Multi-Agent Systems: Technical & Ethical Challenges of Functioning in a Mixed Group

Kobi Gal & Barbara J. Grosz

In today's highly interconnected, open-networked computing world, artificial intelligence computer agents increasingly interact in groups with each other and with people both virtually and in the physical world. AI's current core challenges concern determining ways to build AI systems that function effectively and safely for people and the societies in which they live. To incorporate reasoning about people, research in multi-agent systems has engendered paradigmatic shifts in computer-agent design, models, and methods, as well as the development of new representations of information about agents and their environments. These changes have raised technical as well as ethical and societal challenges. This essay describes technical advances in computer-agent representations, decision-making, reasoning, and learning methods and highlights some paramount ethical challenges.

For many decades after its inception, AI's most pressing question, its core challenge, was to determine whether it was possible to build computer systems able to perform intelligent behaviors like engaging in a conversation, playing chess, or fixing a complex piece of machinery. By the twenty-first century, the use of computer systems had evolved from a single person with computing expertise interacting with a single system to a highly interconnected, open-networked computing world in which people's online activities connect them instantly with many different systems and people. There are thus ever more situations in which AI agents interact in groups with each other and with people both virtually and in the physical world. AI's most pressing questions today – its core challenges – center on determining ways to build AI systems that function effectively and safely for people and the societies in which they live. Concomitantly, research in the multi-agent systems area of AI increasingly addresses challenges of building capabilities for AI agents to act effectively in groups that include people: for instance, investigating robot-human collaborations in industrial settings, coordinating health care for patients seeing multiple providers, and adapting educational content to individual students' needs. We refer to these as *mixed-agent groups*.

https://doi.org/10.1162/DAED_a_01904

AI research traditionally modeled the behavior of an individual computer agent, whether embodied in a physical system (such as robots) or embedded in a software system (such as recommendation systems or customer service chatbots), as an act-observe-update-decide cycle: the agent does something in its world, observes the ways that world changes, revises its beliefs about the world based on those observations, and determines what action, if any, to take next. Some AI agent models determine next actions based on maximizing a utility function, while others reason logically. These individual-agent models have regarded other agents, whether computer agents or people, as part of the agent's environment. To enable agents to participate effectively in mixed-agent groups required two significant modeling changes: the design of ways to represent the mental state of other agents and the development of models of human decision-making and communication capacities that respect the complementarities of human and computer-agent capabilities. For instance, computer systems have vastly greater ability than humans to access and summarize large amounts of data, while people's capabilities for causal and counterfactual reasoning far outstrip those of AI systems.

Mental state representations enable computer agents to treat other agents (whether human or computer) as full-fledged actors that have beliefs and abilities to make decisions, to act on those decisions, and to reason about the beliefs and actions of other agents in their environment. Computer agents can then recognize ways that actions of one agent may affect the beliefs and influence subsequent actions of other agents. Research on standard multi-agent models, including both logic-based belief-desire-intention models and probabilistic Markov decision process models, has generated a variety of techniques for multi–computer agent groups, for both competitive and cooperative settings, yielding a diverse range of successfully deployed systems.[1]

To develop realistic models of human decision-making has required changes to every component of the traditional act-observe-update-decide cycle. AI researchers have developed new models, methods, and agent designs that incorporate reasoning about people for both machine-learning–based systems and logic-based systems. While agents in mixed-agent groups, like those in multi-agent systems generally, might compete, the focus of research has been on settings in which computer agents cooperate or fully collaborate with people in their mixed-agent group. These changes have raised not just new technical challenges, but also paramount ethical and societal-impact challenges.

Research on AI models of collaboration laid the foundations for reasoning about people as participants in mixed-agent groups.[2] These models stipulate as a defining characteristic of collaboration that all team participants share an overarching goal. The models provide theoretical frameworks for representing and reasoning about the mental state and communication requirements

for successful teamwork activities. Related work in AI and cognitive science specifies the obligations collective intentionality entails.[3]

Significant recent research focuses on settings in which computer agents need to coordinate with people, but absent an overall teamwork goal. For instance, an autonomous vehicle does not share an overarching destination goal with other drivers or pedestrians it encounters; the autonomous vehicle and others on the road do not make a team. Aspects of the early frameworks are also relevant to such settings, as is early work specifying the roles of social norms in coordinating behavior of multiple agents without a shared goal who nonetheless need to avoid conflict.[4] Key insights of this early work include establishing the need to explicitly design agents for collaboration, showing that the requisite capabilities could not be patched on, and the need for revisions of plan representations and decision-making algorithms for them.[5]

Subsequent work in both logical and machine learning paradigms has demonstrated the benefits of developing algorithms that consider the combined performances of people and agents rather than focusing on the autonomous performance of a computer agent in isolation.[6] For example, methods that optimize for agents to complement human capabilities or to balance human and computer agent preferences outperformed individual human and computer performances.[7] Other work deploys cross-training to improve human-robot team performance.[8] A consensus is emerging from this research of the importance of bringing insights from the social sciences to bear in designing agents for working with people.[9]

The advent of large-scale Internet activities – from citizen science to online learning and question-and-answer sites – has provided researchers with significantly more data than ever before about people's behaviors and preferences, creating new technical opportunities and raising new AI research questions. Not only do people's decision-making processes often not adhere to standard assumptions about optimizing for utility, but these larger-scale settings require computer agents to operate in the "open world," rather than in well-defined, constrained, and therefore more easily specifiable environments ("closed worlds").[10] As a result, agent designs need to accommodate both *scale* – a significant increase in the number of people an agent may work with – and operating "in the wild": that is, in open worlds in which computer agents have only partial information about other agents and much less control. Further challenges arise from the need for computer-agent behaviors and explanations to mesh with people's expectations.[11]

We briefly describe AI researchers' advances on three core computer-agent capabilities that are enabling agents to participate more effectively in mixed-agent groups: 1) *decision-making* about what to do next, considering the potential effects of an agent's actions on other agents' beliefs and decision-making, as well as on the environment; 2) *reasoning* to draw conclusions about the effects of an agent's actions on that environment, including any causal connections; and 3) *learning* from

the effects it observes in the environment and on others' actions. This research has led to paradigmatic shifts in a variety of AI methods and algorithms, as well as to the development of *new representations* of information about agents and the environments in which they act.

New representations of actions, plans, and agent interactions enable agents to reason about their human partners despite having limited information about their beliefs and capabilities. For instance, a digital personal assistant may not know which route a person is taking to get home, and a health care coordination system may need to learn interaction patterns among medical providers as they evolve.

Novel ways of representing task and plan knowledge – for instance, with Ece Kamar, we expanded the SharedPlans specification of teamwork – enable collaboration when an agent does not know which of several plans a person is following.[12] To enable computer agents to reason effectively about information sharing when they lack *a priori* knowledge of other agents' plans (as required by standard information-sharing algorithms), Ofra Amir and colleagues developed a representation of "mutual influence potential" networks for teams that operate over long periods of time (such as project management and health care teams).[13] To address the need for computer-agent collaborators to adapt to their human partners' actions, Stefanos Nikolaidis and colleagues developed a representation for Markov decision processes that evolves through cross-training, and they demonstrated that cross-training outperforms other training regimes.[14]

New methods of decision-making have been designed by AI researchers to reason about social influences on people's behavior in negotiation; to determine when to share information with partners in a group activity; and, for large-scale groups, to identify the best people for certain tasks and to provide incentives for them to contribute to group activities.

For computer agents to negotiate effectively with people, they need to take into account findings in the social sciences that have revealed social influences on people's negotiation strategies. Research incorporating such findings into agent negotiation strategies – by representing social attributes in the decision-making model – has demonstrated the ability of such socially aware agents to reach agreements that benefit all participants. For instance, through empirical investigations, we showed that people's willingness to accept offers is affected by such traits as altruism and selfishness, and that agents incorporating these traits into their negotiation strategies outperform traditional game-theoretic equilibria strategies.[15] Amos Azaria and colleagues improved agent success in advising a person on the best route to a destination by incorporating a model of people's behavior in repeated negotiations.[16] And Arlette van Wissen and colleagues found that although

people trust computer agents as much as other people in negotiations, they treat them less fairly.[17] Agents negotiating for people also need to model their preferences. For example, an agent might assist a consumer in negotiating the best deal for an item available from multiple online sellers who offer similar products at varying prices and characteristics – used or new, full or delayed payment – saving the consumer time and money. If the consumer is price sensitive, the agent could negotiate a lower price while agreeing to make the payment in advance.

To coordinate their activities, participants in mixed-agent groups typically must share information with each other about their activities, environments, and tasks. Decisions about what information to share, and when, are more complicated when computer agents are not privy to important task-related information that people hold. For example, a driver-assist system considering whether to alert the driver to unexpected traffic ahead on a possible route that allowed for a side-trip to a pharmacy may not be certain about the driver's current preferences with respect to making that stop. As a result, it may not know if this traffic situation is on the route the driver is taking and thus whether notifying the driver would be useful or an unnecessary interruption. Information exchanges – whether an unneeded alert or a request for irrelevant information – generate cognitive and communication costs. Research on managing information exchange to avoid overburdening people includes theoretical model development and empirical studies.

With Ece Kamar, we identified the class of "nearly decomposable" settings, in which computer agents need to reason about only that subset of their human partners' actions that interact with the agent's actions.[18] We developed a multi-agent Markov decision process for such settings that enables more efficient inference for interruption management. An empirical study using this method identified factors influencing people's acceptance of an agent's interruptions.

In work on information sharing for team settings in which agents have very limited information about their human partners, Ofra Amir and colleagues developed an algorithm that identifies the information that is most relevant to each team member using the influence potential networks described earlier.[19] The results of a laboratory study using this algorithm demonstrated that information-sharing decisions based on the influence-potential representation yielded higher productivity and lower perceived workload compared with standard human-computer interaction approaches.

In such large-scale settings as disaster response and online forums, the standard multi-agent systems' role assignment problem – the problem of identifying the best agent for a particular task – is more difficult because less information is directly available about (human) participants' capabilities. These settings also introduce a new role-assignment challenge : namely, keeping people engaged.

Methods that integrate behavior prediction into decision-making processes enable inferring people's capabilities from their prior interactions and thus pre-

dicting the best person to assign a task to. Research on engagement includes the use of reinforcement learning to generate motivational messages.[20] The benefits of these approaches have been demonstrated in citizen science applications such as classifying celestial bodies and identifying EEG patterns.[21]

Reasoning and learning are tightly coupled. We discuss them together because new methods developed to jointly learn and use models of people's behavior have been consequential for mixed-agent group settings. Important new reasoning capabilities include 1) methods for predicting people's behavior from data about the actions they have taken in the past, their causal effects, and the outcomes that result; 2) techniques for agents to use advice and feedback from people to learn more effectively; and 3) methods for agents to explain their choices and recommendations well enough that people understand them. For example, to find the sequence of math problems that maximizes students' learning gains, an AI tutor needs to predict their responses to math problems. It also needs to be able to explain its problem choices to students, and possibly their teachers.[22]

Computer agents in mixed-agent groups need to model people's past actions and to predict their likely future actions. Machine learning algorithms face a compatibility-performance trade-off: updating machine learning systems with new data may improve their overall performance, but the updated predictions may decrease trust in the system by individuals for whom the predictions no longer work. To address this problem, Jonathan Martinez and colleagues defined machine learning algorithms that personalize their updates to individual users, which not only yields higher accuracy but also makes models more compatible with people's expectations.[23] They established the efficacy of this approach empirically by comparing it with a baseline method that did not personalize the model's updates.

People "in the wild" also make computer agents' plan recognition – the ability to determine what others are doing and why – more difficult, since they often exhibit complex planning behaviors: they may follow multiple plans, interleave actions from different plans, or perform actions that are redundant, wrong, or arbitrary. Novel plan and goal recognition algorithms have been developed to enable agents to adapt to people's exploratory and error-prone behavior. They use various techniques, including heuristics and approaches that replace predefined libraries of possible plans with generating plans on the fly.[24] To enable agents to support people's understanding of plans of other agents (human and computer) in their groups, researchers have designed new types of visualizations for presenting inferred plans to people in ways that facilitate their understanding of others' plans.[25]

Reinforcement learning algorithms enable agents to learn about their environment and about other agents through exploration and trial and error. Mixed-agent groups introduce a new possibility: algorithms can incorporate guidance and feedback from people who have relevant task expertise or knowledge of the

agent's environment and thus significantly facilitate agent learning. W. Bradley Knox and Peter Stone combined feedback from human teachers, who give positive or negative signals to the agent trainee, with autonomous learning about the environment.[26] Travis Mandel and colleagues augmented a reinforcement algorithm with a method for querying people about the best action to perform.[27] Their empirical studies demonstrated significant improvements to algorithm performance for domains with large numbers of actions. Matthew E. Taylor and colleagues showed that agents could adapt a policy to a new domain more effectively if a person first demonstrates how to act in that domain.[28] In this work, short episodes of human demonstrations led to rapid savings in learning time and policy performance for agents in different robot soccer simulation tasks.

For people to trust agents, the models they use to predict people's behavior not only need to perform well according to machine learning systems' metrics, but also to produce interpretable predictions – their action choices need to make sense to the people with whom they interact.[29] As all applications of AI machine learning methods have this need for "interpretability," a variety of research studies have investigated the design of "interpretable models" as well as ways to measure the interpretability of machine learning models in practice.[30]

The evaluation of multi-agent systems becomes significantly more complicated when an agent group includes people. Testing in the wild – that is, in the actual intended situations of use – may be costly both practically and ethically. In response to this challenge, researchers have developed various testbed systems that enable initial evaluation of effectiveness of computer-agent decision-making algorithms in lab (or lab-like) settings. They enable testing of new methods on intended user populations without such costs, allowing agent designers to better determine responses to agents' decisions as well as to compare the performance of different computational decision-making strategies. Some testbed systems have also been used to gather information about people's decision-making strategies to help improve the performance of learning algorithms.

Colored Trails, one of the first such testbeds, enabled the development of a family of games that facilitated the analysis of decision-making strategies, including negotiation strategies and coalition formation in widely varying settings.[31] The Genius testbed (General Environment for Negotiation with Intelligent multi-purpose Usage Simulation) advances research on bilateral multi-issue negotiation by providing tools for specific negotiation scenarios and negotiator preference profiles and for computing and visualizing optimal solutions.[32] The IAGO testbed (Interactive Arbitration Guide Online) provides a web-based interaction system for two-agent bargaining tasks. It has been used to study the role of affect and deception on negotiation strategies in mixed-agent groups.[33] Both Genius and IAGO testbeds have been used in competitions that compare computational strategies for negotiating with people.[34]

Research and development of computer agents capable of participating effectively in mixed-agent groups raise various ethical issues. Some are inherited from AI generally: for instance, avoiding bias, ensuring privacy, and treating people's data ethically. Others result from the mixed-agent group setting entailing that people and computer agents work together and, in some cases, share decision-making. Further, computer agents may be designed to influence people's behavior, make decisions related to people's futures, and negotiate successfully with people. While the roles computer agents and people assume vary within and across application domains, that people are inherent to the definition of "mixed-agent group" makes addressing particular ethical challenges of the utmost importance. We briefly discuss three challenges mixed-agent group research raises, all of which will require research done in concert with social and behavioral scientists and ethicists. We note that choices among ethical values and setting of norms are responsibilities of the societies in which these agent systems are used. Our discussion of ethical challenges thus presumes norms are established by communities of use, policy-making organizations, governmental bodies, or similar entities external to the research effort.

Challenge 1: Inclusive design and testing. The testing of new mixed-agent group algorithms and systems must involve the full range of people expected to participate in group undertakings with such agents. Further, whether for research or for system development, in designing mixed-agent group agents to align with societal values, designers must consider and engage at all stages of the work with the full spectrum of people with whom these agents are intended to interact. For instance, in the initial design stage, researchers should conduct informative interviews or observations to determine system goals and characteristics appropriate for the intended user population.[35]

Inclusivity generates particular challenges when designing new representations, whether models are explicitly designed or derived by machine learning methods. For instance, when developing new representations of tasks and plans, designers need to engage not only the kinds of people agents are likely to work with on a task, but also the kinds of people potentially affected by agent actions and decisions: for example, in a health care setting, the design of an agent that will work with physicians, nurses, and patients, as well as hospital administrative staff, should include physicians, nurses, and patients in the design cycle.

The need for inclusivity at the design stage also arises in areas of learning and reasoning. For example, when developing models of people's behavior, it is crucial for agents to handle adequately all types of people whose behavior it may need to track.

Challenge 2: Avoiding deception and exploitation. The use of social science factors in negotiation algorithms or for behavior modification (like nudges) may have purposes that engender unethical behavior. Mixed-agent group work on negoti-

ation may raise significant questions if the negotiation algorithm focuses only on improving the computer agent's outcome and deploys deception, rather than balancing good for all parties.[36] Similarly, role assignment in some ride-sharing applications has raised significant questions of deception and exploitation.

For agents in mixed-agent groups to be trustworthy, any use of deceptive strategies must be revealed. Researchers developing and deploying negotiation and behavior modification strategies must explain the rationale for them and make evident the ethical challenges they raise for any system that deploys them in applications and possible mitigations.

Challenge 3: Preventing or mitigating unanticipated uses of models and algorithms. The development of new representations and algorithms (such as for information sharing, role assignment, or behavior modeling) is typically driven by an intended application. The resulting learned representations and models may not be appropriate for other applications or may have consequences that were not anticipated when design was focused on the initial intended application. For example, a ride-sharing company might decide to adopt one of the "motivational" algorithms developed in the context of citizen science to attempt to keep drivers working when the system predicts they are close to quitting for the day. While there may be no serious downsides to encouraging someone to continue working on a science project despite being tired, there can be serious consequences from drivers working when fatigued. In some cases, the technology may be sufficiently unreliable or human oversight may be sufficiently inadequate that the unanticipated use should not be allowed. Researchers, system designers, and developers all bear responsibility for preventing the misuse of these technologies.

As mixed-agent groups become the norm in ever more multi-agent domains, advances in multi-agent systems research provide foundations for developing computer agents able to be effective partners in such settings. This work has also revealed a variety of new research challenges and raised important questions of ethical and societal impact.

For these reasons and others, successes in laboratory settings have not yet been translated into deployed systems on a large scale. The inadequacies of automated call centers and the difficulties Amazon fulfillment center workers have experienced working with robots illustrate the problems that arise when computer agents' activities do not mesh well with their human coworkers'. Perhaps the greatest challenge of developing computer agents technically and ethically adequate for participation in mixed-agent group undertakings is to fully recognize the sociotechnical nature of such activities. This recognition should lead not only to different kinds of algorithms, but also to processes for system development and deployment that take account of human capabilities, societal factors, and human-computer interaction design principles.

These challenges do not belong to research alone. If AI systems are to function effectively and safely for people and the societies in which they live, they require attention through the full pipeline from design through development, testing, and deployment. Addressing these challenges is all the more important given the recent broad range of national-level calls for developing effective methods for human-centered AI and for human-AI collaborations.

ABOUT THE AUTHORS

Kobi Gal is an Associate Professor in the Department of Software and Information Systems Engineering at the Ben-Gurion University of the Negev, and a Reader in the School of Informatics at the University of Edinburgh. His contributions to artificial intelligence include novel representations and algorithms for autonomous decision-making in heterogeneous groups comprising people and computational agents. They have been published in various highly refereed venues in artificial intelligence and in the learning and cognitive sciences. Among other awards, Gal is the recipient of the Wolf Foundation's Krill Prize for Israeli scientists and a Marie Curie European Union International Fellowship. He is a Senior Member of the Association for the Advancement of Artificial Intelligence.

Barbara J. Grosz, a Fellow of the American Academy since 2004, is the Higgins Research Professor of Natural Sciences at the John A. Paulson School of Engineering and Applied Sciences at Harvard University. Her contributions to the field of artificial intelligence include fundamental advances in natural language dialogue processing and in theories of multi-agent collaboration and their application to human-computer interaction, as well as innovative uses of models developed in this research to improve health care coordination and science education. She cofounded Harvard's Embedded Ethics program, which integrates teaching of ethical reasoning into core computer science courses. Grosz is also known for her role in the establishment and leadership of interdisciplinary institutions and for her contributions to the advancement of women in science. She is a member of the American Philosophical Society and the National Academy of Engineering, an elected fellow of several scientific societies, and recipient of the 2009 ACM/AAAI Allen Newell Award, the 2015 IJCAI Award for Research Excellence, and the 2017 Association for Computational Linguistics Lifetime Achievement Award.

ENDNOTES

[1] David C. Parkes and Michael P. Wellman, "Economic Reasoning and Artificial Intelligence," *Science* 349 (6245) (2015): 267–272; and Michael Wooldridge, *An Introduction to Multiagent Systems* (Hoboken, N.J.: John Wiley & Sons, 2009).

[2] Barbara J. Grosz and Sarit Kraus, "Collaborative Plans for Complex Group Action," *Artificial Intelligence* 86 (2) (1996): 269–357; Hector J. Levesque, Philip R. Cohen, and José H. T. Nunes, "On Acting Together," in *Proceedings of the Eighth National Conference on Artificial Intelligence* (Menlo Park, Calif.: AAAI Press, 1990); and David Kinny, Elizabeth Sonenberg, Magnus Ljungberg, et al., "Planned Team Activity," in *Artificial Social Systems*, ed. Cristiano Castelfranchi and Eric Werner (New York: Springer, 1992).

[3] Rosaria Conte and Cristiano Castelfranchi, *Cognitive and Social Action* (London: UCL Press, 1995); and Barbara J. Grosz and Luke Hunsberger, "The Dynamics of Intention in Collaborative Activity," *Cognitive Systems Research* 7 (2–3) (2006): 259–272.

[4] Yoav Shoham and Moshe Tennenholtz, "On the Emergence of Social Conventions: Modeling, Analysis, and Simulations," *Artificial Intelligence* 94 (1–2) (1997): 139–166.

[5] Barbara J. Grosz and Candace L. Sidner, "Plans for Discourse," in *Intentions in Communication*, ed. Philip R. Cohen, Jerry Morgan, and Martha E. Pollack (Cambridge, Mass.: MIT Press, 1990); and Milind Tambe, "Towards Flexible Teamwork," *Journal of Artificial Intelligence Research* 7 (1997): 83–124.

[6] Sarvapali D. Ramchurn, Trung Dong Huynh, Feng Wu, et al., "A Disaster Response System Based on Human-Agent Collectives," *Journal of Artificial Intelligence Research* 57 (2016): 661–708; and Matthew Johnson, Jeffrey Mark Bradshaw, Paul J. Feltovich, and Catholijn M. Jonker, "Coactive Design: Designing Support for Interdependence in Joint Activity," *Journal of Human-Robot Interaction* 3 (1) (2014): 43–69.

[7] Bryan Wilder, Eric Horvitz, and Ece Kamar, "Learning to Complement Humans," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, ed. Christian Bessiere (Santa Clara, Calif.: International Joint Conferences on Artificial Intelligence Organization, 2020), 1526–1533; and Amos Azaria, Zinovi Rabinovich, Sarit Kraus, et al., "Strategic Advice Provision in Repeated Human-Agent Interactions," *Autonomous Agents and Multi-Agent Systems* 30 (1) (2016): 4–29.

[8] Stefanos Nikolaidis, Przemyslaw Lasota, Ramya Ramakrishnan, and Julie Shah, "Improved Human-Robot Team Performance through Cross-Training, an Approach Inspired by Human Team Training Practices," *The International Journal of Robotics Research* 34 (14) (2015): 1711–1730.

[9] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, and Josh Bongardet, "Machine Behaviour," *Nature* 568 (7753) (2019): 477–486; and Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere, in "Explainable Reinforcement Learning through a Causal Lens," *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence* (Palo Alto, Calif.: AAAI Press, 2020).

[10] Amos Tversky and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science* 185 (4157) (1974): 1124–1131.

[11] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati, "Plan Explanations as Model Reconciliation: Moving Beyond *Explanation as Soliloquy*," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17* (Santa Clara, Calif.: International Joint Conferences on Artificial Intelligence Organization, 2017), 156–163.

[12] Ece Kamar, Ya'akov Gal, and Barbara J. Grosz, "Incorporating Helpful Behavior into Collaborative Planning," in *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, ed. Ryszard Kowalczyk, Quoc Bao Vo, Zakaria Maamar, and Michael Huhns (Budapest: International Foundation for Autonomous Agents and Multiagent Systems, 2009); and Grosz and Kraus, "Collaborative Plans for Complex Group Action."

[13] Ofra Amir, Barbara J. Grosz, Krzysztof Z. Gajos, and Limor Gultchin, "Personalized Change Awareness: Reducing Information Overload in Loosely-Coupled Teamwork," *Artificial Intelligence* 275 (2019): 204–233.

[14] Nikolaidis et al., "Improved Human-Robot Team Performance."

[15] Ya'akov Gal, Barbara Grosz, Sarit Kraus, et al., "Agent Decision-Making in Open Mixed Networks," *Artificial Intelligence* 174 (18) (2010): 1460–1480.

[16] Azaria et al., "Strategic Advice Provision in Repeated Human Agent Interactions."

[17] Arlette van Wissen, Ya'akov Gal, Bart Kamphorst, and Virginia Dignum, "Human-Agent Teamwork in Dynamic Environments," *Computers in Human Behavior* 28 (1) (2012): 23–33.

[18] Ece Kamar, Ya'akov Kobi Gal, and Barbara J. Grosz, "Modeling Information Exchange Opportunities for Effective Human-Computer Teamwork," *Artificial Intelligence* 195 (2013).

[19] Amir et al., "Personalized Change Awareness."

[20] Avi Segal, Kobi Gal, Ece Kamar, et al., "Optimizing Interventions via Offline Policy Evaluation: Studies in Citizen Science," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (Palo Alto, Calif.: AAAI Press, 2018).

[21] Ece Kamar, Severin Hacker, and Eric Horvitz, "Combining Human and Machine Intelligence in Large-Scale Crowdsourcing," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)* (Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems, 2012); and Shengying Pan, Kate Larson, Josh Bradshaw, and Edith Law, "Dynamic Task Allocation Algorithm for Hiring Workers that Learn," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-16*, ed. Subbarao Kambhampati (Santa Clara, Calif.: International Joint Conferences on Artificial Intelligence Organization, 2016).

[22] Avi Segal, Yossi Ben David, Joseph Jay Williams, et al., "Combining Difficulty Ranking with Multi-Armed Bandits to Sequence Educational Content," in *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27 – 30, 2018, Proceedings, Part II*, ed. Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. Ulrich Hoppe, et al. (New York: Springer, 2018), 317–321.

[23] Jonathan Martinez, Kobi Gal, Ece Kamar, and Levi H. S. Lelis, "Personalization in Human-AI Teams: Improving the Compatibility-Accuracy Tradeoff," in *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence* (Palo Alto, Calif.: AAAI Press, 2021).

[24] Reuth Mirsky, Ya'akov Gal, and Stuart M. Shieber. "CRADLE: An Online Plan Recognition Algorithm for Exploratory Domains," *ACM Transactions on Intelligent Systems and Technology* 8 (3) (2017): 1–22; Chakraborti et al., "Plan Explanations as Model Reconciliation"; and Mor Vered, Gal A. Kaminka, and Sivan Biham, "Online Goal Recognition through Mirroring: Humans and Agents," presented at the Fourth Annual Conference on Advances in Cognitive Systems, Evanston, Illinois, June 23–26, 2016.

[25] Ofra Amir and Ya'akov Gal, "Plan Recognition and Visualization in Exploratory Learning Environments," *ACM Transactions on Interactive Intelligent Systems* 3 (3) (2013) : 1–23 ; and Nicholas Hoernle, Kobi Gal, Barbara Grosz, and Leilah Lyons, "Interpretable Models for Understanding Immersive Simulations," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (Santa Clara, Calif. : International Joint Conferences on Artificial Intelligence Organization, 2020).

[26] W. Bradley Knox and Peter Stone, "Combining Manual Feedback with Subsequent Reward Signals for Reinforcement Learning," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, vol. 1 (Toronto : International Foundation for Autonomous Agents and Multiagent Systems, 2010), 5–12.

[27] Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popovic, "Where to Add Actions in Human-in-the-Loop Reinforcement Learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (Palo Alto, Calif. : AAAI Press, 2017).

[28] Matthew E. Taylor, Halit Bener Suay, and Sonia Chernova, "Integrating Reinforcement Learning with Human Demonstrations of Varying Ability," *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems* (Taipei : International Foundation for Autonomous Agents and Multiagent Systems, 2011).

[29] Avi Rosenfeld and Ariella Richardson, "Explainability in Human-Agent Systems," *Autonomous Agents and Multi-Agent Systems* 33 (3) (2019) ; and Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach, "Understanding the Effect of Accuracy on Trust in Machine Learning Models," in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (New York : Association for Computing Machinery, 2019).

[30] Isaac Lage, Andrew Slavin Ross, Been Kim, et al., "Human-in-the-Loop Interpretability Prior," *Advances in Neural Information Processing Systems* 31 (2018) ; and Hoernle et al., "Interpretable Models for Understanding Immersive Simulations."

[31] Gal et al., "Agent Decision-Making in Open Mixed Networks" ; and van Wissen et al., "Human-Agent Teamwork in Dynamic Environments."

[32] Raz Lin, Sarit Kraus, Tim Baarslag, et al., "Genius : An Integrated Environment for Supporting the Design of Generic Automated Negotiators," *Computational Intelligence* 30 (1) (2014).

[33] Johnathan Mell, Gale Lucas, Sharon Mozgai, and Jonathan Gratch, "The Effects of Experience on Deception in Human-Agent Negotiation," *Journal of Artificial Intelligence Research* 68 (2020) : 633–660.

[34] See Automated Negotiating Agents Competition (ANAC), http ://ii.tudelft.nl/nego/node/7 (accessed March 1, 2022).

[35] Ofra Amir, Barbara J. Grosz, Krzysztof Z. Gajos, et al., "From Care Plans to Care Coordination : Opportunities for Computer Support of Teamwork in Complex Healthcare," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York : Association for Computing Machinery, 2015), 1419–1428 ; and Mary L. Gray and Siddharth Suri, *Ghost Work : How to Stop Silicon Valley from Building a New Global Underclass* (New York : Eamon Dolan Books, 2019).

[36] Mell et al., "The Effects of Experience on Deception in Human-Agent Negotiation" ; and Azaria, "Strategic Advice Provision in Repeated Human Agent Interactions."