Multi-modal Agents for Business Intelligence

Blue Sky Ideas Track

Jeffrey O. Kephart
IBM Thomas J. Research Center
Yorktown Heights, NY 10567 USA
kephart@us.ibm.com

ABSTRACT

Given their installation on nearly a billion consumer devices around the world, consumers clearly enjoy using voice-driven assistants such as Alexa, Siri and Google Home, and find it compelling to interact with AI agents as quasi-human entities. Inevitably, people who are accustomed to using voice-driven assistants at home and in the car will expect to use such technologies in the workplace. What form will this take? Simple extrapolations from consumer space (e.g. running meetings or presentations) promise only modest value. I propose that AAMAS and the AI research community should pursue a bolder vision in which software agents act as quasi-human collaborators on core business intelligence tasks that entail analyzing data, diagnosing problems, and making decisions. Moreover, these business intelligence agents should communicate *multi-modally*, i.e. they must understand and employ speech complemented by non-verbal behaviors such as pointing, eye gaze, and facial expression. I outline requisite agent, MAS and AI technologies and pose several fundamental research challenges raised by this vision[1].

KEYWORDS

AI for Data Analytics; Intelligent User Interfaces; Multi-modal Human-Computer Interaction; Planning and Decision Support for Human-Machine Teams; Multi-agent Systems

ACM Reference Format:
Jeffrey O. Kephart. 2021. Multi-modal Agents for Business Intelligence: Blue Sky Ideas Track. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 6 pages.

1 INTRODUCTION AND VISION

Voice-driven assistants such as Alexa, Siri and Google Home are now installed on hundreds of millions of consumer devices worldwide. As of early 2020, there were nearly 90 million smart speakers installed in US homes [26] (up 32% from 2019) and 130 million in-car voice assistants [27] (up 15% over a 15-month period). The rapid proliferation and wide dissemination of voice assistants strongly indicates that a large proportion of the population finds it convenient and fun to ask an assistant to turn on the TV or radio, find recipes, make online purchases, or help optimize routes through traffic. In essence, an experiment has been conducted in the consumer space, and the results are unequivocal: humans clearly find it natural and compelling to treat AI as quasi-human entities with which they can interact via natural language.

Inevitably, people who use voice assistants in their homes and in their cars will expect to use them in the workplace. What form will such business assistants take? One view, embodied in products that have already reached the market such as *Alexa for Business* [18], sees business voice assistants as a relatively straightforward analog of their counterparts in the consumer space. They can be used for business tasks such as creating and retrieving schedules, starting meetings or presentations, or reporting issues to the IT department.

I suggest a more valuable role for AI assistants in the business world. What is the core purpose of most business meetings? Often, the topic revolves around visualizing data, analyzing it, and making decisions based upon it. The market for Business Intelligence (BI) tools that provide such capabilities is growing at a Compound Annual Growth Rate of 7.6%, and is projected to reach $33.3B USD by 2025 [39]. Such BI tools, offered by companies such as Microsoft, IBM, Tableau, SAP, Oracle, Salesforce, etc. increasingly incorporate AI technologies [47], and this increased uptake is expected to accelerate over the next few years [34, 41]. In parallel with this trend, BI vendors are placing an increasing emphasis on making BI tools more accessible to a broader, less sophisticated class of users.

If we merge and extrapolate these two trends — voice-driven assistant proliferation and growing adoption and sophistication of BI tools — we can envision today's BI tools evolving into software agent *collaborators* that assist people with visualizing data, analyzing it, and making decisions based upon it. Some BI vendors have taken a first step in this direction by incorporating voice-based assistants into their products [47]. But an important ingredient is missing. What qualities do we most desire in human collaborators? We prefer partners who are not just highly competent in skills that complement our own; they must also be *effective communicators*.

What do software agents require to be effective communicators? Consider how people communicate with one another. Speech is a key modality, but non-verbal modalities such as pointing and other hand or body gestures, gaze, and facial expressions are more important than we often realize. Non-verbal communication is particularly relevant in data-rich business intelligence scenarios because people are visual creatures[2] who find it natural to interact with visual representations of data, analytics and decision options [5, 54]. We don't just talk about data; we look at it, point at it, and manipulate it, thereby generating naturally a rich stream of non-verbal actions that accompany and complement what we are thinking and talking about. Verbal and non-verbal channels of communication

[1]The author is grateful to his many colleagues at IBM and RPI, whose names may be found among co-authored references cited herein, for stimulating discussions and joint research that have profoundly influenced his understanding of this subject. He also gratefully acknowledges insightful comments from anonymous AAMAS reviewers.

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

[2]50% of the brain's neural tissue and 2/3 of its electrical activity pertain to vision [49].

are not merely coincident; they are deeply intertwined and deeply dependent on context that has been established by previous communications and actions. People have little difficulty understanding and interpreting multi-modal communication from other people, particularly in extended interactions in which significant context has been established. A central challenge that underpins our vision is to develop software agent technologies that understand contextual multi-modal communication from humans, and respond multi-modally in ways that resonate with humans.

Consider a scenario in which a business person interested in finding suitable business partners says, "I want to see others like this." In the absence of any sort of context, the meaning of this utterance is unclear. However, suppose that this person is simultaneously pointing at a table of companies containing information such as the company's name, the number of employees, the last few years of financial data, and a paragraph of high-level general information about the company. Then one might infer that "this" refers to a specific company to which the speaker is pointing, and consequently "others" most likely refers to companies as well. The word "like" indicates that the speaker seeks companies that are similar to the company to which they were pointing as defined by some similarity metric. The desired metric might be inferred from recent conversational context, or the nature of the columns to which the speaker has pointed recently. For example, if the speaker is pointing at a company's *description* field, a text match to the description fields of other companies would be most appropriate. If numerical attributes are more recent, then a Mahalanobis or other numerical similarity metric might be applied to them.

The natural architecture for multi-modal business agents is a cooperative multi-agent system (MAS) composed of individual agents that encapsulate various business-relevant skills and communicate with one another and with humans. Framing multi-modal business intelligence assistants as MASs suggests a complementary perspective. Rather than thinking of multi-modality as a way to bring agents into the world of humans, take the inverse point of view: it allows us to build MASs in which some agents are people. People are useful additions to multi-agent systems. They have unique skills that software agents don't possess, such as greater general knowledge of the world, stronger general reasoning skills, and better intuition. However, a major barrier is that most humans are not fluent in KQML, FIPA ACL, or any of the more popular agent-to-agent communication languages [17, 44]. In order to build effective MAS teams that include humans, agents must accommodate this unfortunate human deficiency by learning to communicate multi-modally. While this perspective may seem tongue-in-cheek, it connects with an extensive multi-agent literature on human-agent teams [28] and suggests a research thread dedicated to re-examining agent languages, protocols, and other coordination mechanisms from this perspective. Might English or Mandarin become the next ACL?

The next section outlines some essential prior work that serves as a foundation for multi-modal business intelligence agents and multi-agent systems. Then, scenarios are introduced to illustrate how individual agents and multi-agent systems might be employed for business intelligence or related government and science applications. The final section outlines a cross-disciplinary set of essential challenges that the research community must address in order to realize our vision.

2 FOUNDATIONS

An early and prescient vision of humans and machines working in partnership on cognitive tasks such as decision making was described by Licklider [31] in his well-known 1960 paper "Man-Computer Symbiosis." He foresaw the importance of voice-based communication for real-time decision-making: "If computing machines are ever to be used directly by top-level decision makers, it may be worthwhile to provide communication via the most natural means, even at considerable cost."

Two decades later, Bolt [3] implemented a system that interpreted and responded to human speech. Moreover, he demonstrated that human-machine communication could be enhanced synergistically by adding gesture as another modality. Users manipulated graphic objects depicted on a display by pointing at them and issuing commands such as "Put that there." Insightfully, he recognized the power of pronouns such as "this" and "that" in such a scenario: "... the pronoun as verbal tag achieves in the graphical world the same high usefulness it has in ordinary discourse by being pronounced in the presence of a pointed to, visible graphic ..."

In the 1990's, less-encumbered gesture recognition technologies for manipulating objects in a virtual world were developed, including the ALIVE system of Maes et al. [37]. Subsequent researchers explored additional modalities and form factors. Pentland [42] and Ekenel et al. [15] augmented gesture recognition with facial expression recognition to enable smart boards or smart rooms to understand the mental state of users. Brooks [4], Coen [11] and Chen et al. [8] created "intelligent rooms" that freed users from the instrumented chair to which they had been tethered in Bolt's system. More recently, Farrell et al. [16], Kephart et al. [24], and researchers at RPI [1, 13, 52] have explored multi-modality in smart room environments, taking advantage of recent advances in speech recognition, head orientation and gesture recognition based on 3D cameras such as Kinects[3].

Early discussions of software agents as a new paradigm for user interaction that elevates "tools" to "collaborators" include Maes et al. [35] and the COLLAGEN system of Rich, Sidner and Lesh [46].

Also particularly noteworthy is the CALO ("Cognitive Assistant that Learns and Organizes) [50] effort that was supported by the PAL ("Perceptive Assistant that Learns") [51] program of DARPA. During its 5-year tenure from 2003 to 2008, CALO funded 300 researchers and 500 AI papers centered around cognitive assistants and the machine learning and other technologies upon which they are based. While CALO's original intent may have been to develop cognitive battlefield or workplace assistants, its main practical outgrowth turned out to be Siri, now one of the most ubiquitous of today's voice-driven assistants in the consumer space.

In summary, the dream of natural human-machine interaction arose from a desire to integrate machines more intimately into the process of business, government and military problem-solving and decision-making. Yet, while this vision has inspired a nice body of academic work, to date its most visible impact has been in the consumer space. Without doubt, Alexa, Siri and Google Assistant are entertaining and useful, and they will continue to flourish — but it is time to revivify the original vision, and get back to *business*.

[3] https://www.i-programmer.info/programming/hardware/2623-getting-started-with-microsoft-kinect-sdk.html.

3 SCENARIOS

What do I mean by "get back to *business*"? I construe "business" broadly, encompassing business, science and government scenarios that involve data visualization, analysis, and decision-making. I believe that multi-modal assistants that collaborate with humans on such tasks will become as ubiquitous as today's consumer assistants, and will have substantially greater value. A business scenario patterned after an actual mergers and acquisitions (M&A) prototype implemented with my colleagues [16] was outlined in the introduction. We have also implemented a science assistant[4] that helps astrophysicists visualize and analyze data pertaining to exoplanets [24]. Here I introduce a government scenario, using it as a basis for illustrating additional issues that help motivate the research challenges outlined in the next section.

Imagine that a local government's emergency response team is coordinating human resources to mitigate a current or impending natural disaster. Team members sit in a control center with a large display and converse with a cognitive assistant, issuing a sequence of contextually-linked commands like:

(1) "Show me counties with the worst storm damage."
(2) "Show me hospitals in **these areas** (*pointing to map*)."
(3) "Generate disaster relief plans that protect **them**."
(4) "Help me decide which of **these plans** is best."

During this sequence, the agent displays maps with overlays depicting current or predicted damage and the locations of critical facilities, plots of rainfall or damage vs. time, graphic inputs and controls for dynamically steerable simulations, and tabular representations of plans and schedules. As users look at, point at, and manipulate these graphic elements, they generate a rich stream of head poses, pointing, and other natural gestures. The assistant combines these non-verbal cues with co-occurring verbal utterances and the conversational context to infer human intent and respond with appropriate graphics accompanied by synthetic voice. For example, the meaning of "these areas" in the second command is inferred from the regions to which the user is pointing. Identifying "them" as "hospitals" in the third command is inferred from context through coreference resolution [53], and the meaning of "these plans" in the fourth is inferred from pointing or context.

Humans have evolved a very efficient form of communication that relies heavily on deictic pronouns like "this", "that" and "there" plus context that builds up dynamically over the course of extended interactions. We are so expert in this natural form of speech compression that it has descended into our subconscious. It would take a concerted mental effort not to employ it. Problem-solving and decision-making scenarios necessarily entail extended conversations and non-verbal interaction with graphic elements, thereby requiring business assistants to be adept at understanding natural human multi-modal and contextual communication. In strong contrast, today's consumer-space assistants have little need to cope with such subtle complexities of human interaction. For them, unimodal voice communication suffices because typical commands like "Play station WHYY," or "Set a timer for 5 minutes," are one-shot.

Another element that might appear on the display in this and other scenarios are humanoid or non-humanoid avatars. While there is considerable research interest in developing avatars to improve comfort, trust, and persuasion [2, 20], BI avatars can serve other important functions. If distinct agents are responsible for displaying data, running simulations, and planning or scheduling, representing these capabilities as distinct "expert" avatars could reinforce the user's mental model of the system's capabilities and how they can be invoked. Moreover, especially when customized to a user's individual taste or expressive style, avatars could serve as a focal point that non-verbally expresses a state or desire, such as confusion or a desire to speak.

While the cited business and science scenarios and this government scenario posit a shared large display, even a simple laptop possesses all of the basic components required: a microphone, a built-in camera, speakers, and a display with a pointing device. Indeed, if such agents are to become ubiquitous, it is likely that they will become so by being deployed mainly on laptops or even mobile phones, with intelligent rooms being a high-end option.

[4] A video is available at https://www.youtube.com/watch?v=Fg_seQM9T0k.

4 RESEARCH CHALLENGES

Here is a cross-disciplinary set of research challenges that I believe are most important to address in order to realize the vision of multi-modal business agents. The first two general challenge areas recognize that communication is a two-way street: assistants must both *understand* multi-modal communication and *respond* in kind. The third deals with measuring an assistant's effectiveness, while the fourth concerns systemic issues that are important for *any* MAS.

1. Infer human intent and state-of-mind from multi-modal communication and context
   a. Capture verbal human behaviors accurately and inexpensively under natural conditions.
   b. Capture non-verbal human behaviors accurately and inexpensively under natural conditions.
   c. Accurately infer short-term human intent via multi-modal fusion of verbal and non-verbal behavior time series.
   d. Accurately infer long-term human intent and state-of-mind via multi-modal fusion of verbal and non-verbal behavior time series.
   e. Develop adaptive online learning techniques to improve human intent inference accuracy.
2. Convey agent intent and state-of-mind via multi-modal communication and context.
3. Measure extent to which multi-modal AI assistants are helpful. Build and instrument end-to-end prototype and measure, under various multi-modal combinations:
   a. Human cognitive burden required to communicate intent
   b. Accuracy with which human intent is inferred
   c. Human-perceived enjoyment and effectiveness
4. Address issues of security, ethics, and trust.

Challenge area 1, capturing human intent, contains several subchallenges. Challenge 1a is largely an audio and software engineering problem that entails applying filtering and error-correction techniques to the input and output of commercial speech recognition engines [24]. Ironically, low-volume speech from passers-by or TVs poses a greater problem than moderate environmental noise because it tricks speech engines into spending inordinate resource straining to interpret irrelevant signals.

Interpreting non-verbal modalities such as eye-gaze, gesture and emotion is a vigorously-researched subject [29, 32, 33, 36, 38]. Challenge 1b highlights a key aspect of such work that is sometimes overlooked, yet essential to realizing our vision: recognizing such behaviors accurately and inexpensively under natural conditions. For example, dim or variable lighting can make it difficult to recognize gestures or expressions, especially in real time at reasonable frequency on minimal and inexpensive hardware. Behaviors of special interest include head orientation (to help infer whether a human is addressing a specific avatar or another person, or looking at a particular data representation), gesture, body position, lip movement (which can be correlated with audio signals to infer who is speaking), and facial expression. Chen [9] has made a promising start in this direction by developing a real-time, low-cost DNN-based approach to head orientation that copes well with uneven lighting and large angular excursions of the head.

Challenge 1c is about taking inputs from 1a and 1b and accurately inferring short-term human intent via multi-modal fusion. Previously, researchers have approached this as an engineering problem [16, 24] that entails feeding deterministic outputs from speech engines and gesture recognition systems into text classifiers and other NLU (Natural Language Understanding) techniques and applying hand-tuned parameter extraction rules to generate structured system commands. This approach scales poorly and hence is an impediment to achieving ubiquity. The solution may lie in learning-based approaches that combine symbolic and probabilistic neural techniques, of which the Deep Regression Bayesian Network [40] is one example. Symbolic representations like knowledge graphs and symbolic reasoning over them is likely to be necessary, as the final output of multi-modal fusion is a structured command that must satisfy certain logical constraints. While many suitable data sets exist for the various constituent technologies, there is a great need for new *multi-modal* data sets such as the CMU Panoptic Dataset [22]. Since context established by recent interactions and the state of graphic elements on the display is also an important ingredient in interpreting human intent, extending recent work on coreference resolution [10, 19] beyond purely textual NLP to include graphical context and non-verbal modalities may be fruitful.

Challenge 1d is like 1c, except that it addresses intent at the longer time scales that span across individual commands: What is the user actually trying to accomplish? Understanding long-term human objectives would allow agents to behave more proactively by using planning techniques that employ models of human intent and behavior. Early work by Chakraborti [6] suggests that inverse planning techniques might be able to infer human objectives from observed behavior. Challenge 1e has both learning and UI aspects.

Challenge area 2 concerns techniques that allow agents to *employ* multi-modality. One potential benefit is that it will make the agents seem more natural and appealing. Augmenting the generative models that underlie text generators like GPT-2 [45] to include emotion and prosody could prove quite interesting and valuable. If humanoid avatars[5] are employed, such techniques might be extended further to encompass the avatar's visual appearance and behavior. Second, such techniques could endow agents with enough social intelligence to gauge from human speech and gesture when they may interject, or to politely signal their desire to speak or act, thereby influencing humans to pause and let them into the conversation. There is broad scope for exploring non-verbal signalling techniques through which avatars can naturally communicate internal state or desires; for example, Divekar et al. [12] described a non-humanoid avatar that influenced users to direct their gaze so as to help the system ascertain whether it was being addressed.

In the consumer world, voice-driven assistants typically present themselves as individual agents with a single persona. An emerging class of business assistants [7, 16] and conversational business process applications [48] do so as well. However, their architecture is a cooperative multi-agent system that masquerades as a single entity by selecting the agent that is most capable of handling a given command. It might be advantageous to associate agents with individual avatars that can be addressed by name, eye gaze or head orientation. This mimics the way human experts interact with one another, and potentially enables the most appropriate agent to be identified with greater confidence. The use of multiple avatars to represent different *competing* agents was demonstrated recently in the context of the HUMAINE 2020 human-agent multi-lateral negotiation competition held at IJCAI 2020 [14]. If multi-modal BI agents become truly ubiquitous and large scale, the lines between *cooperative* and *competitive* multi-agent systems may blur. One can envision information economies [23] in which agents offered by different vendors will vie with one another in an ever-evolving information supply network to serve business clients.

Challenge area 3 concerns user studies that measure the extent to which multi-modal business assistants serve their intended purpose. Evaluating the resultant improvements in problem-solving and decision-making is more tricky than typical MAS experiments because human behavior is less reproducible than programmed agent behavior. The social intelligence assessment techniques pioneered by Malone et al. [25] and found in the social psychology literature [21] will likely provide important guidance.

Issues of security, ethics and trust must be addressed in *any* AI system. Challenge 4 concerns aspects of these issues that are particularly salient in our context, of which I highlight just three. First, given the need to collect fine-grained audio, video, and mouse event data, solutions will likely have both technological and architectural implications, such as storing and processing most of the data locally. Among the many ethics issues that arise is the need to ensure that decision agents help *reduce* rather than amplify [30] the biases inherent in human decision making [55]. Finally, to establish the foundation of trust on which effective collaboration rests, agents must be able to explain their behavior or rationale [43]. A simple example was implemented in our exoplanets prototype [24], which rendered a dynamically-generated AI plan into an English explanation when the user asked how a calculation was performed.

Addressing the myriad challenges required to realize the vision of multi-modal business agents greatly exceeds the capacity of any single organization. I hope this paper will help inspire a broad-based effort within the agents and AI research community to tackle these challenges. Achieving this vision would elevate AI from a tool to a collaborator, and thereby revolutionize the use of AI by business people, scientists, and policy makers.

[5] See for example the life-like ones developed by Soul Machines https://www.soulmachines.com/

REFERENCES

[1] David Allen, Rahul R Divekar, Jaimie Drozdal, Lilit Balagyozyan, Shuyue Zheng, Ziyi Song, Huang Zou, Jeramey Tyler, Xiangyang Mou, Rui Zhao, et al. 2019. The Rensselaer Mandarin Project—A Cognitive and Immersive Language Learning Environment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9845–9846.

[2] Amy L. Baylor. 2009. Promoting motivation with virtual agents and avatars: role of visual presence and appearance. *Philos Trans R Soc Lond B Biol Sci.* 364, 1535 (2009), 3559–3565. https://doi.org/10.1098/rstb.2009.0148

[3] Richard A. Bolt. 1980. "Put-that-there": Voice and Gesture at the Graphics Interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '80)*. ACM, Seattle, Washington, USA, 262–270. https://doi.org/10.1145/800250.807503 https://www.media.mit.edu/speech/old/papers/1980/bolt_SIGGRAPH80_put-that-there.pdf.

[4] R.A. Brooks. 1997. The Intelligent Room project. In *Proceedings Second International Conference on Cognitive Technology Humanizing the Information Age*. IEEE Comput. Soc, Aizu-Wakamatsu City, Japan, 271–278. https://doi.org/10.1109/CT.1997.617707

[5] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman (Eds.). 1999. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[6] Tathagata Chakraborti, Kartik Talamadupula, Mishal Dholakia, Biplav Srivastava, Jeffrey O Kephart, and Rachel KE Bellamy. 2017. Mr. Jones—Towards a Proactive Smart Room Orchestrator. *2017 AAAI Fall Symp. Ser.* (2017), 258–263.

[7] Praveen Chandar, Yasaman Khazaeni, Matthew Davis, Michael Muller, Marco Crasso, Q Vera Liao, N Sadat Shami, and Werner Geyer. 2017. Leveraging conversational systems to assists new hires during onboarding. In *IFIP Conference on Human-Computer Interaction*. Springer, 381–391.

[8] Harry Chen, Tim Finin, Anupam Joshi, Lalana Kagal, Filip Perich, and Dipanjan Chakraborty. 2004. Intelligent agents meet the semantic web in smart spaces. *IEEE Internet Comput.* 8, 6 (Nov. 2004), 69–79.

[9] Lisha Chen, Hui Su, and Qiang Ji. 2019. Face Alignment with Kernel Density Deep Neural Network. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[10] Kevin Clark and Christopher D. Manning. 2016. Deep Reinforcement Learning for Mention-Ranking Coreference Models. In *Empirical Methods on Natural Language Processing*. https://nlp.stanford.edu/pubs/clark2016deep.pdf

[11] Michael H. Coen. 1997. Building Brains for Rooms: Designing Distributed Software Agents. In *In Proc. of the Conf. on Innov. Appl. of Artif. Intell.* 971–977.

[12] Rahul R Divekar, Jeffrey O Kephart, Xiangyang Mou, Lisha Chen, and Hui Su. 2019. You talkin' to me? - A practical attention-aware embodied agent. In *Human-Computer Interaction – INTERACT 2019*.

[13] Rahul R Divekar, Matthew Peveler, Robert Rouhani, Rui Zhao, Jeffrey O Kephart, David Allen, Kang Wang, Qiang Ji, and Hui Su. 2018. Cira: An architecture for building configurable immersive smart-rooms. In *Proceedings of SAI Intelligent Systems Conference*. Springer, 76–95.

[14] Rahul R. Divekar, Hui Su, Jeffrey O. Kephart, Maira Gratti DeBayser, Melina Guerra, Xiangyang Mou, Matthew Peveler, and Lisha Chen. 2020. HUMAINE: Human Multi-Agent Immersive Negotiation Competition. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3334480.3383001

[15] Hazım Kemal Ekenel, Mika Fischer, and Rainer Stiefelhagen. 2008. Face Recognition in Smart Rooms. In *Mach. Learn. for Multimodal Interact.* 120–131.

[16] Robert G. Farrell, Jonathan Lenchner, Jeffrey O. Kephjart, Alan M. Webb, MIchael J. Muller, Thomas D. Erikson, David O. Melville, Rachel K.E. Bellamy, Daniel M. Gruen, Jonathan H. Connell, Danny Soroker, Andy Aaron, Shari M. Trewin, Maryam Ashoori, Jason B. Ellis, Brian P. Gaucher, and Dario Gil. 2016. Symbiotic Cognitive Computing. *AI Magazine* 37, 3 (Oct. 2016), 81. https://doi.org/10.1609/aimag.v37i3.2628

[17] Tim Finin, Richard Fritzson, Don McKay, and Robin McEntire. 1994. KQML as an agent communication language. In *Proceedings of the third international conference on Information and knowledge management*. 456–463.

[18] Matthew Finnegan. 2018. Alexa for Business: What it does, how to use it. *Computer World* (June 2018). https://www.computerworld.com/article/3279733/alexa-for-business-what-it-does-how-to-use-it.html

[19] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. arXiv:1803.07640 [cs.CL]

[20] Rosanna E. Guadagno, Jim Blascovich, Jeremy N. Bailenson, and Cade McCall. 2007. Virtual Humans and Persuasion: The Effects of Agency and Behavioral Realism. *Media Psychology* 10, 1 (2007), 1–22. https://doi.org/10.1080/15213260701300865 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/15213260701300865

[21] J. Hackman and Nancy Katz. 2010. *Group Behavior and Performance*. Vol. 32. https://doi.org/10.1002/9780470561119.socpsy002032

[22] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2017. Panoptic Studio: A Massively Multiview System for Social Interaction Capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[23] Jeffrey O. Kephart. 2002. Software agents and the route to the information economy. *Proceedings of the National Academy of Sciences* 99, suppl 3 (2002), 7207–7213. https://doi.org/10.1073/pnas.082080499

[24] J. O. Kephart, V. C. Dibia, J. Ellis, B. Srivastava, K. Talamadupula, and M. Dholakia. 2019. An Embodied Cognitive Assistant for Visualizing and Analyzing Exoplanet Data. *IEEE Internet Computing* 23, 2 (2019), 31–39. https://doi.org/10.1109/MIC.2019.2906528

[25] Young Ji Kim, David Engel, Anita Williams Woolley, Jeffrey Yu-Ting Lin, Naomi McArthur, and Thomas W. Malone. 2017. What Makes a Strong Team? Using Collective Intelligence to Predict Team Performance in League of Legends. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2316–2329. https://doi.org/10.1145/2998181.2998185

[26] Bret Kinsella. 2020. Nearly 90 Million U.S. Adults Have Smart Speakers, Adoption Now Exceeds One-Third of Consumers. *voicebot.ai* (April 2020). https://voicebot.ai/2020/04/28/nearly-90-million-u-s-adults-have-smart-speakers-adoption-now-exceeds-one-third-of-consumers/

[27] Bret Kinsella. 2020. U.S. In-car Voice Assistant Users Rise 13.7Nearly 130 Million, Have Significantly Higher Consumer Reach Than Smart Speakers – New Report. *voicebot.ai* (February 2020). https://voicebot.ai/2020/02/20/u-s-in-car-voice-assistant-users-rise-13-7-to-nearly-130-million-have-significantly-higher-consumer-reach-than-smart-speakers/

[28] Glen Klien, David D Woods, Jeffrey M Bradshaw, Robert R Hoffman, and Paul J Feltovich. 2004. Ten challenges for making automation a" team player" in joint human-agent activity. *IEEE Intelligent Systems* 19, 6 (2004), 91–95.

[29] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. 2016. DeepGaze II: Reading fixations from deep features trained on object recognition. arXiv:1610.01563 [cs.CV]

[30] Nicol Turner Lee, Paul Resnick, and Genie Barton. 2019. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. *Brookings Institute: Washington, DC, USA* (2019).

[31] J. C. R. Licklider. 1960. Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics* HFE-1 (March 1960), 4–11. http://groups.csail.mit.edu/medg/people/psz/Licklider.html

[32] J. Lin, C. Wu, and W. Wei. 2013. Facial action unit prediction under partial occlusion based on Error Weighted Cross-Correlation Model. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 3482–3486. https://doi.org/10.1109/ICASSP.2013.6638305

[33] Hongyi Liu and Lihui Wang. 2018. Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics* 68 (2018), 355–367. https://doi.org/10.1016/j.ergon.2017.02.004

[34] Michael Lock. 2018. *Moving Beyond Basic AI*. Technical Report. Aberdeen Group. https://www.ibm.com/downloads/cas/BWGRBGDM

[35] P. Maes. 1994. Agents that reduce work and information overload. *Commun. ACM* 37 (1994), 30–40.

[36] Pattie Maes. 2021. *Overview of Fluid Interfaces: Designing Systems for Cognitive Enhancement*. https://www.media.mit.edu/groups/fluid-interfaces/overview/

[37] Pattie Maes, Trevor Darrell, Bruce Blumberg, and Alex Pentland. 1997. The ALIVE system: Wireless, full-body interaction with autonomous agents. *Multimedia systems* 5, 2 (1997), 105–112.

[38] Y. Miyakoshi and S. Kato. 2011. Facial emotion detection considering partial occlusion of face using Bayesian network. In *2011 IEEE Symposium on Computers Informatics*. 96–101. https://doi.org/10.1109/ISCI.2011.5958891

[39] PR Newswire. 2020. *Business Intelligence Market worth $33.3 billion by 2025*. https://www.prnewswire.com/news-releases/business-intelligence-market-worth-33-3-billion-by-2025---exclusive-report-by-marketsandmarkets-301138062.html

[40] Siqi Nie, Meng Zheng, and Qiang Ji. 2017. Deep Regression Bayesian Network and Its Applications. *CoRR* abs/1710.04809 (2017). arXiv:1710.04809 http://arxiv.org/abs/1710.04809

[41] Michael Norris. 2020. *The Value of AI-powered Business Intelligence*. Technical Report. O'Reilly Media, Inc. https://www.ibm.com/downloads/cas/WWR6MK0X

[42] Alex P. Pentland. 1996. Smart Rooms. *Scientific American* 274, 4 (1996), 68–76. https://www.jstor.org/stable/24989483

[43] P. Jonathon Phillips, Carina A. Hahn, Peter C. Fontana, David A. Broniatowski, and Mark A. Przybocki. 2021. Four Principles of Explainable Artificial Intelligence. (2021). https://doi.org/10.6028/NIST.IR.8312-draft

[44] Stefan Poslad. 2007. Specifying Protocols for Multi-Agent Systems Interaction. *ACM Trans. Auton. Adapt. Syst.* 2, 4 (Nov. 2007), 15–es. https://doi.org/10.1145/1293731.1293735

[45] A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[46] Charles Rich, Candace L Sidner, and Neal Lesh. 2001. Collagen: Applying collaborative discourse theory to human-computer interaction. *AI magazine* 22, 4 (2001), 15–15.
[47] James Richardson, Rita Sallam, Kurt Schlegel, Austin Kronz, and Julian Sun. 2020. *Magic Quadrant for Analytics and Business Intelligence Platforms.* Technical Report ID G00386610. Gartner. https://www.gartner.com/doc/reprints?id=1-1XYUYQ3I&ct=191219&st=sb
[48] Yara Rizk, Vatche Isahagian, S. Boag, Yasaman Khazaeni, M. Unuvar, Vinod Muthusamy, and R. Khalaf. 2020. A Conversational Digital Assistant for Intelligent Process Automation. In *Business Process Management: Blockchain and Robotic Process Automation Forum (Lecture Notes in Business Information Processing)*, A. Asatiani (Ed.). Springer. https://doi.org/10.1007/978-3-030-58779-6_6
[49] S.B. Sells and R. S. Fixott. 1957. Evaluation of Research on Effects of Visual Training on Visual Functions. *American Journal of Ophthalmology* 44, 2 (1957), 230–236. https://doi.org/10.1016/0002-9394(57)90012-0
[50] SRI. 2018. *DARPA Cognitive Agent that Learns and Organizes (CALO) Project.* http://www.ai.sri.com/project/CALO
[51] SRI. 2018. *The PAL Framework.* https://pal.sri.com/
[52] Hui Su. 2017. The Cognitive and Immersive Situations Room. *XRDS: Crossroads, The ACM Magazine for Students* 23, 3 (April 2017), 20–23. https://doi.org/10.1145/3055149
[53] Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion* 59 (2020), 139–162. https://doi.org/10.1016/j.inffus.2020.01.010
[54] Edward R. Tufte. 2001. *The Visual Display of Quantitative Information* (2 ed.). Graphics Press, Cheshire, CT.
[55] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. http://www.jstor.org/stable/1738360