these capabilities are needed to create assistive systems as they stand today. They also enable a new approach to assistive computing that is more collaborative than existing systems, without the need for fine grained micro-management of the AI system.

Human-AI Collaboration

Today, assistive systems are still very rudimentary. They assume that people want to micro-manage every detail of the interaction, and focus on replacing existing human-computer interaction paradigms (keyboard and mouse) with an alternative modality like gaze control. This approach can result in very slow communication and hinder the ability for spontaneous communication. Imagine saying something to a person who is using AAC, and waiting for a long time before you hear a response back from the user due to the inefficiency of input. The AAC user is also typically focused on the text entry, reducing the potential for human connection.

To enable more efficient and spontaneous communication, one approach is to think of a communication system as a highly collaborative human/AI system. The human will not interact with the system at the letter and word level most of the time, but at the topic/sentence level. The AI system can "listen in" on the conversation, utilize automatic speech recognition to understand what the other person is saying, and surface reasonable responses for the user to choose from. The user should be able to nudge the AI system with minimal interaction to guide it in the desired direction, enabling further refinement as the conversation unfolds. Furthermore, the AI system can utilize its memory of previous conversations with this specific person, in addition to its vast knowledge from more generic conversations in the world, to improve its performance.

As the AI system learns from previous interactions, it would continue to refine its predictions and suggestions over time based on experience and feedback. For example, a user might use a recommended sentence for the sake of efficiency in the moment, but flag to the system that the recommended sentence was not an optimal suggestions. The system can then query the user later, when there is more time, to help improve its performance. This capability is important, since otherwise the system will become more constrained over time. The AI system will also need to be empathetic to the needs of the user as well as enable the user to show empathy and affect in communication with others. This can include enabling the user to specify the emotion and reflect that in the voice, understand the sentiment from the text and make suggestions to the user, and/or recognize the emotions of others and utilize this knowledge in the suggested response to the user.

CPAI HUMAN-AI COLLABORATION FRAMEWORK
CASE 4: Assistive Computing and Motor Neuron Disease Patients

I. NATURE OF COLLABORATION

Stage of development or deployment

1. Is the AI fixed once deployed or evolving over time via model updates/continual interaction?
Evolves over time
2. To what extent is there ongoing collaboration between the AI's developer(s) and the AI system? [No collaboration, limited collaboration, moderate collaboration, active collaboration]
Moderate collaboration. I would imagine that the generic prediction models would continue to evolve, but the personalized model will need to be done automatically without developer involvement to allow the system to scale

3. Is the AI system currently used by people other than the AI's original developers?
Yes, we have users engaged in the development phase usually to give feedback. In the original ACAT system, Stephen Hawking was directly involved, but the AI capabilities were limited to word prediction and gesture recognition. In the new system, we are working actively with another user with ALS.

Goals

4. Are the goals of the human-AI collaboration clear or unclear?
Clear: To provide a good prediction in a timely manner, enable the user to nudge the system in real-time to provide a reasonable response.

5. What is the nature of the collaboration's goals? [Physical, knowledge/intellectual, emotional, and/or motivational in nature] Knowledge/intellectual and emotional

6. Is empathy a precondition for the human-AI interaction to function as intended?
Empathy is important both in the interaction as well as the output of the system since it is used for communication with other users

7. Are the human and the AI system's goals aligned? Yes

Interaction pattern

8. Is the collaboration repeated over time or is it a one-time engagement? If over time, at what time-scale? Repeated over time, during conversations with others at the utterance level, and later on at lower frequency to help train the system with more input from the user

9. Is the interaction concurrent – with both human and AI contributing in parallel – or does it depend on taking turns? Depends on what we mean by this actually, both will be working at the same time, but there is a back and forth as well.

Degree of agency

10. Does the AI or human agent contribute more to the system's decision-making? Action-taking? Hard to tell, it depends on how we measure contribution, if we think in terms of letters generated then the AI, but if we think in terms of concepts and high level direction, it would be the human.

11. How much agency does the human have? The AI system? [None, limited, moderate, high, full]
Hard to tell. On one hand, the human will have full agency because they can actually resort to a lower level interaction and type what they want. However, this will require a lot of time due to the disability, which in some sense reduces their agency. The AI system will have moderate agency since it will make recommendations but ultimately the user will choose the output.

II. NATURE OF SITUATION

Location and context

12. Are other people or other AI systems involved as third-parties?This can apply to either 1-1 collaborations or multiple people and an AI (e.g., in the instance of an AI teacher interacting with a classroom of human students, those human students would not be included as third-parties). Yes, since this will enable human to human communication, others communicating with the user will be involved, their speech will be analyzed and they will hear the output of the system.

13. Are the human and AI agents co-located physically or virtually? Physically

Awareness

14. Is the human likely aware that they are interacting with an AI system? This is an interesting question. The user of the system is clearly aware of the interaction with an AI system, since it is making recommendations. However, the other people who are interacting with the user might not know, and simply assume that the system is simply being used for typing the user's thoughts.

15. Does the human need to consent before interacting with the AI system? Yes. Since the system will be recording the other person's speech and using that to recommend responses to the user, I think people should be aware and give their consent.

Consequences

16. How significant are the consequences should the AI fail to perform as designed/expected? What are those consequences? [Low, moderate, high] Moderate. Since the user can always fall back on direct entry, the impact will be a loss of efficiency and spontaneity rather than an inability to communicate. However, due to the limited ability, this could case quite a bit of hardship for the user.

17. How significant are the benefits of the AI to the users should it perform as designed/expected? What are those benefits? [Low, moderate, high] High; This is a major hardship that people with MND state all the time. Connecting with their loved ones and with the world, being able to express their thoughts and be independent is what's at stake here.

18. What are the potential consequences and benefits of the outcome of the collaboration? A good balance between communication efficiency and an ability to express themselves accurately

19. What might be the broader impacts of the human-AI collaboration? Improved independence, human connection, and quality of life

20. To what extent do typical users consider privacy and security when interacting with the AI agent? [Low, Moderate, High] Currently low, because it is mainly focused on language models and prediction. However, as the intelligence increases, users and people who interact with the users should really consider the privacy issues more deeply.

Assessment

21. Who is the main party or individual assessing the nature and effectiveness of the human-AI collaboration? The user

22. Are assessments of the human-AI collaboration's outcome subjective or objective? Largely subjective because it all depends on how people interpret an "acceptable response" versus a "specific response". However, it is possible to articulate objective measures in terms of utilization of automatically generated responses, effectiveness of nudging, relying on letter/word based interaction

Level of trust

23. Are both the human and the AI agent trusting and trustworthy? AI trustworthiness can be defined broadly, driven by task competence, safety, authority, and authenticity, amongst other features (e.g., we know an AI comes from the same affiliation it claims to be from). Not sure how to answer this, while the intention is a good one, such a system can be misused, with risks to the user and other human participants in the interaction.

III. AI SYSTEM CHARACTERISTICS

Interactivity

24. What is the mode of the interaction between the two agents? [Via screen, voice, wearables, virtual reality, or something else] Screen, voice, and possibly virtual reality

25. Could the nature of the data that the AI system operates over impact its interactivity? E.g., a wearable AI may interact with the user via voice, but may make inferences over voice as well as sensor data. Yes

Adaptability

26. Is the AI system passively providing information or proactively anticipating the next steps of the interaction? Proactively anticipating, recommending responses, and possibly predicting what the other person is saying to reduce the latency of interaction and engage the user earlier.

Performance

27. How predictable is the AI system? [Low, moderate, high] Moderate

28. Does the system often produce false-positives? False-negatives? Yes

Explainability

29. Can the AI system communicate its confidence levels to a human? Yes

30. How does the AI system communicate its decision-making process and inputs to that decision-making process to the human? It shows multiple options enabling the user to make the final choice by selecting an option or choosing to enter something different altogether.

Personification

31. How human-like is the AI system? [Not very, moderately, or highly human-like] Highly human-like.

32. How easily anthropomorphized is the AI system? I think this is easy because it is predicting human responses, sounding like a human and expressing affect

IV. HUMAN CHARACTERISTICS

Age

33. Is the person(s) collaborating with the AI system a child (under 18), an adult (18 - 65), or a senior (over 65)? Some mixture of the above, when >1 person collaborating with the AI system?
It could be any age, right now we are targeting adults since we assume a certain level of literacy and most people with MND are adults. Later on this can be tailored for children with AAC needs.

Differently-abled

34. Does the person collaborating with the AI have special needs or accommodations?
Yes, the system is meant for people with disabilities, requiring the AI system to communicate with people

Culture

35. Are there cultural consistencies/norms for those collaborating with the AI system?
These systems are typically needed across cultures, however the predictions should be cognizant of cultures since human responses are highly impacted by the cultural norms

36. What level of previous technology interaction has the user(s) of the system had? [Low, moderate, high] Assuming moderate to high, ideally we should figure out how to bring this to low.

CASE 5: AI Drawing Tools and Artists
Philipp Michel, Project Assistant Professor, (University of Tokyo)

Context/Scenario

Given the strong historical and recent focus of AI methods on visual processing (particularly, detection/recognition/segmentation in images and video), an application of AI methods related to the visual arts is a natural use case.

The arts domain has generally more generously defined success criteria than others (e.g., health), requiring merely that the result of AI-generated visual art be described as interesting or visually pleasing. This domain provides for a more experimental and free-form application of AI technologies and serves to explain the explosion of visual AI applications over the past few years.

Visual AI approaches range from discriminative to generative (e.g., train a neural network on impressionist art, then generate new paintings). Of particular interest in our context, are approaches that take partially complete human artwork and extend, re-shape, or complete these artworks using AI technology in a collaborative fashion. However, although the order may of course be reversed or even interleaved as well.

AI System

Sketch-RNN and Magic Sketchpad are two AI systems that use a recurrent neural network to allow one to draw in tandem with the algorithm. The algorithm allows a human user to select from about 120 categories of objects to draw. Once the user starts drawing, the algorithm will attempt to advance or complete your sketch drawing. The user can continue to take turns with the network as they draw the sketch. The model underlying this practice was trained on millions of sketches from a popular sketch game in a stroke-sequential fashion, attempting to model how humans doodle. The model is thus able to come up with a wide variety of sketches and can also attempt to mimic one's own particular style of drawing. By selecting sketches as an application area, it attempts to capture one of the simplest forms of human visual art. Very similar principles can also be applied to predictive drawing of Chinese and Japanese characters.

Human-AI Collaboration

The main focus in the collaboration between the human and the AI system in the Sketch-RNN case is efficiency/speed of sketch completion and joint creative exploration. The most interesting collaboration happens in cases when the AI draws unexpected strokes that take the sketch in an unplanned direction. While in the Sketch-RNN example the task performed by the human is exactly the same as the task performed by the AI system (drawing via a sequence of sketch strokes), this certainly need not be the case. Other systems may leverage the AI system to perform drawing tasks very quickly that would be time-consuming for humans (e.g. auto-coloring of pencil sketches, 3D models from 2D sketches, etc).

CPAI HUMAN-AI COLLABORATION FRAMEWORK
CASE 5: AI Drawing Tools and Artists

I. NATURE OF COLLABORATION

Stage of development or deployment

1. Is the AI fixed once deployed or evolving over time via model updates/continual interaction?
Fixed once deployed

2. To what extent is there ongoing collaboration between the AI's developer(s) and the AI system? [No collaboration, limited collaboration, moderate collaboration, active collaboration]
Moderate collaboration

3. Is the AI system currently used by people other than the AI's original developers? Yes

Goals

4. Are the goals of the human-AI collaboration clear or unclear? Clear

5. What is the nature of the collaboration's goals? [Physical, knowledge/intellectual, emotional, and/or motivational in nature] Motivational and emotional

6. Is empathy a precondition for the human-AI interaction to function as intended? No

7. Are the human and the AI system's goals aligned? Yes

Interaction pattern

8. Is the collaboration repeated over time or is it a one-time engagement? If over time, at what time-scale? Over time, as user chooses, each interaction lasting on the order of minutes

9. Is the interaction concurrent – with both human and AI contributing in parallel – or does it depend on taking turns? Turn-taking

Degree of agency

10. Does the AI or human agent contribute more to the system's decision-making? Action-taking? or action-taking? Depends on each interaction, can be either

11. How much agency does the human have? The AI system? [None, limited, moderate, high, full] Both limited

II. NATURE OF SITUATION

Location and context

12. Are other people or other AI systems involved as third-parties?
This can apply to either 1-1 collaborations or multiple people and an AI (e.g., in the instance of an AI teacher interacting with a classroom of human students, those human students would not be included as third-parties). No

13. Are the human and AI agents co-located physically or virtually? Co-located virtually

Awareness

14. Is the human likely aware that they are interacting with an AI system? Yes

15. Does the human need to consent before interacting with the AI system? No

Consequences

16. How significant are the consequences should the AI fail to perform as designed/expected? [Low, moderate, high] Low

17. How significant are the benefits of the AI to the users should it perform as designed/expected? [Low, moderate, high] Low

18. What are the potential consequences and benefits of the outcome of the collaboration? Benefits: creativity and entertainment

19. What might be the broader impacts of the human-AI collaboration? Augmented artistic skills, boosted creativity

20. To what extent do typical users consider privacy and security when interacting with the AI agent? [Low, Moderate, High] Low

Assessment

21. Who is the main party or individual assessing the nature and effectiveness of the human-AI collaboration? Human participant

22. Are assessments of the human-AI collaboration's outcome subjective or objective? Subjective

Level of trust

23. Are both the human and the AI agent trusting and trustworthy? AI trustworthiness can be defined broadly, driven by task competence, safety, authority, and authenticity, amongst other features (e.g., we know an AI comes from the same affiliation it claims to be from). Yes

III. AI SYSTEM CHARACTERISTICS

Interactivity

24. What is the mode of the interaction between the two agents? [Via screen, voice, wearables, virtual reality, or something else] Via screen and input

25. Could the nature of the data that the AI system operates over impact its interactivity? E.g., a wearable AI may interact with the user via voice, but may make inferences over voice as well as sensor data. Yes

Adaptability

26. Is the AI system passively providing information or proactively anticipating the next steps of the interaction? Mostly passively providing information

Performance

27. How predictable is the AI system? [Low, moderate, high] Moderately predictable

28. Does the system often produce false-positives? False-negatives? N/A as it is art

Explainability

29. Can the AI system communicate its confidence levels to a human? No

30. How does the AI system communicate its decision-making process and inputs to that decision-making process to the human? Visually (though it might not depict previous drawings when "communicating" why/ how it chose to draw something specifically, rather just display it).

Personification

31. How human-like is the AI system? [Not very, moderately, or highly human-like] Not very

32. How easily anthropomorphized is the AI system? Not easily

HUMAN CHARACTERISTICS

Age

33. Is the person(s) collaborating with the AI system a child (under 18), an adult (18 - 65), or a senior (over 65)? Some mixture of the above, when >1 person collaborating with the AI system?
Any

Differently-abled

34. Does the person collaborating with the AI have special needs or accommodations? No

Culture

35. Are there cultural consistencies/norms for those collaborating with the AI system? No

36. What level of previous technology interaction has the user(s) of the system had? [Low, moderate, high] Moderate

CASE 6:
Magnetic Resonance Imaging (MRI) and Doctors
Bendert Zevenbergen, Visiting Research Fellow (Princeton Center for Information Technology Policy)

Context/Scenario

Magnetic resonance imaging (MRI) is an imaging technique used to observe and detect a variety of diseases by identifying tumors and lesions in body tissues and organs. MRI scanners use radio waves and a strong magnet to generate signals from body tissues, which are then translated into a detailed, three-dimensional image. Patients must lie completely still for up to an hour while doctors and their staff operate the MRI machine and generate the scans. Currently, MRI machine output is a noisy and grainy image which requires a doctor's interpretation. AI image reconstruction technology, due to its ability to produce accurate images from under-sampled data, can potentially speed up and improve the MRI scan process.

AI System

In 2018, Facebook and NYU School of Medicine's Department of Radiology launched fastMRI, a project that investigates AI's potential to improve the MRI scanning process. FastMRI leverages AI to create images of a similar quality to a traditional scan, and to do so by collecting only a fraction of the data typically needed, therefore enabling doctors to conduct the scans more quickly.

With this process, convolutional neural networks (CNNs), specialized for processing image data, are used to create images from reduced amounts of data. The AI system interprets some of the data for the doctor and predicts other missing data during the scan. These features enable faster image creation, and allow a doctor to interpret the data and the image more quickly as well. Increasing the speed of MRI scans would enable increased access to such imaging for a greater number of patients. However, creating an image from less data than is usually used to derive diagnoses may pose a risk if it misses something critical for diagnostic purposes.

Human-AI Collaboration

AI use in MRI involves doctors as well as patients. The doctor and AI system collaborate to interpret a reduced amount of MRI data. AI-driven images allow the doctor to interpret clearer images than the non-AI produced, grainy images; doctors can view the clearer image more quickly and determine whether or not there is an abnormality. Doctors, enabled by this AI system, would likely be able to conduct and analyze more scans per day, increasing their productivity and hopefully, in the process, diagnostic accuracy. AI use in MRI also reduces the time that patients spend in machines to receive a scan. Faster MRI scans may replace X-ray or CT exams, which are currently conducted due to their speed, and come with harmful exposure to radiation not present with MRIs. Other research studies have explored using AI to help doctors analyze tuberculosis in X-rays or detect breast cancer metastasis for pathologists looking at tissue samples.

For this framework, I chose to focus on interactions between the AI MRI system and doctors, recognizing that another interpretation could consider interactions between the AI system and the patient.

CPAI HUMAN-AI COLLABORATION FRAMEWORK
CASE 6: Magnetic Resonance Imaging (MRI) and Doctors

I. NATURE OF COLLABORATION

Stage of development or deployment

1. Is the AI fixed once deployed or evolving over time via model updates/continual interaction?
Likely fixed (although research project may evolve)

2. To what extent is there ongoing collaboration between the AI's developer(s) and the AI system?
Likely active collaboration, because it is an ongoing research project.

3. Is the AI system currently used by people other than the AI's original developers?
Yes, the doctors and nurses using the system

Goals

4. Are the goals of the human-AI collaboration clear or unclear? Clear

5. What is the nature of the collaboration's goals? [Physical, knowledge/intellectual, emotional, and/or motivational in nature]. Knowledge

6. Is empathy a precondition for the human-AI interaction to function as intended?
No, though professional intuition likely is

7. Are the human and the AI system's goals aligned? Yes (detecting anomalies in bodies)

Interaction pattern

8. Is the collaboration repeated over time or is it a one-time engagement? If over time, at what time-scale? One-time per patient. Repeated over time on different patients, or returning patients

9. Is the interaction concurrent – with both human and AI contributing in parallel –or does it depend on taking turns? Concurrent. The AI system likely needs to finish its data collection and analysis before the doctor can interpret the results.

Degree of agency

10. Does the AI or human agent contribute more to the system's decision-making? Action-taking?
The system does not take actions or decisions beyond data collection and analysis. Question is a little unclear in this case

11. How much agency does the human have? The AI system? Full for the human (arguable, though)

II. NATURE OF SITUATION

Location and context

12. Are other people or other AI systems involved as third-parties? This can apply to either 1-1 collaborations or multiple people and an AI (e.g., in the instance of an AI teacher interacting with a classroom of human students, those human students would not be included as third-parties).
Yes. Patients, nurses who do not interpret the results

13. Are the human and AI agents co-located physically or virtually?
Seemingly co-located, though the processing may be physically done elsewhere/decentralized

Awareness

14. Is the human likely aware that they are interacting with an AI system? Yes (although I can't say for sure doctors think of the "new IT system" to be a version of artificial intelligence)

15. Does the human need to consent before interacting with the AI system? Unsure

Consequences

16. How significant are the consequences should the AI fail to perform as designed/expected? What are those consequences? Potentially high if false-negative or false-positive

17. How significant are the benefits of the AI to the users should it perform as designed/expected? What are those benefits? High

18. What are the potential consequences and benefits of the outcome of the collaboration?
Less time spent (patients and doctors). More capacity and resources available in medical facility. Potentially greater accuracy

19. What might be the broader impacts of the human-AI collaboration? More capacity and resources available in medical facility. Patients spending less time at medical facility

20. To what extent do typical users consider privacy and security when interacting with the AI agent?
Likely high (bodily privacy, information privacy, security of data, safety to be subjected to system), would require interviews to be conducted in order to know for sure

.

Assessment

21. Who is the main party or individual assessing the nature and effectiveness of the human-AI collaboration? Researchers & Doctors

22. Are assessments of the human-AI collaboration's outcome subjective or objective?
Depends on how they are analyzed by the research team, I would hope objective

Level of trust

23. Are both the human and the AI agent trusting and trustworthy?
AI trustworthiness can be defined broadly, driven by task competence, safety, authority, and authenticity, amongst other features (e.g., we know an AI comes from the same affiliation it claims to be from). Yes, both should be both trusting and trustworthy.

III. AI SYSTEM CHARACTERISTICS

Interactivity

24. What is the mode of the interaction between the two agents? Screen (likely)

25. Could the nature of the data that the AI system operates over impact its interactivity?
E.g., a wearable AI may interact with the user via voice, but may make inferences over voice as well as sensor data. Probably, though I'm not sure I can articulate how

Adaptability

26. Is the AI system passively providing information or proactively anticipating the next steps of the interaction? To some extent the analyses may guide the doctors interpretation and therefore their decisions

Performance

27. How predictable is the AI system? Probably low, otherwise there would be no need to do MRI scans

28. Does the system often produce false-positives? False-negatives? Both would be hugely problematic if so, so likely not. However, they likely do so at times

Explainability

29. Can the AI system communicate its confidence levels to a human?
Unsure, as I am not a doctor using the system, but it should

30. How does the AI system communicate its decision-making process and inputs to that decision-making process to the human? Unsure, but it should do so clearly to provide doctors with context

Personification

31. How human-like is the AI system? Not very

32. How easily anthropomorphized is the AI system? Not easily

IV. HUMAN CHARACTERISTICS

Age

33. Is the person(s) collaborating with the AI system a child (under 18), an adult (18 - 65), or a senior (over 65)? Some mixture of the above, when >1 person collaborating with the AI system?
Doctor (in training or professional) so likely between 18 and 65

Differently-abled

34. Does the person collaborating with the AI have special needs or accommodations?
Not necessarily.

Culture

35. Are there cultural consistencies/norms for those collaborating with the AI system?
These could be studied (perhaps by PAI)

36. What level of previous technology interaction has the user(s) of the system had?
Not necessarily high (likely)

CASE 7: Autonomous Vehicles and Passengers

In Kwon Choi, Staff Engineer (Samsung)

Context/Scenario

Autonomous vehicles, specifically self-driving cars, promise benefits such as increased safety and mobility, as well as potential reduction in cost, and improved user experience for those using them. Major global automobile manufacturers and software companies are engaged in fierce competition for market share in a future defined by ubiquitous autonomous vehicles. While there are many possible interaction scenarios in an autonomous vehicle beyond the core function of driving, or even beyond cars, we choose to restrict this case example to the interaction between the passenger (who could serve as the driver, depending on the level of autonomy) and the car.

AI System

Innovations in artificial intelligence have contributed to progress towards increasingly safe autonomous vehicles. For example, MIT's MapLite leverages AI models, GPS data, and road condition sensors to expand a vehicle's capacity to navigate unknown environments. The MapLite system enables self-driving cars to drive on unfamiliar roads autonomously and safely, and without the aid of 3-D maps, which are often required for successful navigation.

Human-AI Collaboration

The discussion on autonomous vehicles and their regulation is anchored in frameworks that contrast the levels of autonomy such vehicles possess against degrees of human control and potential for intervention. Five levels for automated driving systems, from zero (complete human driver control) to five, with full autonomy, have been identified by the Society of Automotive Engineers (SAE) and adopted by the U.S. National Highway Traffic Safety Administration. Full human driving is at Level Zero, while at Level One the vehicle assists with some functions like automatic braking, or lane drift prevention, while the driver still handles all major operating tasks. As levels of autonomy escalate, the driver is able to disengage from more and more tasks. By Level Four, vehicles are designed to perform all safety-critical driving functions and monitor roadway conditions for an entire trip. However, this does not cover every driving scenario, and in order to reach Level Five autonomy, no human attention or intervention is required at all.

Despite many innovations in AI that contribute to the safety and the autonomy of vehicles, reports suggest that society – in other words, the very people that would choose to use autonomous vehicles – is largely distrusting of self-driving cars. For instance, a 2018 report from the American Automobile Association revealed that an increasing number of Americans, ~73% of those surveyed, distrust autonomous vehicles. At the same time, the AI systems that assist autonomous vehicles may not yet be at a level of development that ensures human safety, and the question of threshold levels of safety is still not fully answered, as evidenced by several driver fatalities from Tesla's Level Two autopilot system and one pedestrian fatality from Uber's Level Three self-driving vehicles between 2016 and 2018. The implications of safe AI and robust collaboration between human and AI systems is a crucial issue of utmost importance to the safety, use, and eventual adoption of autonomous vehicles. Assessment of the interaction between humans – both passenger and pedestrian – and autonomous vehicles at different ratings along the SAE scale must be undertaken to emphasize and encourage safe, robust, and widespread use of autonomous vehicles.

For the purposes of mapping autonomous vehicles to the framework, we choose to consider how the interaction is handled now, during the ongoing transition to autonomous vehicles. Certain concerns related to consequences, such as a user's willingness to trust the vehicle, present unique challenges at this inflection point in autonomous vehicles' use, and will be vital to their successful and beneficial integration into society.

CPAI HUMAN-AI COLLABORATION FRAMEWORK
CASE 7: Autonomous Vehicles and Passengers

I. NATURE OF COLLABORATION

Stage of development or deployment

1. Is the AI fixed once deployed or evolving over time via model updates/continual interaction?
Evolving over time via model updates/continual interaction.

2. To what extent is there ongoing collaboration between the AI's developer(s) and the AI system? [No collaboration, limited collaboration, moderate collaboration, active collaboration] Active collaboration

3. Is the AI system currently used by people other than the AI's original developers? Yes

Goals

4. Are the goals of the human-AI collaboration clear or unclear? Clear

5. What is the nature of the collaboration's goals? [Physical, knowledge/intellectual, emotional, and/or motivational in nature] Physical

6. Is empathy a precondition for the human-AI interaction to function as intended? No

7. Are the human and the AI system's goals aligned? Yes, if safety is ensured. Of course, the guarantee of safety is a large contingency

Interaction pattern

8. Is the collaboration repeated over time or is it a one-time engagement? If over time, at what time-scale? Over time, likely on a daily basis

9. Is the interaction concurrent – with both human and AI contributing in parallel – or does it depend on taking turns? Concurrent

Degree of agency

10. Does the AI or human agent contribute more to the system's decision-making? Or action-taking? AI takes the lead, but humans can intervene in decision-making and action-taking

11. How much agency does the human have? The AI system? [None, limited, moderate, high, full]
Human: moderate; AI: high

II. NATURE OF SITUATION

Location and context

12. Are other people or other AI systems involved as third-parties? This can apply to either 1-1 collaborations or multiple people and an AI (e.g., in the instance of an AI teacher interacting with a classroom of human students, those human students would not be included as third-parties).
Yes. Possibly other pedestrians/passengers

13. Are the human and AI agents co-located physically or virtually? Co-located physically

Awareness

14. Is the human likely aware that they are interacting with an AI system? Yes

15. Does the human need to consent before interacting with the AI system? Yes

Consequences

16. How significant are the consequences should the AI fail to perform as designed/expected? [Low, moderate, high] High

17. How significant are the benefits of the AI to the users should it perform as designed/ expected? [Low, moderate, high] High

18. What are the potential consequences and benefits of the outcome of the collaboration?
Accident/injury [either potential consequence or lessening a potential benefit]

19.What might be the broader impacts of the human-AI collaboration?
Climate impacts. Altered city design. Safety

20. To what extent do typical users consider privacy and security when interacting with the AI agent? [Low, Moderate, High] Low

Assessment

21. Who is the main party or individual assessing the nature and effectiveness of the human-AI collaboration? Autonomous vehicle companies, transit agencies

22. Are assessments of the human-AI collaboration's outcome subjective or objective?
Objective in the case of safety standards and ubiquity of adoption

Level of trust

23. Are both the human and the AI agent trusting and trustworthy? AI trustworthiness can be defined broadly, driven by task competence, safety, authority, and authenticity, amongst other features (e.g., we know an AI comes from the same affiliation it claims to be from). AI agent in the form of an AV must be trusted for it to be used

III. AI SYSTEM CHARACTERISTICS

Interactivity

24. What is the mode of the interaction between the two agents? [Via screen, voice, wearables, virtual reality, or something else] Physical vehicle.

25. Could the nature of the data that the AI system operates over impact its interactivity? E.g., a wearable AI may interact with the user via voice, but may make inferences over voice as well as sensor data. Yes

Adaptability

26. Is the AI system passively providing information or proactively anticipating the next steps of the interaction? Proactively anticipating the next steps of the interaction

Performance

27. How predictable is the AI system? [Low, moderate, high] Moderate

28. Does the system often produce false-positives? False-negatives? No

Explainability

29. Can the AI system communicate its confidence levels to a human? Yes

30. How does the AI system communicate its decision-making process and inputs to that decision-making process to the human? By direct control of the self driving vehicles, with the action as a result of its decision-making. And/or, via audio/visual/natural-language communication to the human to inform or confirm its decision-making process and/or results to be carried out.

Personification

31. How human-like is the AI system? [Not very, moderately, or highly human-like]
Not very in terms of presence, but moderately in terms of reasoning about the road

32. How easily anthropomorphized is the AI system? [Low, moderate, highly] Low

IV. HUMAN CHARACTERISTICS

Age

33. Is the person(s) collaborating with the AI system a child (under 18), an adult (18 - 65), or a senior (over 65)? Some mixture of the above, when >1 person collaborating with the AI system?
Could be all ages for passengers, but driver likely >16.

Differently-abled

34. Does the person collaborating with the AI have special needs or accommodations? Potentially

Culture

35. Are there cultural consistencies/norms for those collaborating with the AI system?
Driving norms vary around the world [adherence to traffic rules, aggression, etc.]. Accordingly, the same system that performs well in one context may not in another.

36. What level of previous technology interaction has the user(s) of the system had? [Low, moderate, high] Will vary across users/passengers

Acknowledgements

PAI is deeply grateful for the time, insights and expertise provided by the members of the CPAIS Expert Group, and in particular the co-chairs, Amber Story of the American Psychological Association and Hiroaki Kitano of Sony, and the contributors to these Case Studies - Yoonsuck Choe of Samsung, Lama Nachman of Intel, Philipp Michel of the University of Tokyo, Bendert Zevenbergen of the Princeton Center for Information Technology Policy and In Kwon Choi of Samsung.

About Partnership on AI

The Partnership on AI (PAI) is a global multistakeholder nonprofit committed to the creation and dissemination of best practices in artificial intelligence through the diversity of its Partners. By gathering the leading companies, organizations, and people differently affected by artificial intelligence, PAI establishes a common ground between entities which otherwise may not have cause to work together – and in so doing – serves as a uniting force for good in the AI ecosystem. Today, PAI convenes more than 90 partner organizations from around the world to realize the promise of artificial intelligence. Find more information about PAI at partnershiponai.org.

PARTNERSHIP ON AI