Milestones

5 years: AI systems could be designed to study psychological models of complex intelligent phenomena that are based on combinations of symbolic processing and artificial neural networks.

10 years: Integrated architectures are the standard vehicle for modeling the results of complex psychological experiments.

15 years: Progress in AI on neural networks and integrated architectures yields major advances in neural/brain modeling.

Towards Unifying Theories of Natural and Artificial Intelligence

Ultimately, a comprehensive theory of integrated intelligence is desirable that spans all possible forms of intelligence, whether natural or artificial. Such a grand challenge will likely take much more than 20 years to complete, but a start can be made within this time frame by focusing on more modest goals. Recently, a new community has begun to coalesce around the goal of designing human-like cognitive architectures that model both human and artificial intelligence. Among other things, such a common architectural framework could provide a useful intermediary in evaluating and comparing cognitive architectures and create a pathway for unifying models of natural and artificial intelligence.

An essential aspect of those architectures that has increasingly guided their development is the ability to validate them using neural imaging data. Neural imaging data, assembled in large databases, such as the Human Connectome Project covering substantial subject populations performing a diverse range of tasks and using a number of imaging techniques, provides constraints regarding both the structural and functional organization of brain modules as well as the details of knowledge representation within those modules. This new source of data expands on the wealth of existing behavioral data accumulated over more than a century of research, provides converging evidence for and against proposed architectures, and holds the promise to considerably speed up convergence to a consensus theory of natural and artificial intelligence.

Stretch goals: By 2040, a common cognitive architectural model will yield deep understanding across at least one full arc of cognition, from perception to behavior, for a complex task in a real environment. Milestones along this path include—

Milestones

5 years: The full space of existing cognitive architectures (i.e., integrated models of human-like intelligence) is mapped onto a single common model of cognition.

10 years: Strong connections are demonstrated between AI architectures and cognitive models that can be mapped at the level of major brain regions, their functional connectivity and mechanisms, and their communication patterns.

15 years: Shared implemented models of cognition are in wide use by both the AI and computational cognitive science communities.

3.2 A Research Roadmap for Meaningful Interaction

3.2.1 INTRODUCTION AND OVERVIEW

Research on AI for Interaction has resulted in significant advances over the last 20 years. In fact, many of these advances have seen their way into commercial products. In the four focus areas of AI for Interaction, we have seen successes but also major limitations:

- AI systems that act as personal assistants have seen wide-scale success; These systems can interact using spoken language for short, single turn commands and questions, but they are not able to carry on intentional dialog that builds on context over time.

- AI systems can use information from both speech and video when processing streams such as broadcast news to build an interpretation, but they are not able to fuse information from other sources such as smart home sensors.
- AI systems are already able to detect factual claims in text and to distinguish fact from opinion but they cannot reliably determine when claims are false.
- AI systems have seen widespread deployment and trust for a number of practical applications where the system is very accurate or the stakes are very low, as in the case of music recommender systems, but they are not yet trusted in high-stake domains such as medical diagnosis.

Many future AI systems will work closely with human beings, supporting their work and automating routine tasks. In these deployments, the ability to interact naturally with people will be of unparalleled importance. Developing AI systems that have this capacity for seamless interactivity will require major research efforts, spanning 20 years, in four interrelated areas:

1. Integrating Diverse Interaction Channels: AI systems can integrate information from images and their text captions, or from video and speech when processing related streams such as broadcast news. Combining different channels of interaction, such as speech, hand gestures, and facial expressions, provides a natural, effective, and engaging way for people to communicate with AI systems. Jointly considering multiple input modalities also increases the robustness and accuracy of AI systems, providing unique opportunities that cannot be accomplished by considering single modalities. Today's AI systems perform well with one or two modalities. We have seen tremendous advances in speech-based commercial products such as home personal assistants and smartphones. As we move into the future, however, AI systems will need to integrate information from many other modalities, including the many sensors that are embedded in our everyday environments, thus leveraging the advances that have taken place in the Internet of Things. AI systems must also be able to adapt their interactions to the diversity of human ability, enabling their use by people with disabilities, by people with non-standard dialects and by non-English speakers. AI systems must be able to take context into account and handle situations when information is missing from a modality. Multiple modalities also pose a particular challenge for preserving privacy, as much sensitive information is revealed during interactions, including images of faces, recordings of voices, and views of the environment in which people live. AI systems also need to be capable of integrating new, more efficient communication channels, as well as directing and training users in how to best interact with the systems on these new channels.

2. Enabling Collaborative Interaction: Interactions with today's AI systems are often stilted and awkward, lacking the elegance and ease of typical interactions between people. For example, today's AI systems rarely understand the context underlying the user's desires and intentions, the potentially diverging objectives of different users, nor the emotional states and reactions of humans with whom they are interacting. In contrast, when humans interact, they intuitively grasp why someone is asking a question, they seek to find common ground when faced with differing opinions within a team, and they reciprocate in their emotional expressions, offering understanding and encouragement when their teammate seems despondent. Building AI systems that interact with humans as fluently as people work together will require solving challenges ranging from explaining their reasoning and modeling humans' mental states to understanding social norms and supporting complex teamwork.

3. Supporting Interactions Between People: With the increased presence of social media and other online discussion venues, increasingly more human social interactions take place online. AI can help in facilitating these interactions and in mining online data for understanding the emerging social behaviors of online communication. Understanding what people expect from online media, and how they want to use it, can help in developing AI systems that address those needs. We categorize future research in this area into three subsections: deliberation, collaborative creation, and social-tie formation. In the area of deliberation, AI systems are needed that can distinguish fact from falsehood and fact from opinion. AI systems that can detect bots will also be critical moving forward. Online deliberation data can help researchers understand how opinions form and how influence spreads. In the area of collaborative online content creation, Wikipedia already uses AI bots to detect vandalism and enforce standards. Future research should extend this work to new settings such as creative arts or collaborative software engineering. Here

there is a role for AI systems to support collaboration: identifying inconsistencies, assumptions, and terminological differences among collaborators. Finally, social-media analysis is an active area of research, with commercial efforts focusing on identifying phenomena such as hate speech, albeit with the human in the loop. We foresee increasingly sophisticated human-machine hybrid technologies in this space, which may ultimately enable new forms of human-human interaction, and even support emergent social structures.

4. Making AI Systems Trustworthy: Today's AI systems often use methods that offer high performance and accuracy, but are not able to appropriately describe or justify their behaviors when asked. Generating appropriate descriptions of their capabilities and explanations of the resulting behaviors will require research to develop new approaches that align the machine's internal representations with human-understandable explanations. Based on their understanding of an AI system, users should be able to assess and improve any undesirable behaviors when appropriate. AI systems should also have provisions to avoid undesirable persuasion and manipulation. Finally, appropriate mechanisms need to be developed to enable AI systems to act responsibly and ethically, to convey to users the acceptance of responsibility, and to engender and uphold people's trust.

The report's next section contains motivating vignettes for each of the societal drivers, highlighting the research required in each of these four areas. We then discuss the challenge areas in more detail, posing both stretch goals for the 2040 time frame and milestones to track progress along the way.

3.2.2 SOCIETAL DRIVERS FOR MEANINGFUL INTERACTION WITH AI SYSTEMS

The vignettes below illustrate the impacts across human society that the proposed research could enable by the 2040 time frame.

ENHANCE HEALTH AND QUALITY OF LIFE

Vignette 9

Andrei is a precocious 4-year-old who has had difficulty making friends in the neighborhood and complains of interrupted sleep. Their home AI system, embodied as a robot called Nishi and responsible for managing many household duties, has performed routine screenings for ADHD, ASD, and dyslexia based on Andrei's behavior, play patterns, and sleeping patterns (inferred from a wearable activity tracker and in-home monitors). Since Andrei's behavior included traits of autism spectrum disorder, his parents authorized remote screening by a specialist, Dr. Marie, which led to an in-person visit. During this exam, an automated note-taking system records their conversation and highlights specific phrases that indicate medically relevant information and changes from previous checkups. It also notified Dr. Marie of a potential drug interaction correlated with poor sleep, reported in a research study published the previous month.

Andrei now receives an hour of personalized social skills training each day from Nishi, using prescribed therapy games to help Andrei and his parents relate to each other. Nishi also summarizes Andrei's progress to his parents and the doctors, providing personalized views to each physician, highlighting the medical details each one might deem important. The physicians confer to configure personalized updates for the therapy robot.

Research Challenges. Realizing this vision of improved health care will require a number of breakthroughs in the four challenge areas:

1. Integrating Diverse Interaction Channels: Whether conversing with the elderly, a child with emotional challenges or a patient from any walk of life, AI systems must be able to handle a wide diversity of human speech. As we develop systems that use vision, speech, and gesture, we will need a new research ecosystem that makes it possible to create and collect datasets appropriate for these different healthcare settings and that represent all modalities, as well as experimental platforms that enable testing new approaches that were computationally expensive in just one modality. To infer qualities of social interaction and to monitor behavioral health problems such as sleep, interactive systems need to develop new modalities (e.g., facial expressions) and fuse together modalities that have previously not been explored together (e.g., vision, sleep monitors, and speech). With the advent of systems that interact with the elderly to catch cognitive problems and that evaluate a child's ability to socialize, we will need to develop new methods to safeguard their privacy as well as the privacy of those around them.

2. Enabling Collaborative Interaction: To interact with a child like Andrei in a way that is understanding and caring, research in AI must enable systems to carry on a more natural interaction than currently possible and must be better capable of recognizing human emotions and motivations. AI systems must safeguard individuals who are not able to adequately care for themselves, acting reliably and ethically when carrying out critical tasks. To enable appropriate social experiences for Andrei, the AI system must be able to quantify and evaluate Andrei's experiences and progress.

3. Supporting Interactions Between People: One of the robot's goals is to connect Andrei with his neighborhood community, building stronger social ties. Nishi watches in the background as Andrei plays with his friend, but monitors their interaction and may suggest a new activity when discord occurs. The AI system also facilitates collaboration between Andrei's pediatrician and an ADHD specialist, providing different summaries of Andrei's health records corresponding to the key aspects that each specialist requires. Similarly, the AI system tracks recommendations given by different physicians to ensure consistent care and to follow-up on treatment plans.

4. Making AI Systems Trustworthy: To fully take advantage of what interactive AI systems can offer, patients must come to trust the AI systems assisting them. This can only happen if AI systems can explain their actions and suggestions, and understand how to seek the trust of their users.

ACCELERATE SCIENTIFIC DISCOVERY

Vignette 10

Charlie and her distributed team of top-notch materials scientists have a goal of identifying a material with certain properties, obeying various constraints that will render the material practical and cost-efficient.

Charlie manages the project using an AI system that is connected to robotic lab equipment. The AI system makes it easy for Charlie to articulate research goals, hypotheses, and experiment designs. Most experiments are performed automatically with results recorded for full reproducibility. In addition, the AI system tracks inventory and orders replacement materials, tracking the resulting orders. The AI system can mine the literature on materials, gathering information about past experiments and reporting the results in a way that is consistent with Charlie's existing database of past experiments and results. Charlie directs the AI system to find possible patterns in the data and works with it to conjecture causal relationships between factors and outcomes, rather than simply correlative

Vignette 10 Continued

relationships. The AI system also performs a type of optimal experimental design to offer options to the team, carries out the next experiment, given constraints and directives from the scientists. The AI system estimates both an experiment's cost and its possible benefit in the context of the latest published results, providing clear reasons for its uncertainty estimates and accepting redirection and priorities from Charlie's team.

When discussing their project, Charlie's team uses a mixture of online platforms, from text messages and video to immersive augmented and virtual reality. These platforms are moderated by another AI system that supports discussions. This AI system is designed to be an active participant in the conversation that helps synthesize their different contributions into a consensus on the team's goal. The modeled goal is constantly updated and provided in an interpretable way that is consistent with the scientific language of the community. Charlie notices that an important constraint regarding the desired properties of the material is missing from this goal model and she directly updates it. Thanks to dynamic visualizations and an intuitive control interface, her teammate notices that the AI system is using a weak causal relationship and suggests to the system to ignore it. Together the team decides on the next course of action and the resulting experiment proves successful. The IMS drafts many sections of the paper submission, including experimental methods and related work, and the final paper is a landmark.

Research Challenges. Achieving this vision requires a number of breakthroughs in the four challenge areas:

1. Integrating Diverse Interaction Channels: The AI system renders 3D models of the proposed materials, sometimes printing them out and testing their material properties directly and comparing the properties to those predicted by the simulation. The AI system also facilitates virtual meetings, so that the scientists can interact fluently despite being in different locations. Effective use of AR and VR for visualization and collaboration requires coordination of voice, gestural, and other inputs, along with inferring and acting on user intentionality. The AI system must understand natural input and convert user actions between different media (flat screens and video or immersive 3D AR and VR).

2. Enabling Collaborative Interaction: The AI system must be able to use knowledge in this role. In synthesizing contributions, identifying related work, and dividing work among the team, the system must take context into account, communicate naturally, model the team members' mental states, learn to adapt to expectations, and support complex teamwork. In its role of helping the team, the AI system must be able to explain its decisions and to gain the trust of team members.

3. Supporting Interactions Between People: As an active participant in the team members' conversations, the AI system helps build consensus from possibly diverging observations shared by all team members, and helps design metrics based on its knowledge of the most recent metrics developed by that scientific community. The AI system is tasked with using interactive structures to produce documents (e.g., a task list or a consensus on team goals).

4. Making AI Systems Trustworthy: The AI system must be able to explain its rationale to the team members (e.g., why it suggested certain experiments) and it must make its level of uncertainty clear in a way that team members can truly understand. Given the critical outcomes of experimentation, the system must act reliably. Science requires a high standard of ethics and must adhere to IRB regulations and thus, the system must be able to address ethical concerns.

LIFELONG UNIVERSAL ACCESS TO COMPELLING EDUCATION AND TRAINING

Vignette 11

Joe is a worker who was laid off in a company restructuring. He wants to retrain, but needs income in order to support his family and cannot afford to embark on full-time education. A free AI system helps him plan for career change—what is a feasible job he could take that would either build the skills he needs along the way or would pay the bills while giving flexibility to study and advance his career?

To explore his short- and long-term career opportunities, Joe navigates to an interactive AI system and describes his skills and interests. The system visualizes a number of possible career paths for him, including both short- and long-term steps he can pursue to make progress on those paths. The system is sensitive to his individual experiences and goals and builds models of what he wants from a new job. For each path, the system determines missing skills and good first steps to take in achieving that career. In some paths, Joe may enroll in online classes where human instructors supplemented with AI assistants that can provide personal tutoring.

Where a human teacher leads the instruction, the AI system tracks student engagement and attention; prompts the teacher to engage with each student as necessary; assists the instructor in giving feedback to students, even on open-ended work; gives feedback to the teacher on each student's progress, mastery, and challenges; and intervenes with students to increase student retention in the classes. It observes communication between students and the teacher to understand when they agree and disagree. Students are matched together efficiently to collaborate on projects according to their interests, skills, and collaboration style.

Where students must practice material on their own, Joe may use virtual reality to develop some skills. Joe uses an online simulator to practice interacting with simulated customers, and AI helps provide explainable feedback on how he is doing and how he can improve.

The interactive system helps match Joe to open jobs that fit his skills and interests, matching constraints he may have. Joe works with the AI system to identify additional challenges of his new job where he could use further training. The AI system allows Joe to easily and continuously learn and practice a variety of new skills, even including those that require teamwork with other employees.

Research Challenges. Achieving this vision requires a number of breakthroughs in the four challenge areas:

1. Integrating Diverse Interaction Channels: A diversity of students may use the AI system, and their speech must be understood in order to identify their learning goals and evaluate their progress. Extensive datasets that include video, speech, gestures, and other modalities are required for developing AI systems that give on-the-job guidance in real-world situations and that accurately allow students to practice what they learn using virtual reality or augmented reality environments. User models are necessary for understanding and evaluating the interactions students have during their training and providing feedback. Those user models may involve students' movements, vision, speech, facial expressions, etc.

2. Enabling Collaborative Interaction: Common sense is required for the system to recommend career paths (e.g., Joe cannot simply become a manager in a field where he has no experience; he must take an entry-level job in that field) and for on-the-job training in which students must interact with the world or with others. An understanding of context is required for the system to recommend career paths (in this case, students' background and prior interactions with the system), and for understanding students' social and emotional backgrounds when giving teachers feedback on their mastery and challenges. Natural interaction is required when the system interacts with the user to discuss career paths and the users' constraints, and during training when students have questions or must interact with others. Modeling students' mental states is required to model their learning: the system must know what students figure out and do not figure out. Modeling students' emotional states is required to understand their engagement and progress in their classes and training, in career change settings, and when the system is attempting to predict when students may withdraw (so that it can monitor and intervene to increase retention). These models may need to be built with limited data about the students. Personalization and adapting to expectations is required to enable personalized content generation and curriculum generation (in order to select the best materials and best way to convey them for each student). An understanding of human experience is required for the AI systems described to understand users' desires in career paths and their engagement with or enjoyment of training or studying. If multiple students are involved (e.g., in a group project, or a team job), the AI system may need to facilitate complex teamwork among them.

3. Supporting Interactions Between People: AI systems that facilitate constructive online collaboration may be required when multiple students interact to complete work (e.g., a group project). Gathering and linking social media data to real-world experiences may be required when giving users advice about potential career paths or developing practice problems.

4. Making AI Systems Trustworthy: The AI systems described must be trustworthy and explainable so that students learn effectively, in classes or on the job: If Joe is practicing patient interaction, the simulator must be able to explain why his bedside manner is or is not ideal. The feedback that teachers are given on students' engagement and progress must also be explainable for effective teaching. Since the AI systems described may be helping people make personally important decisions about their life paths, or guiding them through critical jobs, ethically designed systems that protect against and do not perform malicious manipulation are necessary.

AMPLIFY BUSINESS INNOVATION AND COMPETITIVENESS

Vignette 12

Hollis runs a small online business, where she sells customized personal devices and customized robots, which she designs and builds on demand. Some objects are aesthetic, such as integrating light and motion sensors with embedded LED lighting to add responsiveness to jewelry; others are more functional, such as customized wristbands that integrate her designs with medical sensors and small displays.

An interactive AI systems allows Hollis to rapidly develop specialized products for her customers, enabling new business opportunities. Hollis can focus on her creative skills, while the AI system monitors her storefront and manages the manufacturing and logistics of her business. Her online storefront shows existing designs that can be customized with different electronics: The AI system ensures requested customization will work (taking into account size, layout, power and other constraints), and notifies Hollis only when a customer requests something that requires adjusting the base designs. The AI system uses smart analytics to inform Hollis about product trends by integrating data from her storefront with opinion data mined from relevant blogs as well as knowledge about what sorts of

Vignette 12 Continued

electronics are popular and available from the suppliers in her supply chain. Hollis uses this information to explore new designs, using another AI service to select appropriate focus groups and collect feedback.

Hollis creates the company's production designs in 3D using direct manipulation in an augmented reality display. The AI system enforces practical constraints in real time and can be taught new guidelines and skills during operation using a combination of demonstration (e.g., showing the system how to do it) and verbal commands (e.g., "Repeat those actions on all sides of the object.")

Hollis's products are made on demand by a set of manufacturers around the world, each with different materials and electronics chosen through automated negotiations that suggest the right supplier for the part. The AI system considers price and build times, but also checks the reputations of the manufacturer and of the equipment they use. The AI system can also be directed to summarize other factors, such as expected shipping times at different times of year, social issues such as political upheaval and working conditions, and so on. The AI system presents its final suggestions to Hollis using an automatically synthesized graphical rendition that highlights the pros and cons of each choice. After asking clarification questions, Hollis makes the final decisions with confidence, knowing that the AI system continuously monitors pricing and availability conditions, alerting Hollis only when there are significant changes to consider. Occasionally, Hollis needs to communicate directly with a supplier; in these situations, the AI system translates between languages during real-time videoconferencing, and advises Hollis during negotiation and conflict resolution.

Research Challenges. Achieving this vision requires a number of breakthroughs in the four challenge areas:

1. Integrating Diverse Interaction Channels: Hollis's AI system finds video blogs depicting users interacting with her products, extracts facial expressions and analyzes acoustic emotional cues to deduce user sentiment. When Hollis uses the design program, the AI system relies on several sensors to augment her creative experience, including haptic feedback and touching sensors. The AI system provides the ability to migrate from augmented to virtual reality, allowing Hollis to see, touch, feel, and wear new prototypes or alternative designs, which the AI system can create in short time. The AI system also supports natural speech interaction between Hollis and suppliers around the world, breaking language barriers. The system handles accent variations from specific locations and analyzes nonverbal behaviors for their meaning when interacting with people from other cultures, avoiding misunderstandings.

2. Enabling Collaborative Interaction: The AI system should have an understanding of the sort of work Hollis is engaged in at any moment (based on watching and learning, but also leveraging her calendar, and so forth), so that it interrupts her at appropriate times. Different people have different preferences (e.g., when Hollis considers different suppliers), which need to be learned by AI systems. AI systems must also support natural cross-cultural communication and interaction (e.g., when responding to customers and negotiating with suppliers).

3. Supporting Interactions Between People: In order to recommend suppliers, the AI system must estimate reliability from a variety of online traces, including noisy or fraudulent reviews. In order to help predict trends and preferences for Hollis's products, the AI system must link social media signals to real-world processes.

4. Making AI Systems Trustworthy: In order for Hollis to trust the AI system, it must be able to explain its reasoning, answer questions about which information it is taking into account and quantify its uncertainty. The AI system also must understand that it must treat customers and suppliers without misleading them and use customer data ethically. All these skills require significant progress over current methods.

SOCIAL JUSTICE AND POLICY

Vignette 13

Susan comes home to an eviction notice on her door, which gives her 24 hours to remove all of her possessions from her apartment. She has had problems with her landlord and feels that the eviction is unjustified and possibly illegal. But her immediate concern is to look for another place to live, so she asks her AI assistant to search for short-term housing in her city. The AI system returns a set of results, but in a side channel (visual or auditory) asks if something is wrong with her current apartment, since it knows that her lease runs for another eight months. She tells the system that she was evicted.

One possible path for the AI system is to provide legal support. The system queries relevant tenant rights law for her city and state. The result is a set of legal documents, which are long and difficult for non-experts to read. The AI system therefore identifies critical parts of each document, specifically relating to eviction. Susan can highlight parts that she doesn't understand, and ask for clarification; this can be done either by generating explanations of an appropriate level of technicality and dialect for Susan, or it can identify snippets of relevant documents. The AI system offers to connect with her with a set of legal aid resources in the morning. The AI system also provides Susan with a set of just-in-time social connections to individuals who have been in a similar situation and are willing to share their experiences. This group can include individuals who have fought an eviction (successfully or otherwise), individuals who have used short-term housing in the same area, others who have used free legal aid services, and other current and former tenants of the same landlord. Keeping in mind that these individuals may not have shared these experiences widely, the AI system needs to decide when and under what conditions to share what information to support Susan's interaction with the others. In addition to person-to-person connections, the AI system can aggregate social media content from these individuals related to their experiences, which will create a set of resources to help Susan deal with her situation both practically and emotionally.

Research Challenges. Achieving this vision requires breakthroughs in all four technical areas.

1. Integrating Diverse Interaction Channels: Susan might interact with the system partially via speech (necessitating accurate speech recognition for a diverse range of speakers and across many contexts), and the system may make additional inferences about the environment and context via video input. Conversely, it can offer information to Susan across multiple modalities, including visual and audio display when appropriate in context. Multimodal integration will support more advanced capabilities, such as a combination of speech and gesture to identify salient features of the environment, such as lack of upkeep of the apartment.

2. Enabling Collaborative Interaction: The AI system works collaboratively with Susan to quickly identify solutions to her situation. There are important contextual factors, and the system must recognize the priority of this situation with respect to other plans and goals that may have been previously indicated by Susan. There is also an affective component to the interaction: Eviction is likely to provoke stress, which may affect the user's decisions as well as her interactions with the system. The system should account for this and adapt its behavior accordingly.

3. People Interacting Online: Several of the system's goals require capabilities relating to online interaction. The AI system must identify factual claims about the legal situation and verify them against external sources. This requires not only finding trusted resources that support or reject existing factual claims, but ensuring that those resources are germane to the specific situation at hand: for example, that the laws and regulations pertain to the correct jurisdiction and type of housing. The system should also be able to distinguish legal consensus from theories or strategies that only some community members believe will work. In proposing just-in-time social connections, the system should optimize for constructive and civil discourse. The scenario also pertains to collaborative creation: The system may help non-experts to assemble a set of guidelines for dealing with eviction, while soliciting targeted advice. The system could identify and help to bridge linguistic gaps between experts and non-experts, including multilingual support.

4. Making AI Systems Trustworthy: Like a trusted human advisor, we expect the system to act entirely on behalf of the user, without the possibility of external manipulation. This includes blocking the acquisition of data about the situation that might impact the user negatively. In this case, because eviction has potential long-term social consequences, the system should not leak information about the situation without explicit user approval. The scenario is relatively high-stakes and time-sensitive, so the AI system's advice should be appropriately scoped and have appropriate justifications. In particular, the advice should be grounded in original sources (laws and regulations), and the connection between these sources and the advice should be made clear. This may necessitate reasoning about linguistic transformations that preserve meaning while helping readers with variable levels of legal experience, formal education, and reading ability. Any risks should be accurately explained, and system uncertainty must be communicated to the user.

3.2.3 TECHNICAL CHALLENGES FOR MEANINGFUL INTERACTION WITH AI SYSTEMS

We group the technical challenges into four high-level capabilities: integrating diverse Interaction channels, enabling collaborative interaction, supporting better interactions between people, and making AI systems trustworthy. While people routinely use two or more modalities (e.g., voice, gestures and facial expressions) when communicating with each other, today's AI systems cannot do this in a robust and flexible manner. Similarly, today's AI systems aren't collaborative; They poorly model the context underlying a human's user's desires and intentions, the human's emotional state, and the potentially diverging objectives of different users. Increasingly, humans are using technology to mediate interactions with other humans, whether through social media or telepresence; There are many opportunities for AI systems to facilitate these interactions if we can solve key technical challenges. Today's AI systems perform well in low-stakes domains but are often unreliable in high-stakes environments such as medical diagnosis or judicial support; to realize the full potential benefits of AI, we need to make these systems more trustworthy and easier to control. We elaborate these challenges below.

Integrating Diverse Interaction Channels

Multi-channel inputs and outputs are everywhere, where wearable sensors can be interconnected with other computing devices, leveraging the advances in the Internet of Things. Day-to-day environments are replete with multimodal sensors in everyday consumer items: For example, doorbells now have video cameras, in-home personal voice assistants contain microphones, and thermostats are equipped with proximity sensors. People wear suites of sensors, such as heart rate monitors on their smart watches and accelerometers on their smartphones, which provide continuous, high resolution, and real-time multimodal data. Information of all kinds is being presented multimodally, as well. For example, smartphone apps use both vibration and text to send messages to users. Online communication, even from traditionally text-based organizations such as news agencies, is increasingly in the form of video and audio in addition to written media. The use of multiple channels for inputs and outputs is increasing as data bandwidth becomes cheaper and sensors become more integrated into common products. People and companies are already using multimodal data to derive insights about sleep, physical activity, traffic, and other information.

In addition, multimodal interaction provides a natural, effective, and engaging way to communicate with AI systems. Jointly considering diverse interaction channels increases the robustness and accuracy of AI systems, providing unique opportunities that cannot be accomplished by considering single modalities. Information incorrectly inferred from one channel can be corrected by another modality, leveraging the complementary nature of multimodal data. For example, compared to a decade ago, voice-based systems have improved in the ability to recognize user input accurately. Improvement in these technologies has led to more mainstream use of some channels (e.g., voice) in systems designed for leisure and entertainment but also work and activities of daily living. The advances in automatic speech recognition and the success of speech-based commercial products that are now ubiquitous in our lives (e.g., watches, smartphones, home personal assistants) have created a paradigm shift in multi-channel interaction. These devices have several heterogeneous sensors that are promoting applications relying on multimodal interactions.

The advances in multimodal sensors has also lead to better models. Over the last years, we have seen emerging algorithmic development for multi-channel processing with machine learning architectures that create representations that can be shared between modalities. Current research efforts have often focused on merging a few modalities, with prominent combinations of audio plus video (e.g., audio-visual automatic speech recognition and audio-visual emotion recognition), text plus images (visual question answering), and text plus video (semantic labeling, image/video retrieval).

At the same time, multimodal technology may be inaccessible to many people. For example, voice activated devices do not work well for people with speech impediments, visual output cannot be easily parsed by a screen reader for the blind, and speech output is unusable by people who are deaf. The fusion of more than two channels is also a challenge, especially when the data is incomplete (e.g., missing facial features due to occlusions).

In the remainder of this section, we identify major challenges for interaction using multiple modalities.

Handling Diversity of Human Ability and Context

Despite progress, many input modalities do not account for diverse human abilities limiting access to some who desire to use these systems or may benefit from them. Voice-based systems still find it challenging to recognize input from users with non-standard speech (e.g., speech impediments, deaf speech, different dialects) limiting access for some users. The ability to handle languages other than English is also critical in today's world, even enabling the possibility of translating between different languages, each spoken by a different conversation participant. Furthermore, some modalities have become commonplace (visual, speech, audio), while others such as haptics, proprioception, and other sensory inputs are underexplored. Multimodal systems will also need to take into account the context in which the signals are created to reach correct conclusions. Context includes the particular environment, interaction partners, prior knowledge, motivation, preferences, and a host of other information that dictates how the data should be interpreted. In the future, single modalities must continue to improve their capabilities for handling diverse human abilities. In parallel, multimodal systems must also take into consideration ways to leverage context as well as other underexplored modalities as opportunities for increasing access and improving interactions.

Stretch goals: By 2040, natural communication with virtual and physically embodied AI systems will be achieved through multiple modalities in both task-specific and open domains. Milestones along this path include—

Milestones

5 years: Multimodal systems will combine multisensory data (e.g., speech and a perceived visual environment) to infer communicative goals and to capture shared experience in constrained environments, communicating shared perceptual experience in language (i.e., generating language descriptions based on first-person view and grounding language to objects using well-trained computer vision models). Human-machine dialog will explicitly model common ground and shared intentionality with respect to the dynamic physical environment in specific domains.

10 years: Multimodal systems will capture long-term past joint experiences, using persistent memory and constructed models of past interactions.

15 years: Multimodal communicative actions will be generated for achieving shared communication goals. This will include capabilities for easily extending and integrating additional communication channels and task-specific languages aimed at improving human-AI communication efficiency.

Explainable, Interpretable, and Data-Limited Multimodal Interfaces

Transparency in AI systems has always been a concern among some in the AI community, beginning with early work on explanation generation for expert systems. Due to the pervasiveness of AI systems in recent years, there has again been increased interest in developing explainable, transparent, and trustworthy AI systems that also respect user privacy needs. Often, multimodal systems are developed to collect as much data as possible about a person using different modalities to improve accuracy or to supplement the limitations of one or more modalities. Some systems may use visual input as an alternative modality to supplement the limitations of speech as an input. Therefore, a system could collect quite a bit of data about an end user through the use of different modalities. It may be difficult, however, for end users to fully understand what data is collected, how it is collected, or how it is being used to support their interactions with the AI system. Some systems that collect data about users attempt to notify users about the data collected about them. However, the choice to opt out of having certain data collected often prohibits use of the system (i.e., if the user does not allow the system to use the data it desires, the user cannot use the system). There is a need to identify methods for creating multimodal systems that allow users to easily understand the data being collected about them, and how that data is used. Systems must also provide enough flexibility to allow users to customize the manner and time in which data about them is collected as well as the kind and amount of data that is shared in a way that aligns with their preferences. AI systems must consider options for end users to customize their interactions to provide access without having to abandon the system completely.

Stretch goals: By 2040, customizable multimodal systems that allow end users to choose how much privacy they retain while allowing for different levels of access to the system. Milestones along this path include—

Milestones

5 years: Methods to explain what data about a person is collected and necessary for interaction.

10 years: Auditing algorithms to efficiently determine how channels of streaming data (or the lack thereof) affects AI prediction capabilities, while ensuring privacy.

15 years: Methods for ensuring privacy for systems that use multiple modalities as input, including speech, vision, and other sensor input.

Plug and Play Multimodal Sensor Fusion

Current solutions for multimodal interactions are not flexible, mainly because: 1) they are often restricted to a small number of interaction channels, 2) they cannot handle missing information, and 3) they assume that the target sensors are predefined, where new (possibly unforeseen) modalities cannot be easily integrated. Yet new input devices are coming: haptic interfaces, brain implants, and more. An important technical problem is increasing the flexibility of multi-channel interaction systems to handle such new and diverse modalities. We envision plug and play sensor fusion approaches that can address these challenges.

Existing fusion algorithms often consider only two or three modalities. With new advances in multimodal sensors, it is important to design machine learning solutions that can robustly and effectively combine heterogeneous data coming from multiple sensors. When adding new channels, current solutions to fuse sensors exponentially increase the model parameter space, which leads to models that either are undertrained due to limited data, or cannot capture the true relationships across modalities. New multimodal fusing solutions should properly scale as more modalities are added. They should also provide capabilities to synchronize data streams, regardless of their sampling rates.

An important technical challenge is to design fusion algorithms that are robust even when data from some modalities are incomplete or absent. There are several scenarios that can lead to missing or incomplete data from modalities. From an accessibility perspective, individuals with physical impairment may not be able to operate or use certain modalities. The users may also decline the use of a sensor to protect their privacy. The placement of the sensors and the environment may temporarily lead to missing information from the users (e.g., occlusion, out of field of vision, and acoustic noise). When this happen, it is important that the next generation of AI systems be able to leverage the remaining modalities to complete their task. This technical challenge requires more effective leverage of the complementary nature of heterogeneous data coming from different sensors and design strategies to train and test models with partial information.

Current multi-channel fusion approaches assume that the sensors are predefined. We envision models that are modality agnostic, so they can work with whatever channels are available. Advances in this area will allow systems to scale to different modalities based on user preferences. The fusion formulation should enable replacement of current sensors with better modalities over time by creating models that easily integrate with new sensors. We envision data agnostic modules that can be interchangeably implemented with different sensors. By defining data modules, these systems can also scale by adding new, possibly unforeseen modalities.

Stretch goals: By 2040, plug and play multimodal sensor fusion systems will adapt to individual users, seamlessly handle the addition of new modalities as the developed, and scale to different modalities based on user preferences. Milestones along this path include—

Milestones

5 years: Multimodal system can learn from realistic datasets, exercising multimodal capabilities in the context of a range of real-world tasks.

10 years: Voice-based interfaces will recognize inputs from user with diverse speech patterns, such as deaf speakers, strong accents, and rare dialects.

15 years: Systems using three or more modalities, fusing input from voice, vision, and sensors allow for natural interaction.

Privacy Preservation for Multimodal Data

Multimodal interaction algorithms leverage sensors such as cameras and microphones to capture inputs like speech and gesture, but the data captured by these sensors can also capture and reveal sensitive information and the user, their environment, and bystanders. Systems should be designed such that the potentially sensitive data collected from these sensors is not shared further than it needs to be, and neither retained nor used for secondary purposes without permission. Some data is needed locally for real-time interaction (e.g., watching someone's hands to detect gestures such as pointing) and can be discarded immediately after use. Other data needs to be used to train personal models to support long-term interaction, and some might be used to train or refine more global models and algorithms. When possible, sensitive data could be processed in the platform (e.g., sensor or web browser) where it is collected, which would then return processed results (e.g., removing parts of video that are not close to interaction areas, or removing faces of bystanders) to applications. Systems that run in the cloud should separate user-identifiable data from anonymized data and try to use algorithms or techniques that can operate with as little access to sensitive data as possible.

Stretch goals: By 2040, multimodal interfaces will respect user privacy. Milestones along this path include—

Milestones

5 years: Privacy preserved for systems using voice and vision as input.

10 years: Multimodal interfaces will work across different levels of privacy and fidelity.

15 years: AI systems will ensure privacy preservation through differential privacy, running only in trusted code, or running models and fusion locally or at the edge.

Collaborative Interaction

In the last decade, we have witnessed a broad adoption of virtual personal assistants such as Apple's Siri, Google's Assistant and Amazon's Alexa. These systems aim to help users to accomplish simple tasks, such as checking the weather, looking up directions, and playing music, through multi-turn interactions. Furthermore, they aim to detect and answer a user's factual questions. However, they are still far from being as efficient as human assistants. Furthermore, these systems are designed domain-by-domain (for example, weather or music), resulting in good coverage of few commonly used domains, but gaps for the much larger set of less common topics. Although there are a few examples of use of context (for example, previous turns of a conversation, identity of the user), these behaviors are laboriously engineered by hand and the broader context is not considered. Except a few example cases, these systems do not integrate common sense (for example, "later today" can mean 4-5 p.m. in the context of business meetings, but 7-8 p.m. in the context of dinner discussions). In general, today's AI systems use special purpose data structures rather than general representations of meaning; such representations have been studied and challenges highlighted, but we don't yet have practical and effective means for representing the content of utterances. In addition, today's AI systems are largely unable to converse with two or more people in a group.

There has also been significant research undertaken in the space of representing, detecting, and understanding human emotions in fields such as computer vision, speech, and language processing. Different representations of emotions have been proposed, including discrete categories (e.g., joy or fear) and the continuum space of valence and activation. The area of sentiment and opinion analysis has also received significant attention from both academia and industry. We have seen significant advances in some areas, for instance systems are now available that can reliably detect the sentiment (positive or negative polarity) of text for certain domains, such as product or movie reviews; we have deployed computer vision tools that can detect smiling or frowning; we have speech models to sense the presence of anger. Progress has also been made in the use of multiple modalities for emotion and sentiment detection, and the fusion of different modalities is an active area of exploration. There is, however still much to be done to understand how to best represent emotions and how to detect emotions of different granularities and in different contexts. Moreover, the research area of emotion generation is just starting, and significant advances will have to be made to develop systems that can produce emotions and empathetic behaviors.

In order to make our AI systems be better partners, they must be collaborative, which requires the following technical advances.

Enabling Natural Interaction

Many of today's interactions with artificial intelligence technologies are limited, stilted, or awkward; they frequently lack the richness, elegance, and ease of typical interactions between people. As AI technologies become more commonplace in our daily lives, there is a need for these technologies to allow for more natural interactions that match the expectations of the human users. Much of the work in this area in the past has focused on the recognition of subtle perceptual cues (such as facial expressions or gaze direction) or on the production of fluid, naturalistic behavior (such as human-sounding speech or smooth, coherent gestures). While most of our systems are currently designed for short, transactional interactions, we must push toward building systems that support long-term interactions, remembering the content and context of previous interactions and shaping current interactions based on this knowledge, adjusting the expectations and behavior of the system as experience grows, and working to maintain long-term stability of behavior and personality. While most of our current systems are reactive (e.g., providing answers to posed questions or responding when tasked) AI systems of the future must be proactive in their activity by initiating conversations with relevant information, asking questions, and actively maintaining trust and positive social relationships.

Finally, in order to enable natural and beneficial interaction and collaboration with people, AI systems should be able to change their behavior based on people's expectations or state. AI systems should be able to provide personalized experiences that provide high levels of improvements on people's success and experiences. Personalization and adaptation capabilities are enabled through an understanding of context and by modeling of users, including their mental models (both described below).

Stretch goals: By 2040, AI Systems that can have extended meaningful, personalized conversation taking advantage of the context of the conversation. Milestones along this path include—

Milestones

5 years: AI systems that can have an extended dialog over a single topic.

10 years: AI systems that can have extended, personalized dialogs over a single topic, using context.

15 years: AI systems that can shift topics and can return to previous dialogs, keeping track of conversational state and conversational participant's intentions.

Alignment with Human Values and Social Norms

AI systems must incorporate the social norms, values, and context that are hallmarks of human interaction. Systems must begin to understand properties about social relationships (e.g., knowing that I am more likely to trust my spouse than a stranger), about objects and ownership (e.g., knowing that it might be acceptable to take and use my pen but not my coffee mug), and about social roles and responsibilities (e.g., knowing that the answer to the question "what are you doing?" should be answered differently when asked by my employer than by my co-worker).

Alignment of AI systems with human values and norms is necessary to ensure that they behave ethically and in our interests. Human society will need to enact guidelines, policies, and regulations that can address issues raised by the use of AI systems, such as ethical standards that regulate conduct. These guidelines must take into account the impact of the actions in the context of the particular use of a given AI system, including potential risks, benefits, harms, and costs, and will identify the responsibilities of decision makers and the rights of humans. Currently, AI systems do not incorporate the complex ethical and commonsense reasoning capabilities that are needed to reliably and flexibly exhibit ethical behavior in a wide variety of interaction and decision making situations. In contrast, future AI systems will potentially be capable of encouraging humans toward ethically acceptable and obligatory behavior.

Stretch goals: By 2040, AI systems will constantly and reliably reason about the ethical implications and social norm adherence of actions they engage in or observe. They will plan and adjust their own behavior accordingly, and will act to prevent others, both human and AI systems, from engaging in unethical behavior. In situations in which the ethical implications are complex and do not lend themselves to simple ethical acceptability verdicts, AI systems will engage in thoughtful conversations about these complexities. Milestones along this path include—

Milestones

5 years: AI systems will take ethical requirements and contextual clues into consideration when deciding how to pursue their goals.

10 years: AI systems will effectively reason about how to behave when faced with conflicting social norms and ethical acceptability tradeoffs.

15 years: AI systems will reason broadly about the ethical implications of their actions and the actions of others, perhaps more efficiently than humans.

Modeling and Communicating Mental States

Humans are capable of creating a mental model of intentions, beliefs, desires, and expectations of others. Understanding what someone else is thinking makes it easier to interact and collaborate with them. As AI systems become more widely accessible, it is important for them to accurately model mental states of their users and enable frictionless interactions.

Current, widely used AI systems have only very limited modeling a user's mental states. They mainly focus on determining which of a small pre-defined set of possible intentions the user might have. Furthermore, these systems aim to track user's intentions throughout conversations, the user specifies more information and sometimes change their mind. However, even in a simple interaction, the possible number of mental states is much larger than what these systems can model. Furthermore, as a user does not necessarily know about the set of intentions that were pre-defined by the AI system builders, they often try to interact with these systems with different intentions. The mismatch between what builders include in these systems and a user expects to result from inefficient interactions ends up limiting these interactions to a few simple use cases and prevents the application of these AI systems to many scenarios and domains. While the work on modeling and tracking users' intentions should continue to improve in these limited scenarios, future AI systems need to extend to the scale and complexity of real-world intentions.

To support collaboration, AI systems need to be able to reason and generate persuasive arguments that are not only relevant to the current situational context, but are also tailored to individual users based on shared personal experience. Argumentation and negotiation methods could be learned through extracting and learning from text.

Stretch goals: By 2040, AI systems will work collaboratively with multiple human and AI-based teammates using verbal and nonverbal communications, engage in mixed-initiative interactions, have shared mental models, be able to predict human actions and emotions, and perform collaborative planning and negotiation in support of shared goals. Milestones along this path include—

Milestones

5 years: Persuasive arguments will be generated based on shared experiences and goals.

10 years: Human-machine teams will regularly tackle hard problems collaboratively, using defined measures of engagement and perseverance in human-machine teams. AI systems will communicate with human teammates in task-specific contexts, using verbal and nonverbal modalities to establish shared mental models.

15 years: AI systems will exhibit persistent memory, enabling them to serve as long-term teammates. They will take both reactive and proactive actions to complement their human teammates' rational and emotional actions, and will communicate in natural language dialogs to plan, re-plan, and negotiate alternatives to support the team's goals.

Modeling and Communicating Emotions

Building systems that can represent, detect, and understand human emotions is a major challenge. Emotions must be considered in context—including social context, spatial and temporal context, and the context provided by the background of those participating in an interaction. Despite multiple proposed representations of emotions to date, there is still no agreement on what are the ideal representations that would fully cover the multiple facets of human emotion. The detection and tracking of emotions is also challenging, and it will require significant advances in data construction, unimodal and multimodal algorithms, and temporal models. The understanding of human emotions will assume reliable context representations, and the ability of the algorithms to account for the variability brought by different contexts.

People often reciprocate in their emotional expressions. Empathetic communication forms the ties of our society—both in personal settings and in more formal settings (e.g., patient-doctor relations, student-instructor interaction). Our expectation is that the people we communicate with will understand and eventually respond to our emotions: if we are sad, we expect understanding and encouragement; if we are happy, we expect similarly joyful reactions. Yet, most current AI systems lack the ability to exhibit emotions, and even more lack the ability to create empathetic reactions. Thus, significant research will need to be devoted to emotion generation, coupled with methods for emotion understanding. Since human emotions arise in the context of experiences that a person is undergoing, research on emotion must connect with that research for modeling context and common sense. A key technical challenge is how to model human experiences in a way that the AI system can assess its quality. This is difficult

because experiences are extremely varied, nuanced, and involve subjective judgments. To deal with the subjective aspects of experiences, we need to develop methods for modeling and reasoning with subjective data, whereas in contrast, previous work has focused almost exclusively on objective data. Ultimately, we hope to produce empathetic AI systems that act in a manner that is emotionally congruent with the users they interact with.

In addition to modeling a human's mental state, an AI system must ensure that its behavior is explicable and comprehensible to the humans with whom it is interacting. One way of ensuring explicability is for the AI system to explicitly consider the mental models of humans when planning its own behavior. In some cases, the system may decide that adapting to people's expectations is either infeasible or poses undue burden (cost) on the AI system or the team. For example, a robot in a collaborative search and rescue scenario might find that it is unable to rendezvous with the human partner in the agreed location because of collapsed pathways and debris. In such cases, the AI system should be able to provide an explanation, which involves communicating to the human the need to change the rendezvous point. These explanations need to be tailored to the mental models of the humans.

Stretch goals: By 2040, AI systems will reason about the emotional components of encountered situations and about the emotional effects of their actions and the actions of others involved in a wide range of situations, both real and fictional (e.g., stories, films, worst-case scenarios). Milestones along this path include—

Milestones

5 years: AI systems will reason about how their actions emotionally affect people with whom they interact.

10 years: AI systems will predict human emotions based on observing an individual at work, play, or during training, for as little as 30 seconds. When reading a short story, AI systems will construct models of the emotions of the characters in the story and how they change as a consequence of the story's events.

15 years: AI systems will reason about complex affect-aware interactions, e.g., anticipating how a person might be impacted by a team interaction, personal history, or particular context.

Supporting Interactions Between People

Human social interactions are increasingly conducted through online media and networks. Artificial intelligence can play an important role in facilitating these interactions and in mining online data for insights about social phenomena. While artificial intelligence already has a strong presence in online interaction, there are a number of possibilities for future research. These research possibilities can be structured by the purpose of the online interaction, which can include 1) deliberation, 2) collaborative creation, and 3) social-tie formation.

In deliberative online communication, existing state-of-the-art techniques use supervised machine learning to identify factual claims in text and distinguish statements of fact from opinions. However, like all learning-based approaches to natural language, there is still a question of domain specificity, as well as the development of technology for "low-resource" languages that lack labeled data. Online deliberations also provide valuable data for social scientists who are interested in understanding the dynamics of opinion formation, and we see a trend toward interdisciplinary research that leverages such data. However, online data poses unique challenges, and methodological advances are required to ensure the validity of the resulting inferences.

Artificial intelligence already plays a limited role in collaborative creation. Wikipedia, the online encyclopedia, uses AI bots to prevent vandalism and ensure that content meets existing standards. More advanced capabilities are at the level of research prototypes: for example, recommending collaborators and summarizing the existing literature in an area of scientific research.

Social media analysis is an active area of research, with substantial progress toward commercialization. For example, AI is already widely used to detect objectionable content such as hate speech. However, deployed solutions for content filtering in online media are typically human-machine hybrids, with AI used only as a preliminary filter, reducing the workload of human moderators. Questions remain about whether online platforms and communities can reach consensus about definitions of objectionable

content; this is a baseline requirement for the deployment of any artificially intelligent system. AI is also playing an increasingly active role in content generation and manipulation: for example, through predictive text entry and machine translation. We foresee increasingly sophisticated human-machine hybrid technologies in this space, which may ultimately enable new forms of human-human interaction and emergent social structures. In order to realize this vision, we need progress in all three high-level directions. We also foresee more sophisticated machine-machine interactions with AI systems that represent people and that can help to promote new human-human and human-machine modes of interaction, be this with new or strengthened social ties (including coordinated activities), new possibilities for discourse, and new opportunities collaborative creation. Machine-machine interaction is also, in and of itself, an important component of AI interactivity and it is important to continue to develop theory, algorithms, and formalisms for multi-agent systems so that ecosystems of interacting AI systems will achieve intended outcomes and represent successful extensions (and reflections) of human society.

Another active area of current research involves linking online social communication to external phenomena, such as economics, politics, and crisis events. However, existing approaches are typically "one-off" technical solutions that are customized for the event and data sources of interest. We still lack a set of robust, general-purpose methods for jointly reasoning about these very different types of data.

Improving Deliberative Interactions Between People

Individuals are increasingly using online platforms to participate in deliberative discussions. This trend provides new opportunities as well as new challenges for new AI systems to utilize and address. Below, we identify and elaborate on four technical challenges.

Identifying factual claims

The Web and social media have democratized access to information. However, the decentralized nature of these platforms, the lack of a filtering mechanism, and the impossible challenge of inferring credibility of individuals and information online also enable the widespread diffusion of misinformation and disinformation, threatening to the health of online conversations. Given the growing pace with which misinformation and disinformation spread online and the societal implications of this trend, it is crucial to build AI systems that detect and combat these phenomena. Such systems need to be scalable, accurate, and fair. One important building block here is the detection and validation of factual claims in online data. For instance, given a text (e.g., a post on social media), what are the claims of facts (e.g., "the moon is made of cheese"), and can we check these facts against external resources (e.g., Wikipedia, textbooks, research articles, knowledge bases, etc.)?

Current research in claim detection often relies on fact databases and conventional statistical methods. These solutions have limited accuracy and generalizability—due to the limited scope of fact databases. Future research in claim detection should broaden the application domains, work across different communication modes, and be interactive and responsive. The claim validation step is currently at its infancy. We expect the next 10 years to lead to more accurate and adaptive AI systems. Such systems should have a strong fairness emphasis (e.g., how are the false positives of a false-claim classifier distributed across individuals from different communities?). Furthermore, current systems focus on batch claim validation. Yet in cases of misinformation, timely action is crucial. This requires investment in real-time detection methods.

Understanding and designing metrics

Current research on the health of online deliberation has a strong emphasis on civility. There are systems (e.g., the Google Perspective API) that accurately model civility of text data across different domains (e.g., Wikipedia, news commenting systems). However, there are various other qualities beyond civility—identified by social science research and grounded in the principles of deliberation—that are currently unexplored by AI systems. Some of these qualities concern how participants treat one another over the course of the discussion (e.g., acknowledging others' contributions, feelings, or beliefs). Other qualities concern the ways that participants engage with the topic (e.g., providing evidence for held beliefs). We expect future research in AI to model such dimensions to provide a richer understanding of conversation health. Furthermore, current research mostly focuses on modeling content at the individual

content level (e.g., a single tweet, a Reddit post). We expect future research to move beyond this building block and accurately model users, discussions, and communities. Such granularity of modeling will enable better intervention systems.

The diversity of online discussion spaces introduces yet another challenge. Markers of conversation health revered by one community can be irrelevant for another. Communities might have different significance they assign to each conversation quality (e.g., Do we value civility or diversity of thoughts more?). We expect future AI systems to learn these community priorities and goals through passive and active interactions with community members (which will map to revealed and stated preferences) and label content accordingly. In addition, the interpretation of conversation qualities might differ across communities (e.g., What counts as civil in the "change my view" and "parenting" subreddits may be quite different.). Current AI systems heavily rely on crowd-workers to label content according to the codebooks constructed by researchers with limited community input. Accordingly, the models do not reflect the values of the communities they are meant to model/guide. We would like future AI systems to engage the community both at the codebook construction and data labeling to identify community-level deliberation health measures.

It is important to note that online communities do not exist in a bubble. What should an AI system do in the case of a community that aims to share and spread health misinformation? Identifying when a local community-level measure of conversation quality is to be preferred over the more traditional definition put forth by social scientists will be yet another important challenge for responsible AI.

Identifying consensus on facts and goals

Given the transcript of an online multi-party discussion (or set of discussions), can we identify the facts that are universally agreed upon? For example, a transcript of a conversation between economists might reveal strong agreement on the rational choice model of decision making. Current human-computer interaction research leverages human intelligence to perform discussion summarization (e.g., Wikipedia discussions, slack channels). Future AI+HCI+network science collaborative solutions can use text and network features to scale these solutions up.

Similarly, given the transcript of an online multi-party discussion (or set of discussions), can we identify the goals that are shared by the participants/community members? Identifying community goals is a more challenging task, as the revealed and stated preferences might diverge. Furthermore, individual participants may not be cognizant of their own goals. Future AI systems that integrate user modeling with text (stated goal) summarization techniques can provide better models. In the long term, AI could help communities elicit shared goals (e.g., a mission statement for a subreddit).

Affecting and facilitating online conversations

Future AI systems should go beyond simply identifying and understanding. The aforementioned measures and models should feed into systems that facilitate healthier, higher quality, and more fruitful conversations online. For instance, future AI systems can help online users improve the quality of conversation spaces they participate in by helping them craft messages that 1) directly contribute to quality and 2) indirectly inspire others to behave similarly. A message-crafting AI system might pull parts of the conversation thread that need attention (e.g., an unanswered question) or pull relevant external information from credible sources to enrich conversations. In order to convince individuals to follow algorithmic message-crafting suggestions, such systems should model users accurately and identify the optimal messaging and timing for these algorithmic suggestions. Furthermore, in order to indirectly inspire others, such systems should accurately model relationships and contextual cues—predicting the likely impact of a given message on the quality metrics for subsequent conversations. We believe the next 20 years will bring similar applications for facilitating factually correct conversations, helping communities identify community-specific health metrics, as well as facilitating community goal-setting processes.

Supporting Complex Teamwork

Some scenarios of collaboration move beyond short-term interactions with atomic tasks. People are often involved in collaborations that span a long time horizons and require team members to carry on complex activities while supporting each other's work. AI

systems becoming effective partners within teamwork opens up new challenges in modeling of context and people, and requires algorithmic advances in planning and multi-agent reasoning. Early work in models of collaboration and teamwork relied on constructs such as joint intentions or intention toward teammates' success. These important breakthroughs illustrated that teamwork is more than the sum of individuals doing their tasks, but these early models lack the ability to handle uncertainties and costs present in more complex real-world settings. Subsequent models offer a rich representational framework, but face problems due to computational complexity. Future research must combine these expressive representations with explicit models of intention, while yielding computationally practical algorithms. Furthermore, we must address other crucial aspects in effective human-machine teamwork, such as delegation of autonomy, modeling the capabilities of others, authority, responsibility, and rights.

Collaboration and Interaction in Social Networks

Social interactions are crucial for addressing many societally beneficial interactions, such as spreading information about HIV prevention in an at-risk population of homeless youth in a city. Facilitating this type of operation is a problem of collaborative interaction on a massive scale, which raises new challenges that bridge theory and practice. For example, consider the challenge of spreading information about health. In AI, this problem is recognized as one of influence maximization, i.e., given a social network graph, recruit a set of nodes in the graph (opinion leaders) who would most effectively spread this information. Decades of work have yielded important models of influence spread with theoretically attractive properties and algorithms with formal guarantees. However, it is still unclear whether these models of influence spread are reflective of how influence actually spreads in the real world. We need real-world studies to validate these models in physical (as opposed to virtual or electronic) social networks; the work should be interdisciplinary, combining AI methods with efforts in social work, sociology, and other related areas.

Stretch goals: By 2040, AI systems can monitor human discussions and automatically identifying consensus on facts and goals, as well as collaborative workflows and emergent plans. Milestones along this path include—

Milestones

5 years: Defining metrics for measuring various dimensions of the quality of online discussions and constructing methods for automatically estimating these metrics.

10 years: AI systems that can reliably identify factual and fraudulent claims from a conversational transcript, making use of background material available on the Web.

15 years: AI systems that can understand domain-specific conversational contexts, where it is also necessary to model communities (in terms of communication patterns, goals, and processes).

Collaborative Creation of New Resources

Organizational structures and collaborative creation: Online interactions are a vital part of collaborative creation of many different types of artifacts. While there have been studies on both the artifacts themselves (e.g., Wikipedia articles or open-source software), and the process of their creation (e.g., talk pages or pull requests), there is a significant opportunity in better connecting the social structures that result in collaborative creations. Clearly, the creations themselves are of significant interest and value (Wikipedia articles, open-source code, collaboratively crafted papers, legal documents, etc.). However, the process by which they were created is equally important. The process here includes the various online interactions of the participants as mediated by their organization (the social network, corporate structure, etc.). The relationship between the interactions and the created artifacts are both important in understanding the creative process but also in building AI-driven interventions that lead to better creation.

The collaborative process includes workflows that allow individuals and groups to manage and coordinate the creative goal. Individuals and groups coordinate to decide on how the work should be split, what each sub-component will look like (in form and function), and how the pieces will come together as a whole. Mining this data represents a unique challenge today as both the artifacts themselves and the interactive traces are varied in structure and form (text, images, video, code, votes, change logs, discussion forums, social networks,

community tags, etc.). A key AI challenge is in extraction and the modeling of the creations by people, their online interactions, and the connections between interaction and the creative process. Doing so effectively requires improving natural language understanding and text mining for everything from Wikipedia text, collaboratively created scientific text, discussions around code (e.g., Stack Overflow), the code itself (e.g., Github). However, an increase in multimedia content such as images, videos, and fonts (e.g., Behance) and audio (e.g., Soundcloud) presents new modeling challenges. Modeling the interactions will also require addressing new challenges. Existing interactive media (e.g., email) are being supplanted by innovative social platforms that must be modeled in different ways. More advanced extraction tasks may include identifying causal structures between the process itself and the creative product.

A consequence of better models of the interactive process is better AI-driven interventions to support a community or individuals in their creative goals. By determining which workflows result in higher quality outcomes, a challenge will be to build interventions that can appropriately provide support. For example, an AI system that is aware of both the goal (e.g., the design of a video) and the way the community suggests changes can intervene in a group discussion to suggest task breakdowns, bring in related clips from search engines, or produce alternative video sequences as they are discussed.

Synthesis and Context Bridging

Retrieving information as needed or anticipating needs is a valuable contribution to the creative process. However, a more significant benefit is enhancing this participation to contribute novel synthesis and to bridge context. For example, an AI-system that can model what software engineers are discussing and what they have previously built may be constructed to provide alternative architectures for a new module. Such intervention requires a significant understanding of the goals of the community, where they are in the process, and the ability to bridge this model to other knowledge bases. The opportunity is an interactive system that can more actively participate in the creative process. Rather than simply retrieving relevant information, this information can be synthesized and translated into an appropriate form. The "query" for such a system is the ongoing interactions of the participants, a model of their goals, and the state of their creations (e.g., the text, software, lab notebooks, etc.). The research challenges include improvements to text modeling, goal and knowledge modeling, integrating causal models, understanding the network structures of the interactions and how the organization makes decisions. Further interactive challenges require creating the appropriate kind of query response, a process that requires synthesis, summarization, and translation across domains.

An example of this type of system could be one that supports scientists in generating hypotheses. For example, given a single research paper or a set of research papers, a task might be to mine collections of research articles to identify relevant prior work that might have been missed and then to provide a summarized synthesis of this information. A specific challenge for this is in bridging the vocabulary differences across communities (e.g., all the different names for factor analysis). A more ambitious intervention (20-year time frame) would be for the system to understand the discussions around the hypotheses, model the hidden assumptions of the researchers and then to identify additional hypotheses that have not been tested. These may naturally follow from existing results or tacitly underpin prior work. This requires not only analyzing the text of the documents but extracting the hypothesized causal models and reasoning about them.

Stretch goals: By 2040, AI systems model the hidden assumptions of participants and identify additional information (e.g., hypotheses) that have not been discussed, using information from external resources as needed. AI systems engage human collaborators by suggesting possibly fruitful activities and task decompositions. Milestones along this path include—

Milestones

5 years: Improved reputation systems for gauging the quality of human contributors and the effectiveness of team interactions.

10 years: Goal-sensitive monitoring of human team activity that learns which workflows are most productive for a given team and objective task.

15 years: Interactive systems that participate in the creative process while modeling the participants, their goals, their interactions, and the state of their "creations."

Building Stronger Social Ties

Linking social media to real-world social data: The increasing prevalence of social interaction in online media has enabled a new wave of social scientific research that aims at linking real-world events to their human causes and consequences. For example, metadata can be used to link messages on Twitter to real-world events such as natural disasters and civil unrest. The relevant messages can then be analyzed using natural language processing, computer vision, and social network analysis. This application of artificial intelligence produces real-time situational awareness of rapidly unfolding crises; and in the long term, it can offer new insights about the causes and consequences of events of interest. However, a key challenge for such research is making valid and reliable inferences based on online data. Online data is typically incomplete and non-representative; it may be confounded by existing algorithms such as content recommenders; and it can be sensitive with respect to the privacy of the individuals who are included in the data. Similarly, event logs may also be incomplete: for example, we usually have data only for crimes that were reported, rather than all crimes that were committed. Working with this data — and linking it to real-world events — therefore requires new methodological principles. We see an opportunity for the development of more sophisticated techniques for reasoning about causal structures and potential confounds. This research will require new bridges between the predictive technologies that are typically favored in computer science and the explanatory methods that are typical of the social sciences. Interdisciplinary teams of AI researchers and social scientists may be best positioned to make such advances. Future research in this space will also depend critically on the availability of data. One possibility is for public-private partnerships that might enable researchers to access data held by social media corporations, while preserving the privacy guarantees that these platforms offer their users. There is also a role for increasing accessibility to government data, so that social media can be linked against records of real-world trends and events.

Social Assistants: In a world of digital, multimodal communication with ubiquitous connectivity, smartphones, collaboration platforms, and online social networks, an increasing challenge is to manage the vast amount of information that is available and find ways to hold meaningful conversations. There is a role for AI in managing this communication and enabling healthy interactions, whether among our friends and acquaintances, in public spaces, or while at work. Today, we have simple email assistants that suggest that you sleep on an angry message before sending it, recommend possible responses to an email, and remind you when you should respond to an email. But these suggestions have limited personalization and don't grow to understand the human's social relationships in any deep way. Future research in this space could develop AI with the capability to more fully understand social context, to identify which kinds of communication, with whom, when, and on what topic will be especially valuable, and to automatically handle as much communication as possible. This can be done in a personalized way by AI that is responsive to our goals and our desire for privacy, is cognizant of our social and business relationships, and is considerate of our current context so as to know when to interrupt and when not, as well as what can be automatically handled and where interaction is required. AI can also play a role in helping us to identify opportunities for new kinds of interactions. A future capability is for personalized AI systems to better understand the preferences, goals, and current context of individuals (as described in Section 3.2) so that AI systems can interact with each and identify new opportunities. Examples of this include: suggesting a conversation with a stranger on a particular topic, identifying a goal that an ad hoc team of individuals are motivated to achieve (along with an outline of a plan for how to achieve it), or even enabling new forms of democracy by identifying representative groups of people who are well informed and willing to respond to a poll about an important issue that is affecting their city or state.

Stretch goals: By 2040, AI assistants will facilitate healthier, higher quality, and more fruitful conversations online and better processes for creative teamwork, as well as work to enable people to identify and then collaborate in achieving shared goals. Milestones along this path include—

Milestones

5 years: Learning methods for modeling human interaction, predicting emotional reactions, and tracking long-term happiness of users.

10 years: AI systems will track simultaneous news and sensor feeds, linking real-world events to their human causes and consequences.

15 years: AI systems can identify and automate valuable communication opportunities in the context of deep models of human relationships, social context, and privacy concerns

Making AI Systems Trustworthy

AI systems have already seen wide deployment, adoption, and trust for a number of practical problems. They have been particularly successful when the AI system is very accurate, the problem has low stakes, or when there is an easy way for the human to supervise and correct the system's mistakes. For example, music streaming services allow people to steer recommendations via likes and dislikes, spell checkers and auto-completion systems display multiple alternatives to allow people to select correct suggestions, and personal voice assistants typically ask for confirmation of spoken commands when they are uncertain.

Transparency and Explainability

If we cannot construct a system that can explain its behavior to users, a set of complicated questions come into play. If we are designing an inscrutable AI system, how do we transparently incorporate values into the design? How do we formally define and capture the different types of explanation and interpretation depending on the setting (explanation for recourse, explanation for understanding, explanation for persuasion, and so on)? How would we capture the subjective nature of many notions of trust, for example the fact that multiple stakeholders in a system might have different expectations from a system? These challenges are fundamentally socio-technical, in that they have to recognize the ambiguous and often messy human element at the core of any AI system. But this is a challenge that is in the spirit of AI and interaction, especially when viewed in conjunction with other ideas such as modeling human mental states and collaborations.

Explanation will be key to developing mechanisms to build trust between AI systems and their human interlocutors. While developing these methods is important, there are broader design concerns that also need to be addressed. These include: how to transparently incorporate human values into the design of AI systems, breaking down their inscrutability barriers; how to formally define and capture the different types of explanation and interpretation in ways that are tailored to a given setting (e.g., explanation for recourse, explanation for understanding, explanation for persuasion, etc.), and how to capture the subjective nature of many notions of trust, for example, the fact that multiple stakeholders in a system might have different expectations.

These are fundamentally socio-technical challenges, and effective solutions must recognize and handle the ambiguous human elements involved in interactions.

Stretch goals: By 2040, AI systems will be able to reason about trust in both one-on-one interactions and in teams, including how a person's trust might be impacted by a team interaction, personal history, or specific context. The operations of AI systems will be well enough understood by their human partners to build trust in the system over time, even as the system learns and evolves over months, years, and decades. Milestones along this path include—

Milestones

5 years: AI systems will generate concise, human-understandable explanations for machine learning algorithms, summarize specific reasons for independent actions, and explain recommendations.

10 years: AI systems will analyze their decisions according to confidence, potential, uncertainty, and risks, and will compare alternatives along these dimensions with detail or at a high level of abstraction. Trust levels in AI systems will be measurably increased based on these explanations.

15 years: Human partners will be able to understand a new AI system based on explanations of its operations provided at a level that an untrained person can understand.

User Control of AI System Behaviors

Users should be able to accurately assess and modify AI system behaviors when appropriate. Ultimately, AI systems are intended to improve human lives and it is therefore imperative we consider the human factors involved in correctly understanding their

capabilities and limitations in order to prevent undesirable outcomes. Engendering appropriate levels of trust requires considering the psychology of people basing their actions or decisions on the outputs and explanations of AI systems. For example, research has shown that the mere length of an explanation can have a significant impact on whether or not a person believes an AI system. Moreover, because AI systems learn and therefore evolve over time, it is important to ensure people remain vigilant in vetting AI behavior throughout continued use: An AI system that worked in a specific situation in the past may not work the same way after it is updated.

It is thus critically important to develop approaches that AI system creators can use to convey model uncertainty, help people explore alternative interpretations, and enable people to intervene when AI systems inevitably make mistakes. Further, it is important to realize that the accountability for the effects of the AI systems resides with the developers. Policies should be enacted that will require research oversight through institutional review boards and industry equivalents. Doing so requires consideration of several factors, including characteristics of the target scenario and risk level. Furthermore, while confidence intervals may be effective for skilled end users interacting with AI systems through graphical interfaces, they may be inappropriate for dialog-based settings or time-pressured situations. Similarly, while AI behaviors can be easily adjusted in low-risk scenarios (e.g., email clients that automatically sort or prioritize emails based on importance often support correcting individual misclassifications), it is vital that we make advances in how to best support understanding and intervention in increasingly prevalent high-risk scenarios as well.

Stretch goals: By 2040, AI systems can be used in high-risk scenarios where users trust the system to convey uncertainty and allowing them to intervene as appropriate. Milestones along this path include—

Milestones

5 years: AI systems can be corrected by users when exhibiting undesirable behaviors.

10 years: AI systems can convey uncertainty of their knowledge and enable users to intervene accordingly.

15 years: AI systems communicate to users significant changes in their behaviors as they evolve over time.

Preventing Undesirable Manipulation

People are socially manipulative and persuasive as part of normal social interaction. We influence and are influenced by others' opinions, behaviors, emotions, values, desires, and more. We tend to be more influenced by those we trust and with whom we share a positive relationship. We have also long created technologies, artifacts, and media with the power and intent to persuade, influence behavior, and evoke emotion, such as movies, literature, music, TV, news, advertisements, social media, and more. Socially interactive, personified AI technologies can be designed to use multiple methods of persuasion: leveraging the interpersonal persuasiveness of a known and trusted other along with the technology-enabled persuasiveness of interactive media. To date, narrow socially persuasive AI systems have been designed and explored in research contexts for benevolent purposes, such as acting as health coaches, intelligent tutors, or eldercare companions to promote health, social connection, learning, empathy, and more. But just as social manipulation can be used to benefit people, it can also be used to exploit or harm. This raises important ethical considerations in the design of socially persuasive AI. For instance, a personal health coach robot that provides social support to help users change their behavior to be healthier is a welcome innovation. However, it would be undesirable if that same robot was operated by a company that exploited the robot's persuasive abilities to try to get users to sign up for expensive meal plans that are not needed. Whose interests does the AI serve and how would one really know? How can we mitigate unintended negative consequences or prevent exploitation by those who do not have our best interests at heart? There exist ethical and moral considerations in the design of socially persuasive AI, such as how to design and ensure fairness, beneficence, transparency, accountability, explainability, respect for human dignity and autonomy, promoting justice, etc.. We need to develop methods, technologies, and practices for responsible and trustworthy socially persuasive AI that will enable people to design, monitor, characterize, verify, and improve their behavior over time to benefit people and society.

Stretch goals: By 2040, users control and trust their AI systems to use persuasion when ethically appropriate, and to alert and protect them from intentional manipulation by others (humans or machines). Milestones along this path include—

Milestones

5 years: AI systems follow ethical guidelines when knowingly attempting to persuade users.

10 years: AI systems understand and notify a user when others (humans or machines) are acting persuasively against the preferences of the user.

15 years: AI systems detect and control malicious persuasion or intentional manipulation by others (humans or machines).

3.2.4 CONCLUSIONS

There is a long way to go before AI systems can interact with people in truly natural ways and can support interaction between people online. Building AI systems that are transparent and that can explain their rationales to end users and developers alike will help enable the development of systems that are unbiased and can be trusted, for it is through explanation that biases can be discovered. In 20 years, if research is properly supported, we will see AI systems that can communicate naturally and collaboratively with a diverse range of users regardless of language, dialect, or ability. We will see AI systems that can communicate using multiple modalities understanding how gesture, speech, images, and sensors complement each other. We will see AI systems that can support the human creative process, suggesting hypotheses and enabling better collaboration. We will see unbiased, transparent AI systems performing high-stakes tasks that humans can trust.

3.3 A Research Roadmap for Self-Aware Learning

3.3.1 INTRODUCTION AND OVERVIEW

The field of machine learning (ML) seeks to provide learning and adaptation capabilities for AI systems. While classical computing based on programming relies on human ingenuity to anticipate the wide range of conditions in which a system may find itself, AI systems must learn autonomously from a variety of sources including labeled training data, tutorial demonstration, instruction manuals and scientific articles, and through interactions with the physical and social world. They must also be able to adapt in order to compensate for changes in the environment and context, changes in the goals of the users, and changes in software, sensors, and computer and robotic hardware. All AI systems are limited by their sensors, their training experiences, and the vocabulary and representations in which their knowledge of the world is expressed; nonetheless, they need to behave robustly in the face of these limitations. This requires a deep understanding of their own uncertainty and a mandate to confront difficult decisions with caution. This is especially important in the face of adversarial attacks that seek to find and exploit the weaknesses and blind spots of all computer systems. Avoiding unintended biases that may be inherent in the way a system has been trained is another important consideration.

While machine learning methods transform the way we build and maintain AI systems, they do not eliminate the need for software engineering. Programming is still required, but at a higher level, where a human expert selects the training materials, designs the representation (i.e., vocabulary and structure) in which the learned knowledge will be captured, and specifies the measures of success. It is often the case that very large quantities of labeled data are needed for training and testing purposes, introducing significant labor and expense. Another critical task of AI designers is to check the correctness of what has been learned. This can be a real challenge, as it requires the AI system to be able to visualize and explain the learned knowledge. Ideally, an AI system would be aware of its own capabilities—it should be able to analyze what it has learned, characterize its boundaries and limitations, and proactively seek opportunities to improve its performance through further learning.