DOI: 10.1111/nyas.14783

REVIEW

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

Predictive models for human–AI nexus in group decision making

Omid Askarisichani[1] | Francesco Bullo[2,3] | Noah E. Friedkin[3,4] | Ambuj K. Singh[1]

[1]Department of Computer Science, University of California, Santa Barbara, California, USA

[2]Department of Mechanical Engineering, University of California, Santa Barbara, California, USA

[3]Center for Control, Dynamical Systems and Computation, University of California, Santa Barbara, California, USA

[4]Department of Sociology, University of California, Santa Barbara, California, USA

Correspondence
Ambuj K. Singh, Department of Computer Science, University of California, Santa Barbara, CA 93106, USA.
Email: ambuj@cs.ucsb.edu

Funding information
Army Research Office

Abstract

Machine learning (ML) and artificial intelligence (AI) have had a profound impact on our lives. Domains like health and learning are naturally helped by human–AI interactions and decision making. In these areas, as ML algorithms prove their value in making important decisions, humans add their distinctive expertise and judgment on social and interpersonal issues that need to be considered in tandem with algorithmic inputs of information. Some questions naturally arise. What rules and regulations should be invoked on the employment of AI, and what protocols should be in place to evaluate available AI resources? What are the forms of effective communication and coordination with AI that best promote effective human–AI teamwork? In this review, we highlight factors that we believe are especially important in assembling and managing human–AI decision making in a group setting.

KEYWORDS
decision making, human—AI teams, machine learning

INTRODUCTION

Intelligent agents have become fundamental to everyday life. Examples of such systems include social assistants on mobile devices, pedagogical agents in tutoring systems, social robots collaborating with humans, and multimodal interface agents for smart appliances and environments. Combining state-of-the-art machine learning (ML) and understanding of human behaviors may excite major scientific discoveries at the interface of natural and artificial intelligence (AI).

As noted in the National Science Foundation's 10 Big Ideas,[1] *The Future of Work at the Human–Technology Frontier*: "we have a unique opportunity to actively shape the development and use of technologies to improve the quality of work while also increasing productivity and economic growth in manufacturing and in service sectors such as healthcare and education." While combining a single human with a single AI agent has been explored much in the literature,[2–10] this review concentrates on the group setting of human and AI agents. The intellectual challenges here include integrating group theoretic constructs from the social sciences and AI/ML methods to understand the dynamic behavior of groups with AI involvement. A key barrier in this endeavor is the current limitation in data, models, and theories that explain their dynamic behavior, coordination, and performance. Existing research indicates that group performance cannot be understood by studying the components (individuals and networks) in isolation. It is not simply a sum of individual performance, but ruled by patterns of interactions, influence, and other relationships among group members. Yet, we do not fully understand the dominant sociocognitive processes that determine the dynamic, adaptive, and learning behavior of human–AI groups. In this article, we concentrate on the open problem of theory development on optimal coordination of AI and groups of humans in decision making.

Sociocognitive constructs for decision making in human groups

Over the last decades, scientists have made meaningful headway in understanding collective group behavior of humans. Researchers have examined social processes of groups on single issues and sequences of issues, and have understood the implications of these social

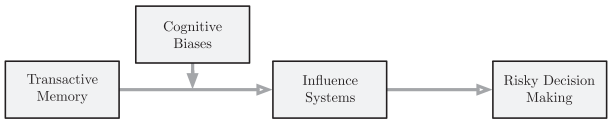

FIGURE 1 We formalize group interactions as the composition of sociocognitive constructs. This review considers the three constructs of transactive memory (Section "Transactive Memory Systems in Human–AI Groups"), cognitive biases (Section "Cognitive Biases in Human–AI Groups"), and influence systems (Section "Influence Systems in Human–AI Groups"), and investigates how these are realized in the task of risky decision making (Section "Risky Decision Making in Human–AI Groups")

processes for group performance on objective measures of performance, and their emergent effects given the characteristics of groups and the actors in them. As shown in Figure 1, this review focuses on three selected sociocognitive constructs to provide a quantitative understanding of group decision-making processes in uncertain environments. Transactive memory systems (TMS)[11] are automatically activated in the appraisal of group members' levels of expertise and potential contributions to tasks. Cognitive biases, heuristic and bounded rational processes augment and interfere with the recognition and correct appraisal[12] of other individuals' skills. Finally, interpersonal influence systems, with various weighted digraph structures of $i \rightarrow j$ arcs of accorded influence, result from the composition of learning, biases, and a number of other antecedent factors and cognitive processes.[13]

Understanding and modeling decision making in groups remains highly complex. For example, research from psychology[14,15] suggests that people process information using dual processes: an explicit (controlled) conscious process and an implicit (automatic) unconscious process. The first process is encoded by analytic algorithms, rules, and reasoning systems, and is amenable to ML modeling. The second implicit automatic system and its interaction with the explicit system is harder to model in humans, and poses considerable challenges for a theoretical understanding of mixed human–AI groups.

Another source of complexity in modeling and understanding decision making in groups has to do with uncertainty. This uncertainty can be separated into two kinds: aleatoric and epistemic.[16,17] Aleatoric uncertainty refers to the notion of randomness (as in coin flipping): the variability in the outcome of an experiment that is due to inherently random effects. Epistemic uncertainty refers to uncertainty due to lack of knowledge of the group environment. This uncertainty can in principle be reduced by a proper recognition of expertise on groups and protocols that reveal explanations on why a fact may be true.

Both aleatoric and epistemic uncertainty require groups to decide under varying amounts of risk and reward under conditions that are not completely rational. The most successful model for explaining risky decision making is prospect theory.[18,19] According to it, individuals make decisions based on the potential value of losses and gains among the set of available options. It proposes that individuals compute an internal evaluation for each prospect that is determined by a value function and a probability weighting function. The value function is S-shaped and asymmetrical, capturing loss aversion. The probability weighting function encodes the hypothesis that individuals over-react to small probability events, but under-react to large probability events. The theory deviates from its rational competitor, expected utility theory,[20] which assumes that people evaluate the outcome of a decision in terms of the expected reward, independent of any cognitive biases (such as risk aversion). Other recent theories explaining individual choices under risk/uncertainty include dynamic decision models,[21] such as dynamic field theory.[22]

Overall, numerous efforts have focused on explaining and modeling group decision making; this is a vast field tackling a multifaceted problem. Motivated by some of our own experimental and analytical work, this review focuses on the above-mentioned key sociocognitive constructs and presents an explicit way of interconnecting them quantitatively in the context of risky decision making.

While we focus only on a few selected sociocognitive constructs, it is important to recognize that other modeling results have also been put forth. For example, collective intelligence refers to a group's ability to produce intelligence and behaviors beyond the individual;[23] in this body of work, human groups can display magnified cognitive capacity and unique cognitive abilities that emerge from the interaction between the group members.[24] Theory of mind[25–29] broadly refers to humans' ability to represent the mental states of others, including their desires, beliefs, and intentions. Finally, the seminal work by Reference 30 and the influential References 31 and 32 have introduced and popularized the concept of group mental model to focus on the overlapping knowledge and shared cognition among group members. Broadly speaking, the theory of group mental models focuses on a broader content domain than the theory of TMS.

Sociocognitive constructs for decision making in human–AI groups

Humans and AI are clearly different in their cognitive and processing capabilities.[33] Groups with AI involvement should be designed so that the raw computational and search power of computers for state-space reduction can be combined with group inductive reasoning, especially in uncertain environments. What is the optimal group-AI design for a given decision? This is a question pervading all kinds of groups that oriented to specific types of issues. Taxonomies and ontologies for characterizing group decision making have been defined[34–36] in order to investigate the optimal composition of groups. The behavior of groups with AI involvement must be observable and predictable. This is challenging in complex uncertain environments. While groups often adopt satisficing strategies,[37,38] AI utilizes search space reduction strategies, such as limited look-ahead, constraint relaxation, and heuristics. Both humans and AI are subject to bias and faulty information: humans by their members' beliefs and AI by the available data and training protocols. Since observability and predictability have ramifications on the level of trust,[39] groups with AI involvement must have confidence that the behavior of their AI is consistent with an

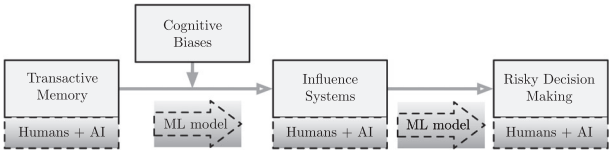

FIGURE 2 Our simplified representation in Figure 1 is now augmented to include (1) the presence of AI agents in the group, and (2) the adoption of ML models to predict the behavior of a human+AI group. In this review, we confine our ML modeling efforts to the specific tasks of predicting the emergence of influence systems (the left big arrow) and the prediction of decisions under multiple risk/reward scenarios (the right big arrow). Note that ML techniques can be utilized broadly in other tasks in the group environment, such as deciding optimal group composition and the design of interventions

acceptable common ground whatever the displayed initial beliefs of the group's members might be.[40–45]

An excellent survey of the broad ideas on human–AI autonomy teaming has been put forth in a recent paper.[46] The authors discuss different levels of autonomy and summarize the recent literature. They note that most of the papers have concentrated on dyads (one human and one AI agent in a group). Furthermore, the authors found a "relatively haphazard collections of independent and dependent variables considered in relatively narrow (rather than integrative) empirical studies"; likely because we are in the relatively early days of studying and building an understanding of such groups. A related review[47] discusses the implications of leadership in human–AI groups. Reference 48 discusses the mechanisms for enhancing the performance of human–AI groups and outlines the critical scientific questions that must be addressed to enable this vision. Reference 49 outlines a research agenda for exploring the potential risks and benefits of human–AI groups; the agenda is separated into three design areas: machine artifact, collaboration, and institution, and augmented with a set of possibilities that have significant potential for benefit or harm.

There is evidence that the same sociocognitive constructs relevant in human groups play an important role when the group is composed of multiple humans and AI agents; we illustrate this concept by introducing AI agents in Figure 2. As illustrated there, this review focuses on the adoption of ML models to quantitatively predict the behavior of a human+AI group on two specific tasks: the emergence of influence systems and the prediction of which decisions a group will reach among options with varying risk/reward tradeoffs. However, much work[48] remains to be done in order to understand and quantify how these constructs are affected by the introduction of AI agents in human groups. How to integrate AI into human groups in order to produce cognitive abilities that go beyond the individual or the group of humans, and allow meaningful interactions is an important question.[50]

Group constructs have also been proposed in ML—experts, weak learners, crowd-sourced workers—to achieve goals that no single individual can accomplish on its own. In the case of boosting,[51] one can obtain a "strong learner" that is able to predict arbitrarily accurately based on an ensemble of "weak learners" whose predictions are slightly better than random guessing. In the case of "learning from expert advice,"[52] an algorithm works with a group of K arbitrary "experts" who give daily "stock predictions" and who perform nearly as well as the "expert" that has the best "track record" at any given time. It is an iterative game in which in each iteration the "player" must make a decision and the experts with the best track record may change over time. The "Multi-armed Bandits" (MAB) problems[53] can be thought of as a variant of the problem of "learning from expert advice" in which a "player" can only observe the payoff of the "expert" at each iteration. In the case of "crowd sourcing,"[54] an algorithm aggregates the inputs of a large group of unreliable "participants," evaluates each "participant," and then infers the ground truth.

Some recent human–AI group experiments

To explore sociocognitive constructs for decision making in human and human–AI groups (Sections "Socio-cognitive constructs for decision making in human groups" and "Socio-cognitive constructs for decision making in human–AI group"), we carried out a number of controlled experiments. In the context of expertise recognition, the experiments documented in References 55 and 56 investigate how a group answers a sequence of intellective questions with the help of the Platform for Online Group Studies;[57] here is a synopsis of these experiments. The group's task is to answer intellective questions from different categories, such as history, science and technology, and so on. Every group consists of four individuals and each individual has access to their own AI agent. Each question is answered in four timed phases. In the first phase, every group member records their individual response for the question. In the second phase, the response of every group member is displayed on the screen and a chat plugin (the only communication channel) is enabled for communication. Group members then record their choices and decide whether or not to use an AI agent (and which AI agent to use) in an optional third phase. In the fourth and final phase, the group submits an answer. The correct answer to each question is displayed at the end of each round. Note that if the group has relied on the incorrect response of an AI agent, then the group's trust in that agent (and their other available AI agents) may be eroded. After every few questions, subjects are asked to record the influence of their teammates in their decision-making process as a percent value, such that the sum of all values adds up to 100. Every subject assumes they are given a total of 100 chips and instructed to distribute these chips to indicate the relative importance of each member in determining their own final answers on past problems. The number of chips that subjects allocate to themselves indicate the extent to which their final answer was not affected by the conversation. After normalization, the self-reported interpersonal influences form a row-stochastic appraisal/influence matrix. This approach to measuring interpersonal appraisals and influence is standard in the study of influence systems, for example, see the classic work documented in Reference 13 as well as the recent studies.[37,58–60] Additionally, the group members are asked to rate the accuracy of all four AI agents based on their interactions with them. Finally, the platform collects a log of all the instant messages, including time of message and content during every

question, the individual and group answers, and the self-reported influence matrices.

This experimental protocol and attending software implementation allow us to study (1) expertise recognition and learning phenomena, (2) cognitive biases in human–AI groups, (3) influence systems resulting from learning and biases, and (4) risky decision-making process in human–AI groups. We elaborate on each of these four aspects in the next four sections.

The rest of this review is organized as follows. Section "Transactive Memory Systems in Human–AI Groups" considers the theory on TMS in which the distributed knowledge of the members of a group is made apparent and efficiently exploited. Section "Cognitive Biases in Human–AI Groups" considers the effects and mitigation of cognitive biases. Section "Influence Systems in Human–AI Groups" considers the emergence of influence systems. Section "Risky Decision Making in Human–AI Groups" considers the integration of the above three constructs into group decision making. We end with a brief discussion in Section "Conclusions and Future Work."

TMS IN HUMAN–AI GROUPS

TMS is a conceptual model of group cognition, learning, and performance. This model originates in the seminal work by Reference 11 and is by now well-established in organization science, for example, see the influential References 61–65. From Reference 63, TMS theory models how members of long-tenured groups rely upon one another to obtain, process, and communicate information from specialized knowledge domains; this theory is used to understand the functioning of specialized teams in organizations, such as consulting teams, product development teams, research teams, and ad hoc project teams. In other words, a TMS is a collective "memory" system that emerges in groups engaged in tasks and captures how the distributed knowledge of the members of a group is made apparent, appraised, and efficiently exploited by every member of the group. Empirical research across a range of group types and settings[64,66,67] demonstrates a strong positive relationship between the development of a TMS and group performance.

A key question in TMS theory is how do individuals and AI agents estimate expertise levels of each other in order to rationally assign influence in the decision-making process. Naturally, a reliable appraisals of expertise may be obtained from a sequence of issues, under the assumption that expertise is stationary, in which case expertise can be estimated from the accuracy/success of prior predictions/decisions. An elaboration of TMS includes appraisals of the usefulness of one or more AI-ML resources in decision making on a particular issue. In general, TMS systems provide a basis for assigning more weight to some members than others,[68] and more weight to some AI-ML algorithms than others. Note that trustworthy ML algorithms may be employed in the selection of members for a decision-making group.

Our work on TMS systems includes empirical and theoretical contributions. Our recent experiments on memory-based intellective tasks, as summarized in Section "Some recent human–AI group experiments" and documented in References 55 and 56, provide novel evidence about interpersonal appraisals, memory systems, and social influence in groups. We found empirical evidence for longstanding theories of TMS and confidence heuristics, regarding the origins of social influence and group performance. Specifically, we quantified how, along an issue sequence with feedback, individuals with higher expertise and social confidence are accorded higher interpersonal influence. We modeled how higher-performing individuals better recognize experts in their group, whereas lower-performing individuals assign more uniform evaluations and influence using a "central tendency/reversion to the mean" bias (more about these concepts in Section "Cognitive Biases in Human–AI Groups"). On the theoretical side, building on early simulation-based computational models,[69–71] we have proposed collective learning models for human groups that explain how interpersonal appraisals evolve when individuals have access to a performance signal; this work[72,73] is documented in Figure 3. These models are the first quantitative multiagent mathematical models for TMS. In these models, interpersonal appraisals and workload changes occur simultaneously: each member elaborates personal appraisals of neighboring members based on the performance exhibited on previous tasks, while the workload is redistributed based on the current appraisal estimates. We establish rigorous results characterizing the ability (or the inability) of the group to correctly learn each other's expertise and thus converge to an allocation maximizing the group's performance.

We conclude this discussion about TMS in human–AI groups by reviewing some open questions in this area. Broadly speaking, despite all theoretical work so far, it remains valuable to investigate whether the formation of a TMS can be monitored and whether TMS is a valid predictor of high performance of the human–AI group. Accordingly, it would be valuable to collect rich data sets on the emergence and evolution of TMS in human–AI groups. Such data would allow us to develop a detailed quantitative understanding of how accurate and shared assessments are achieved and how they operate to elevate group performance. Second, it would be important to design software agents that can (1) monitor the real-time development of a TMS, based on data on the communication among the individuals and the decisions taken by the group, and (2) intervene in appropriate ways to facilitate the learning process. Such software agents are potentially very useful in practical applications.

COGNITIVE BIASES IN HUMAN–AI GROUPS

Psychological research has established that human decision making is fundamentally based on cognitive biases and heuristics. Cognitive biases and heuristics are ways for the human brain to quickly respond, without having to recall and elaborate all relevant evidence. For example, *implicit confirmation*[a] biases are widely established and irrational; for example, see the famous study by MacNell *et al.*[74] on gender bias in teaching evaluations. The *overconfidence effect*[b] was originally investigated by Oskamp,[75] and it is now recognized as one of the most "prevalent" and "potentially catastrophic" problems in decision making

[a] *Confirmation bias*: The tendency to favor information that confirms previously held beliefs or, similarly, to believe previously learned misinformation even after it has been corrected.
[b] *Overconfidence bias*: The tendency for a person's subjective confidence in his or her judgments to be reliably greater than the objective accuracy of those judgments.

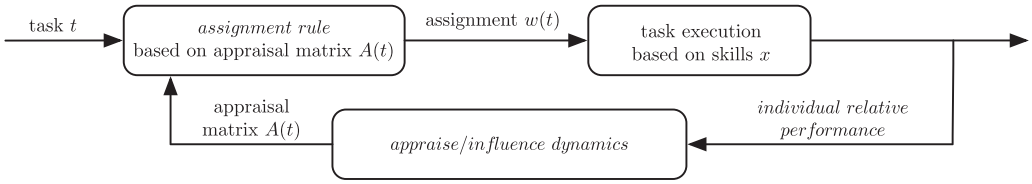

FIGURE 3 In References 72 and 73, we propose mathematical models of group learning in groups of individuals who collectively execute a sequence of tasks and whose actions are determined by individual skill levels and networks of interpersonal appraisals and influence. As emphasized by italics in the figure, mathematical rules need to be specified for how (1) an interpersonal appraisal matrix *A* determines the work assignments *w*, (2) the task performance of the individuals depends upon their latent skills *x* and their assignment *w*, and (3) the feedback performance signal affects the appraisal matrix *A*. Our models propose low-complexity idealized mechanistic representations of interpersonal appraisal, work assignment, learning of interpersonal skills, and influence systems. The mathematical models allow us to identify conditions under which rational and optimal group behavior arises (or does not arise) along the task sequence

(Reference 76, Chapter 19). Related to overconfidence is the so-called *confidence heuristics*, which postulates an expertise–confidence link and a confidence–persuasion link; game theoretical models are given by Reference 77 and empirical evidence is established by References 78 and 79. Finally, the classic study by Kruger and Dunning[80] establishes how difficulties in recognizing one's own incompetence lead[c] to inflated self-assessments. Despite controversies and interpretation challenges, cognitive biases are widespread and their study is an area of active research, for example, see the recent edited handbook.[81] It is important to note that cognitive biases and heuristics are not only sources of errors in judgment, but rather they may arise as bounded rational deviations from logical thought[82] and may lead to "useful attitudes or behavior."[83] Reference 12 gives a compelling description of "ecological rationality"; for example, the *recognition heuristic*[d] is efficient and optimal under certain assumptions.

At the group level, a number of factors may affect decision making. Individuals' cognitive biases may be amplified or attenuated. Moreover, there exist social cognitive biases that affect groups so that their deliberation may be dysfunctional or suboptimal because of poor communication, irrational reasoning, or interpersonal influence processes. The *groupthink phenomenon* described by Janis[84] is a widely known theory of how a bias toward social conformity and cohesion elevates the risk of ill-considered decisions. The *central tendency bias*[85] is the tendency of low-performing individuals to provide evaluations with low differentiation—this phenomenon is akin to a social Dunning–Kruger effect, whereby unskilled individuals not only overestimate their own ability, but also fail to recognize different levels of ability in others. Inaccurate evaluations in turn elevate the risks of suboptimal decisions. This phenomenon is also understandable in the context of social comparison theory.[86,87] The influential work by Golub and Jackson[88] proposes a "naive learning model" explaining mathematically how social influence systems may decrease the decision-making accuracy of a group based on the existence of prominent individuals and information cascades. Note that, while Reference 88 focuses on large populations, similar concepts related to biased influence centralities apply to small group decision making. A recent empirical and theoretical work on the negative influence of information cascades is given in Reference 89. In simple intellective tasks (such as memory or estimation tasks), a rational strategy is arguably to adopt expertise (e.g., measured as the rate of correct answers reached by the individual) as the main driver of accorded interpersonal influence and, therefore, accorded social power.

Experiments in the literature[90] and in our own laboratory, as summarized in Section "Some recent human–AI group experiments," demonstrate that other cognitive processes affect and potentially distort decision making in intellective tasks. Specifically, in our experiments (Section "Some recent human–AI group experiments"), we found statistically significant evidence in support of the following hypotheses:

H1 Individuals with higher expertise are accorded higher interpersonal influence from the group. [This effect is consistent with expertise-based TMS and influence systems.]

H2 Individuals with higher confidence are accorded higher interpersonal influence from the group. [This effect is consistent with overconfidence bias.]

H3 Individuals with lower expertise have diminished ability to recognize experts in the group. [This effect is consistent with a central tendency bias and a social Dunning–Kruger effect.]

While the first effect describes a correct learning process, the second and third effects lead the group to inaccurate performance evaluations, inaccurate accorded influence, and, ultimately, to performance deterioration.

Interestingly, in our experiments[56] on human–AI groups (discussed in detail in Section "Risky Decision Making in Human–AI Groups"), we found evidence of a *risk-aversion bias*[19,91] in the exploration-versus-exploitation tradeoff: individuals were risk averse and did not sufficiently explore the abilities of the AIs that they invoked. This may be a form of *pseudo-certainty effect*,[e] namely, the tendency[19] to make risk-averse choices if the expected outcome is positive. Consistent

[c] *Dunning–Kruger effect*: The tendency for unskilled individuals to overestimate their own ability and the tendency for experts to underestimate their own ability.

[d] *Recognition heuristic*: If one of two objects is recognized and the other is not, then infer that the recognized object has the higher value with respect to the criterion; see Reference 12.

[e] *Pseudo-certainty effect*: The tendency to make risk-averse choices if the expected outcome is positive, but make risk-seeking choices to avoid negative outcomes.

with the literature, we also found evidence that exploration choices can be understood via *prospect theory* (including framing biases) and Bayesian approaches (we discuss these concepts in Section "Risky Decision Making in Human–AI Groups"). Such experimental settings are also valuable to study *automation bias*[f] [92–94] and we plan to do so in future studies.

We conclude this discussion about cognitive biases by reviewing some open questions in this area. First, a broad important open question is to identify the leading cognitive processes, heuristics, and biases that underlie the formation of sociocognitive structures in the human–AI group. In turn, these processes and structures will naturally affect the accuracy of human–AI decision making. Second, it remains unclear how to design software agents that help the group overcome individual and social biases, such as overconfidence, central tendency, and conformity pressure. In other words, it would be valuable to design AI agents that can help the group accurately learn individual abilities and achieve agreement despite the presence of cognitive biases.

INFLUENCE SYSTEMS IN HUMAN–AI GROUPS

The seminal work by References 95 and 96 in social psychology catalogs the bases of social influence. Social influence is broadly understood as a "change in the belief, attitude, or behavior of a person (the target of influence) that results from the action of another person (an influencing agent)", and social power is defined as the potential for such influence. French and Raven explain how social power is accorded on the grounds of six possible dimensions: coercive power, reward power, legitimate power, referent power, expertise power, and information power. It cannot be generally assumed that interpersonal influence systems generate faulty or regrettable decisions, and it cannot generally assumed that a group of disagreeing or like-minded experts will settle on the correct or optimal decision. Conflicting positions on scientific issues exist among physicists, economists, indeed, in every field of science. Hence, while groups of experts are a desirable platform, the mitigation of misleading conclusions remains an open and difficult problem.

In the context of group decision making and forecasting, various models and intervention strategies have been proposed in the literature to enhance the performance of decision-making groups. The ground-breaking work at RAND Corporation in the 1960s led to the design of the first "engineered influence system": the Delphi method. Key references include the seminal work by Dalkey and Helmer[97] the elaboration,[98] and an influential survey by Telesford *et al.*[99] The Delphi method is widely studied and accepted as one of the most successful and rigorous design for expert forecasting and rational decision making. Paraphrasing the survey,[99] the Delphi method is a controlled iterative process to encourage a group of individuals (possibly experts in the subject matter) to develop informed opinions about a topic and converge to closer evaluations, possibly consensus. At each iteration:

(i) individuals express an opinion and arguments in favor/against the various alternatives,
(ii) a coordinator anonymizes the various responses, aggregates them in some statistical sense, and shares them within the group, and
(iii) individuals may adjust their opinion in response to the information they receive.

The ultimate result is meant to represent the best possible forecast. The Delphi iterative process has numerous critical features: (1) group members are provided anonymity, (2) information and feedback is shared in a controlled manner, and (3) appropriate statistical analysis techniques are adopted. From Reference 99, these features are engineered to "offset the shortcomings of conventional means of pooling opinions obtained from a group interaction (i.e., influences of dominant individuals, noise, and group pressure for conformity and other spurious effects)." The Delphi process is an attempt to systematize and engineer the opinion dynamics process inside an influence system and to lend it greater objectivity. Numerous variations have been proposed, for example, the wideband Delphi,[100] the wisdom of select crowds,[101] and the resistance to social influence method.[102]

Extinguishing individual cognitive biases is a central concern in the above studies. Approaches to improving the judgment and decision-making abilities of individuals are reviewed in a recent empirical study;[103] these include debiasing training, incentive design, and nudging strategies. Special attention is given to debiasing training, in which individuals are made aware of cognitive biases and their implications. In the context of debiasing training, Reference 104 encourages the consideration of information that is likely to be underweighted in intuitive judgment. For example, Reference 105 suggests training people on statistical reasoning and normative rules of which they may be unaware. Similarly, in the context of training to adopt decision-making strategies, Reference 106 focuses on how to reduce stereotypes (gender stereotypes, in particular). Along with debiasing training, Reference 106 recommends a structured recall strategy in which (1) explicit evaluation criteria are established, (2) specific evidence of positive and negative behavior is recalled, (3) options are rated on each criterion, and (4) only finally a summary evaluation is publicly expressed and obtained by the group.

A general structural theory of social influence with deep connections to the broad area of social psychology, social networks, and network science is described in Reference 13. Here is a synopsis of the central mathematical model in this theory. Starting with the seminal French–Harary–DeGroot[107–109] weighted-averaging opinion update mechanism, the Friedkin–Johnsen[13,110] generalization describes the evolution of the opinion $x_i(k)$ of individual i at discrete time k by

$$x_i(k+1) = (1 - w_{ii}) \sum_{j=1}^{n} w_{ij} x_j(k) + w_{ii} x_i(0),$$

where $x_i(0)$ denotes the initial opinion of individual i, $x_j(k)$ denotes the opinion of all individuals j, and w_{ij} are the interpersonal influence weights that individual i accords to individual j and that satisfy $0 \le w_{ij} \le 1$ and $\sum_{j=1}^{n} w_{ij} = 1$. To the best of our knowledge, this

[f] *Automation bias*: The tendency for humans to favor suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct.

Friedkin–Johnsen formalization is currently the only model on which a sustained line of human-subject experiments has confirmed the model's predictions of opinion changes. This line of experiments includes our early foundational work,[110,111] as well as a sequence of experiments on reflective appraisal and social power in risky issues,[59] problem solving and intellective tasks,[58] resource allocation and bounded rationality,[37] multidimensional attitudes and appraisal spaces,[60] and intellective memory tasks.[55,56] Indeed, this model is consistent with the dynamics we observed in our experiments described in Section "Some recent human–AI group experiments." Overall, this body of work provides a comprehensive influence systems theory, grounded in social psychology.

Specifically, in Reference 55, we study interpersonal influence in small groups of individuals who collectively execute a sequence of intellective tasks. As discussed in Sections "Transactive Memory Systems in Human–AI Groups" and "Cognitive Biases in Human–AI Groups," our experiments provide empirical evidence in support of TMS and cognitive biases theory as well as on their impact on social influence systems.

Inspired by these theories and based upon the three empirically validated hypotheses presented in Section "Cognitive Biases in Human–AI Groups," we propose (1) a cognitive dynamical model that describes the process by which individuals adjust interpersonal influences over time, and (2) a deep neural network model based on a pretrained text embedding model for predicting the influence of individuals. Using message contents, message times, and individual correctness collected during tasks, we are able to accurately predict individuals' self-reported influence over time. Extensive experiments verify the accuracy of the both models compared to baselines. While the neural networks model is the most accurate, the dynamical model is the most interpretable for influence prediction. In summary, these results illustrate how ML models can be used to quantify influence systems arising from TMS theory and cognitive biases theory; in other words, these results instantiate the first "ML model" arrow in Figure 2.

We conclude this discussion about influence systems by reviewing some open questions in this area. The broad open question is how to monitor the process that leads from expertise to interpersonal influence and, ultimately, to social power. First, a key question relates to understanding what cognitive processes and heuristics dominate discussions about intellective issues: does the human–AI group's influence system mitigate or exacerbate the effect of inefficient heuristics? Precisely as we discuss in the previous section about TMS, we hypothesize that influence system formation in the presence of AI agents can be monitored and is a predictor of high performance in the human–AI group. But this hypothesis needs further validation. Second, it is yet unclear if it is possible to design software agents that help the formation of efficient and accurate influence systems. For example, it would be valuable to design a Delphi-style AI moderator that will suggest (or force, by a software redesign) that the group follows a discussion procedure similar to the Delphi method. This would be an important step, as we transition from analysis of human and human–AI groups, to the design of coordination and supervision strategies. After the introduction of a Delphi-style moderator, it would be important to monitor the dynamics of the influence system and, specifically, monitor phenomena, such as reflected appraisal, expertise learning, exploration/exploitation of expert AI agents and decision processes.

RISKY DECISION MAKING IN HUMAN–AI GROUPS

We first recapitulate existing work on risky decision making in human groups and its relationship to influence systems (Figure 1). Starting with the seminal works,[112,113] a line of research has developed on choice dilemmas: these are issues on which individuals decide on the minimum probability of success they require to choose an option with greater rewards and greater chance of failure over an option with smaller rewards and smaller chance of failure. Individuals have heterogeneous initial positions on the minimum chance of success that they require to select the more risky option, and when a group of individuals is considering such issue, a choice shift usually occurs (the mean of a group's settled position differs from the mean of its members initial positions). The type of shift varies. It may be movement toward greater or smaller risk tolerance. Reference 114 shows that such choice shifts depend on the influence system and the relative influence centralities of the group's members. It remains an open question whether additional information, provided by an AI-agent during deliberation, on the probability distributions of success for the risky option (1) importantly affects the emergent relative influence centralities of a group's members and (2) alters individuals' prospect theory S-shaped risk tolerance traits along a sequence of issues.

In a series of experiments, we modeled individuals' risk/reward profiles using prospect theory, the change in risky behaviors when individuals arrive in a group setting, and explanations of the shift using influence systems. We first asked each individual a series of questions (such as whether they prefer gamble 1 vs. gamble 2) to estimate their prospect theory parameters. Then, we engaged them in a group environment and again asked them a series of questions: first before a group discussion and second following a group discussion. In individual settings, we found that prospect theory-based models are more predictive than alternate models based on utility theory,[20] or models that maximize gains or minimize losses. Let IND denote a prospect theory-based model for individuals before they arrive in a group environment. Let PRE denote a prospect theory-based model when the human subjects assembled into groups but prior to a group discussion. Finally, let POST denote a prospect theory-based model when the human subjects assembled into groups and after a group discussion. We found that individuals become less risk-averse and become more sensitive to gain/loss increases in a group environment. Interestingly, these shifts correlate with the initial magnitudes of the parameters.

The prospect theory parameters of IND, PRE, and POST can also be used to compare the similarity of individuals' behaviors. For most

groups, pairwise IND distances were higher than pairwise PRE distances that were themselves higher than pairwise POST distances. This suggests that the behavior of individuals shifts toward consensus in a group setting. We also find that the distance of an individual's behavior between IND and POST correlates with the average influence that the group exerts on the individual, and that this shift occurs mostly during the initial rounds of group discussions.

As discussed in Section "Some recent human–AI group experiments," we designed another series of experiments to understand the predictive power of ML in mixed human–AI groups. In these experiments, groups were asked a sequence of intellective questions (with a verifiable answer) from different domains. The human groups were also assisted by AI agents with heterogeneous accuracy levels. The groups attempted to answer each question without consulting an agent; if they were unsure, they could consult one of the agents and use the obtained answer to provide the final response. Thus, there were two subsequent rounds of decision making: whether to consult an AI agent and next how to incorporate the answer obtained into a final response. In summary, these results illustrate how ML models can be used to quantify risky decision-making strategies arising from influence systems theory; in other words, these results instantiate the second "ML model" arrow in Figure 2.

We proposed four predictive models for what the groups would do based on ML and prospect theory. The first two models capture the appraisal process in a group, while the last two capture the appraisal process as well as prospect theory-based risk/reward tradeoff. The first model, NB (Naive Bayes), captures the accuracy of a human/AI-agent using a beta distribution that is updated at each round (after observing whether it was correct or incorrect) using the Bayes rule. A Naive Bayes assumption is used to integrate the responses of the human/AI-agents. The second model, CENT (centrality), integrates individual responses through an interpersonal influence system. The probability of the group choosing an option is computed as the sum of the eigenvector centrality values of each individual choosing that option. A similar weighting process is used to integrate the group's evaluation of the AI-agents. The third model, PT-NB (prospect theory coupled with Naive Bayes), uses prospect theory to analyze the actions of the group as a set of prospects. The probabilities of success and reward of each prospect are computed as in the model NB. The group chooses among these prospects based on prospect theory parameters of the group (learned through an initial training sequence). The final model, PT-CENT (prospect theory coupled with centrality), again uses prospect theory to analyze the actions of the group as a set of prospects. The probabilities of success and reward of each prospect are computed as in the model CENT. A group chooses among these prospects based on prospect theory.

We found that though appraisal-based models NB and CENT perform adequately in explaining a human–AI group's decision making, the prospect theory- based models PT-NB and PT-CENT are better, implying that modeling the inherent risk in decision making improves the models. Overall, we observed that humans develop accurate interpersonal appraisals but have a difficult time appraising the AI agents. Furthermore, we found that when a group decides to consult an AI agent, it is ultimately over-reliant upon the AI agent. Inaccurate appraisals and over-reliance upon the AI agents clearly lead to poor performance. Human subjects in a human–AI group find it difficult to properly reconcile an incorrect response from the AI agent with potentially correct answers from other group members. A similar observation extends to how a group performs in comparison to its constituent humans and AI agents. The groups exceed individual performances when the inputs come from humans or when the AI agents provide correct responses.

The above findings open up interesting directions for future research. How to incorporate resource limits (as in the number of queries to agents and the time remaining) as a part of the group decision making? The problem now becomes similar to a Markov decision process[115] in which resources can be modeled as part of the current state. The group now needs a policy that integrates the consumption of resources as part of its value/action function. How a group reaches a consensus on such a policy is of utmost interest. Questions to be examined include: How far does a group look ahead? How does a group compute the expected reward in a future state? How do the influence system and the prospect theory parameters affect the choice of actions? It is likely that the modeling of an agent's risk/reward tradeoff through prospect theory parameters and the modeling of an agent's biases through the influence matrix provide the right representation of an agent's beliefs. It is also likely that maximizing the total reward is the "desire" of every agent and as such, there is uniformity in the group's and the individuals' actions. But these hypotheses need to be validated.

The other general research direction is that of active participation by AI agents in decision making. If the reward functions are known for the group, then the AI agents can decide the optimal policy by computing the expected rewards at every future step and backing up to the current state using dynamic programming. An AI agent can intervene during group deliberations by suggesting such a policy. The adoption of such a policy will depend on the influence structure among the group, on how a group appraises itself, and the group's risk/reward profile. However, one key open question is how does the group establish a common reward function for the entire group through composition of the reward functions of the individual members. A mechanistic explanation of this process through theory and empirical validation will be extremely useful.

CONCLUSIONS AND FUTURE WORK

This review has highlighted some fundamental cognitive processes and psychological/algorithmic constructs that provide a possible framework to model, analyze, and design mixed human–AI groups. We have borrowed from existing mechanistic models in social science theory (that explain human groups) and AI models (that explain algorithmic behavior). How and when to combine these dual models to explain decision making in mixed human–AI groups remains a challenge. Future research directions include (1) how to validate these models longitudinally, over long periods of time, for dynamically evolving

human–AI groups, and (2) how to design AI that facilitate group learning processes and the establishment of effective shared mental models.

All of the above proposed designs for group decision making entail strong assumptions that specify the conditions under which they are justified. In general, the empirical evaluation and validation of particular designs of group decision making require data on a population of groups, their within-group networks, their members' displayed initial opinions on issues, their decisions, and the performance consequences of their decisions, under various designed or natural conditions that have constrained or guided each group's process of reaching a consensus decision. Anything less than such a massive enterprise will not serve to advance the reliable evaluation and exploitation of ML and AI components in group decision making. The mere availability of ML and AI tells us nothing unless they are subject to an evaluation of their contributions to group decision making. The usefulness of ML and AI components in group decision making depends entirely on their particular construction, and their construction is based on the human groups that have designed the ML and AI involved. An institutionalized decision-making group may demand the construction of an AI with particular properties.

Human–AI decision-making groups present a set of complex problems related to optimizing group performance. When the issue being considered can be reduced to a matter of gathering information and partitioning facts and fictions, human–AI groups are ideal. However, group decision making becomes challenging when variables related to social justice cannot be ignored. While technology advances, its proper applications are subject to negotiation.

ACKNOWLEDGMENTS

This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under Grant number W911NF-15-1-0577.

AUTHOR CONTRIBUTIONS

A.S., F.B., and N.F. took the lead in writing the paper. O.A. contributed to the models. All authors reviewed the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

PEER REVIEW

The peer review history for this article is available at: https://publons.com/publon/10.1111/nyas.14783

REFERENCES

1. NSF. NSF's 10 big ideas. https://www.nsf.gov/news/special_reports/big_ideas/.
2. Beck, M. R., Scarlata, C., Fortson, L. F., Lintott, C. J., Simmons, B., Galloway, M. A., Willett, K. W., Dickinson, H., Masters, K. L., Marshall, P. J., & Wright, D. (2018). Integrating human and machine intelligence in galaxy morphology classification tasks. *Monthly Notices of the Royal Astronomical Society*, *476*(4), 5516–5534.
3. Bien, N., Rajpurkar, P., Ball, R. L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B. N., Yeom, K. W., Shpanskaya, K., Halabi, S., Zucker, E., Fanton, G., Amanatullah, D. F., Beaulieu, C. F., Riley, G. M., Stewart, R. J., Blankenberg, F. G., Larson, D. B., … Lungren, M. P. (2018). Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Medicine*, *15*(11), e1002699.
4. Chen, J. Y. (2018). Human–autonomy teaming in military settings. *Theoretical Issues in Ergonomics Science*, *19*(3), 255–258.
5. Kamar, E., Hacker, S., & Horvitz, E. (2012). Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1* (pp. 467–474). International Foundation for Autonomous Agents and Multiagent Systems.
6. Kerrigan, G., Smyth, P., & Steyvers, M. (2021). Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems*, *34*.
7. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *Quarterly Journal of Economics*, *133*(1), 237–293.
8. Trouille, L., Lintott, C. J., & Fortson, L. F. (2019). Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human–machine systems. *Proceedings of the National Academy of Sciences*, *116*(6), 1902–1909.
9. Unhelkar, V. V., Lasota, P. A., Tyroller, Q., Buhai, R.-D., Marceau, L., Deml, B., & Shah, J. A. (2018). Human-aware robotic assistant for collaborative assembly: Integrating human motion prediction with planning in time. *IEEE Robotics and Automation Letters*, *3*(3), 2394–2401.
10. Wilson, H. J., & Daugherty, P. R. (2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*, *96*(4), 114–123.
11. Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Goethals (Eds.). *Theories of group behavior* (pp. 185–208). Springer.
12. Gigerenzer, G. (2006). Bounded and rational. In R.J. Stainton, (Ed.), *Contemporary debates in cognitive science* (pp. 115–133). Blackwell.
13. Friedkin, N. E., & Johnsen, E. C. (2011). *Social influence network theory: A sociological examination of small group dynamics*. Cambridge: Cambridge University Press.
14. Chaiken, S., & Trope, Y. (1999). *Dual-process theories in social psychology*. New York: Guilford Press.
15. Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
16. Der Kiureghian, A., & Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Structural Safety*, *31*.
17. Fox, C. R., & Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. In W. Brun, G. Keren, G. Kirkebøen, & H. Montgomery (Eds.). *Perspectives on thinking, judging and decision-making* (pp. 1-14). Universitetsforlaget.
18. Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 363–391.
19. Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.
20. Fishburn, P. C., & Fishburn, C. S. (1970). *Utility theory for decision making*. Operations Research Society of America. Publications in operations research. Wiley.
21. Pleskac, T. J., Diederich, A., & Wallsten, T. S. (2015). Models of decision making under risk and uncertainty. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.). *The Oxford handbook of computational and mathematical psychology* (pp. 209–231). New York: Oxford University Press.
22. Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432–459.
23. Woolley, A. W., Chabris, C. F., Pentland, A., Hasnmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*, 686–688.

24. Amelkin, V., Askarisichani, O., Kim, Y. J., Malone, T. W., & Singh, A. K. (2018). Dynamics of collective performance in collaboration networks. *PLoS One*, *13*(10), e0204547.
25. Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.
26. Cuzzolin, F., Morelli, A., Cîrstea, B., & Sahakian, B. J. (2020). Knowing me, knowing you: Theory of mind in AI. *Psychological Medicine*, *50*(7), 1057–1061.
27. Oguntola, I., Hughes, D., & Sycara, K. P. (2021). Deep interpretable models of theory of mind for human–agent teaming. *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. https://doi.org/10.1109/RO-MAN50785.2021.9515505
28. Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.
29. Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. In *International Conference on Machine Learning* (pp. 4215–4224).
30. Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management*, *20*.
31. Mohammed, S., & Dumville, B. C. (2001). Team mental models in a team knowledge framework: Expanding theory and measurement across disciplinary boundaries. *Journal of Organizational Behavior*, *22*(2), 89–106.
32. Mohammed, S., Ferzandi, L., & Hamilton, K. (2010). Metaphor no more: A 15-year review of the team mental model construct. *Journal of Management*, *36*(4), 876–910.
33. Cummings, M. (2014). Man versus machine or man + machine? *IEEE Intelligent Systems*, *29*(5), 62–69.
34. Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462–492.
35. Rasmussen, J. (1987). Skills, rules, and knowledge; Signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics*, *13*, 257–266.
36. Sheridan, T. B., & Verplank, W. L. (1978). Human and computer control of undersea teleoperators. Technical report. Cambridge, MA: MIT, Man–Machine Laboratory.
37. Friedkin, N. E., Mei, W., Proskurnikov, A. V., & Bullo, F. (2019). Mathematical structures in group decision-making on resource allocation distributions. *Scientific Reports*, *9*(1), 1377.
38. Simon, H. A. (1957). *Models of man: Social and rational*. Mathematical Essays on Rational Human Behavior in a Social Setting. Wiley.
39. Stubbs, K., Hinds, P. J., & Wettergreen, D. (2007). Autonomy and common ground in human–robot interaction: A field study. *IEEE Intelligent Systems*, *22*, 42–50.
40. Chen, J., & Barnes, M. (2014). Human agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human–Machine Systems*, *44*(1), 13–29.
41. Garcia, J. O., Brooks, J., Kerick, S., Johnson, T., Mullen, T. R., & Vettel, J. M. (2017). Estimating direction in brain–behavior interactions: Proactive and reactive brain states in driving. *Neuroimage*.
42. Li, S., Sun, W., & Miller, T. (2016). Communication in human–agent teams for tasks with joint action. In V. Dignum, P. Noriega, M. Sensoy, & J. S. Sichman (Eds.). *Coordination, organizations, institutions, and norms in agent systems XI* (pp. 224–241). Springer.
43. Marathe, A. R., Metcalfe, J. S., Lance, B. J., Lukos, J. R., Jangraw, D., Lai, K.-T., Touryan, J., Stump, E., Sadler, B. M., Nothwang, W., & McDowell, K. (2018). The privileged sensing framework: A principled approach to improved human–autonomy integration. *Theoretical Issues in Ergonomics Science*, *19*(3), 283–320.
44. Marathe, A. R., Schaefer, K. E., Evans, A. W., & Metcalfe, J. S. (2018). Bidirectional communication for effective human–agent teaming. In J. Y. C. Chen & G. Fragomeni (Eds.). *Virtual, augmented and mixed reality: Interaction, navigation, visualization, embodiment, and simulation* (pp. 338–350). Springer.
45. Telesford, Q. K., Lynall, M. E., Vettel, J., Miller, M. B., Grafton, S. T., & Bassett, D. S. (2016). Detection of functional brain network reconfiguration during task-driven cognitive states. *Neuroimage*, *142*, 198–210.
46. O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2020). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors*. https://doi.org/10.1177/0018720820960865
47. Larson, L., & DeChurch, L. (2020). Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *Leadership Quarterly*, *31*, 101377.
48. DeCostanza, A. H., Marathe, A. R., Bohannon, A., Evans, A. W., Palazzolo, E. T., Metcalfe, J. S., & McDowell, K. (2018). Enhancing human–agent teaming with individualized, adaptive technologies: A discussion of critical scientific questions. US Army Research Laboratory Technical Report.
49. Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, *57*(2), 103174.
50. Bansal, G., Nushi, B., Kamar, E., Lasecki, W., Weld, D., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human–AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*.
51. Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227.
52. Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D. P., Schapire, R. E., & Warmuth, M. K. (1997). How to use expert advice. *Journal of the ACM*, *44*(3), 427–485.
53. Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*(2–3), 235–256.
54. Welinder, P., Branson, S., Perona, P., & Belongie, S. J. (2010). The multidimensional wisdom of crowds. In *Advances in neural information processing systems*.
55. Askarisichani, O., Huang, E. Y., Sato, K. S., Friedkin, N. E., Bullo, F., & Singh, A. K. (2020). Expertise and confidence explain how social influence evolves along intellective tasks. *arXiv*. https://doi.org/10.48550/arXiv.2011.07168
56. Ye, W., Bullo, F., Friedkin, N. E., & Singh, A. K. (2022). Modeling human-AI team decision making. *arXiv*. https://doi.org/10.48550/arXiv.2201.02759
57. Kim, Y. J., Bachhuber, J., Botelho, C., Friedkin, N., Kim, Y., Malone, T. W., Patil, A., Ramamoorthy, S. M., Singh, A., Upadhyaya, C., Woolley, A. W., & Yang, M. J. (2020). Platform for online group studies (POGS) user manual. https://github.com/CCI-MIT/POGS/wiki/Platform-for-Online-Group-Studies-%28POGS%29-User-Manual.
58. Friedkin, N. E., & Bullo, F. (2017). How truth wins in opinion dynamics along issue sequences. *Proceedings of the National Academy of Sciences*, *114*(43), 11380–11385.
59. Friedkin, N. E., Jia, P., & Bullo, F. (2016). A theory of the evolution of social power: Natural trajectories of interpersonal influence systems along issue sequences. *Sociological Science*, *3*, 444–472.
60. Friedkin, N. E., Proskurnikov, A. V., & Bullo, F. (2021). Group dynamics on multidimensional attitudes. *Social Networks*, *65*, 157–167.
61. Austin, J. R. (2003). Transactive memory in organizational groups: The effects of content, consensus, specialization, and accuracy on group performance. *Journal of Applied Psychology*, *88*(5), 866.
62. Choi, S. Y., Lee, H., & Yoo, Y. (2010). The impact of information technology and transactive memory systems on knowledge sharing, application, and team performance: A field study. *Management Information System Quarterly*, *34*(4), 855–870.

63. Lewis, K. (2003). Measuring transactive memory systems in the field: Scale development and validation. *Journal of Applied Psychology*, *88*(4), 587–604.
64. Liang, D. W., Moreland, R., & Argote, L. (1995). Group versus individual training and group performance: The mediating role of transactive memory. *Personality and Social Psychology Bulletin*, *21*, 384–393.
65. Wegner, D. M., Erber, R., & Raymond, P. (1991). Transactive memory in close relationships. *Journal of Personality and Social Psychology*, *61*, 923–929.
66. Lewis, K. (2004). Knowledge and performance in knowledge-worker teams: A longitudinal study of transactive memory systems. *Management Science*, *50*, 1519–1533.
67. Yuan, Y. C., Carboni, I., & Ehrlich, K. (2010). The impact of awareness and accessibility on expertise retrieval: A multilevel network perspective. *Journal of the American Society for Information Science and Technology*, *61*(4), 700–714.
68. Libby, R., Trotman, K. T., & Zimmer, I. (1987). Member variation, recognition of expertise, and group performance. *Journal of Applied Psychology*, *72*(1), 81.
69. Anderson Jr, E. G., & Lewis, K. (2014). A dynamic model of individual and collective learning amid disruption. *Organization Science*, *25*(2), 356–376.
70. Palazzolo, E. T., Serb, D. A., She, Y. C., Su, C. K., & Contractor, N. S. (2006). Coevolution of communication and knowledge networks in transactive memory systems: Using computational models for theoretical development. *Communication Theory*, *16*, 223–250.
71. Ren, Y., Carley, K. M., & Argote, L. (2006). The contingent effects of transactive memory: When is it more beneficial to know what others know? *Management Science*, *52*, 671–682.
72. Huang, E. Y., Paccagnan, D., Mei, W., & Bullo, F. (2022). Assign and appraise: Achieving optimal performance in collaborative teams. *IEEE Transactions on Automatic Control*. Advance online publication. https://doi.org/10.1109/TAC.2022.3156879
73. Mei, W., Friedkin, N. E., Lewis, K., & Bullo, F. (2018). Dynamic models of appraisal networks explaining collective learning. *IEEE Transactions on Automatic Control*, *63*(9), 2898–2912.
74. MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, *40*(4), 291–303.
75. Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology*, *29*(3), 261.
76. Plous, S. (1993). *The psychology of judgment and decision making*. New York: McGraw-Hill.
77. Thomas, J. P., & McFadyen, R. G. (1995). The confidence heuristic: A game-theoretic analysis. *Journal of Economic Psychology*, *16*(1), 97–113.
78. Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, *17*(1), 39–57.
79. Pulford, B. D., Colman, A. M., Buabang, E. K., & Krockow, E. M. (2018). The persuasive power of knowledge: Testing the confidence heuristic. *Journal of Experimental Psychology: General*, *147*(10), 1431.
80. Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121.
81. Pohl, R. F. (2004). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press.
82. Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology*, *2*(1), 83–115.
83. Dardenne, B., & Leyens, J.-P. (1995). Confirmation bias as a social skill. *Personality and Social Psychology Bulletin*, *21*(11), 1229–1239.
84. Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes*. Boston, MA: Houghton Mifflin.
85. Davison, H. K., Mishra, V., Bing, M. N., & Frink, D. D. (2014). How individual performance affects variability of peer evaluations in classroom teams: A distributive justice perspective. *Journal of Management Education*, *38*(1), 43–85.
86. Buunk, A. P., & Gibbons, F. X. (2007). Social comparison: The end of a theory and the emergence of a field. *Organizational Behavior and Human Decision Processes*, *102*(1), 3–21.
87. Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140.
88. Golub, B., & Jackson, M. O. (2010). Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, *2*(1), 112–149.
89. Frey, V., & van de Rijt, A. (2020). Social influence undermines the wisdom of the crowd in sequential decision making. *Management Science*, *67*, 4273–4286.
90. Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P. M., Moussaid, M., & Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, *117*(21), 11379–11386.
91. Dow, J., & Ribeiro da Costa Werlang, S. (1992). Uncertainty aversion, risk aversion, and the optimal choice of portfolio. *Econometrica*, *60*(1), 197–204.
92. Cummings, M. L. (2015). Automation bias in intelligent time critical decision support systems. In D. Harris & W.-C. Li (Eds.), *Decision making in aviation* (pp. 289–294). Routledge.
93. Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human–Computer Studies*, *58*(6), 697–718.
94. Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human–Computer Studies*, *51*(5), 991–1006.
95. French Jr., J. R. P., & Raven, B. (1959). The bases of social power. In D. Cartwright (Ed.). *Studies in social power* (pp. 150–167). Institute for Social Research, University of Michigan.
96. Raven, B. H. (1965). Social influence and power. In I. D. Steiner & M. Fishbein (Eds.). *Current studies in social psychology* (pp. 371–382). Holt, Rinehart, Winston.
97. Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science*, *9*(3), 458–467.
98. Dalkey, N. C. (1969). The Delphi method: An experimental study of group opinion. Technical report. RAND Corporation.
99. Hsu, C.-C., & Sandford, B. A. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research & Evaluation*, *12*(10), 1–8.
100. Boehm, B. W. (1981). *Software engineering economics*. Hoboken, NJ: Prentice-Hall.
101. Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276.
102. Madirolasand, G., & de Polavieja, G. G. (2015). Improving collective estimations using resistance to social influence. *PLoS Computational Biology*, *11*(11), e1004594.
103. Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 129–140.
104. Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology*, *69*(6), 1069.
105. Larrick, R. P., Morgan, J. N., & Nisbett, R. E. (1990). Teaching the use of cost–benefit reasoning in everyday life. *Psychological Science*, *1*(6), 362–370.
106. Bauer, C. C., & Baltes, B. B. (2002). Reducing the effects of gender stereotypes on performance evaluations. *Sex Roles*, *47*(9), 465–476.

107. French Jr., J. R. P. (1956). A formal theory of social power. *Psychological Review*, *63*(3), 181–194.
108. Harary, F. (1959). A criterion for unanimity in French's theory of social power. In D. Cartwright (Ed.). *Studies in social power* (pp. 168–182). Ann Arbor, MI: University of Michigan.
109. DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, *69*(345), 118–121.
110. Friedkin, N. E., & Johnsen, E. C. (1990). Social influence and opinions. *Journal of Mathematical Sociology*, *15*(3–4), 193–206.
111. Friedkin, N. E., & Cook, K. S. (1990). Peer group influence. *Sociological Methods & Research*, *19*(1), 122–143.
112. Cartwright, D. (1971). Risk taking by individuals and groups: An assessment of research employing choice dilemmas. *Journal of Personality and Social Psychology*, *20*(3), 361–378.
113. Cartwright, D. (1973). Determinants of scientific progress: The case of research on the risky shift. *American Psychologist*, *28*(3), 222.
114. Friedkin, N. E. (1999). Choice shift and group polarization. *American Sociological Review*, *64*(6), 856–875.
115. Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

How to cite this article: Askarisichani, O., Bullo, F., Friedkin, N. E., & Singh, A. K. (2022). Predictive models for human–AI nexus in group decision making. *Ann NY Acad Sci*., *1514*, 70–81. https://doi.org/10.1111/nyas.14783