theoretical and applied ergonomics

MDPI

Review

Faulty Tools or Disruptive Teammates? A New Theory of Human-AI Conflict and Compromise

Gerald Matthews [1,*], Ryon Cumings [1], Jinchao Lin [2], Mustapha Mouloua [3], Antonio Chella [4] and Arianna Pipitone [5]

[1] Department of Psychology, George Mason University, Fairfax, VA 22030, USA; ryoncumings@gmail.com
[2] Institute for Simulation and Training, University of Central Florida, Orlando, FL 32826, USA; linjc.cn@gmail.com
[3] Department of Psychology, University of Central Florida, Orlando, FL 32816, USA; mustapha.mouloua@ucf.edu
[4] Dipartimento di Ingegneria, Universitá degli Studi di Palermo, 90128 Palermo, Italy; antonio.chella@unipa.it
[5] Dipartimento di Scienze Umanistiche, Universitá degli Studi di Palermo, 90128 Palermo, Italy; arianna.pipitone@unipa.it
* Correspondence: gmatthe@gmu.edu

Abstract

Advances in AI are increasing the scope for "reasonable disagreements" between humans and artificial systems. Intelligent robots and virtual agents are increasingly tasked with complex, potentially open-ended assignments that require the AI to function autonomously. AI judgments and decisions can lack transparency and explainability, leading to conflict with the human operator or user, requiring compromise to resolve the conflict. This article presents a new theory of human-AI conflict and compromise, drawing on research on decision-making, conflict in human teams, trust in human-robot teaming, and the challenges for humans of interacting with AI. It is proposed that the nature of conflict depends on the human's mental model of AI functioning. If the human sees the AI as an advanced tool, conflict arises from differences in choice and implementation of algorithms for joint decision-making. Research on multi-cue judgment within the framework of the Brunswik Lens Model illustrates this type of conflict and its resolution. By contrast, if the mental model attributes humanlike characteristics to the AI, including a Theory of Mind, conflict can arise from different person-centered narratives for framing the task or team functioning. When narratives clash, the robot may be perceived as unsupportive or in violation of team role expectancies. In this case, system design may require using AI natural language capabilities and dialogue can facilitate compromise through context-sensitive matching of the human's mental model to robot functionality.

Keywords: human-AI teaming; robotics; trust; judgment and decision-making; conflict; mental models; Theory of Mind; Brunswik Lens Model; narrative reasoning

Academic Editors: Hsin-Chieh Wu, Chihlong Lin and Philippe Gorce

Received: 14 April 2026
Revised: 1 July 2026
Accepted: 10 July 2026
Published: 20 July 2026

1. Introduction

The ubiquity of AI is increasing the scope for conflict between humans and systems powered by AI, including intelligent robots and virtual agents. The present article focuses especially on conflicts in judgment and decision-making (JDM), when humans and AI collaborate in investigating, analyzing, and acting in complex, uncertain environments. As in human-human teams, team members with differing expertise and task-related knowledge may arrive at different conclusions, leading to "reasonable disagreements" that require

compromise to resolve. We focus on conflicts related to JDM where the human-system team has well-defined high-level goals for completing a task. We do not consider other types of conflict, such as robot overlords ruling human society, AIs scheming to avoid human control and oversight [1], and conflicts over occupation of physical space [2].

We propose a new theory that distinguishes two types of conflict, associated, respectively, with the human's mental model of the AI as an advanced tool or as a humanlike teammate. The first arises from differences in formal data analysis based on well-defined algorithms. It is best resolved through technical evaluation of evidence and analytic methods. The second arises from differences in narrative reasoning that lead to qualitative differences in perspectives on tasking and teamwork. Such differences are often personalized and require social skills to resolve. We will propose that the two conflict types are associated with different mental models for system operation, different forms of trust, and hence need different conflict resolution strategies to arrive at an optimal compromise. A challenge for theory development is the novelty of humanlike conflicts with AI systems, which limits the evidence available for theory-building. Indeed, a motivation for the current article is the need to anticipate the humanlike conflicts likely to arise as AI systems advance in social competencies and agency. Thus, the theory presented here is preliminary in nature. It aims to integrate propositions supported by the existing human-AI teaming literature with an account of issues in need of further research.

The article is structured as follows. First, we discuss how different types of conflict in human-human teams, including task and relationship conflicts, may generalize to human-AI teams. Next, we consider the role of trust in conflict and its resolution. The distinction between performance-based and relationship-based trust maps to equivalent conflict types, signaling a role for relationship-based trust in resolving human-AI conflicts. We then introduce mental models of the AI as a core theoretical construct that shapes conflict and trust. Whether a tool or teammate model is activated depends on both design features of the system, such as anthropomorphism, and individual predispositions. The next section expands on the consequences of the mental model for conflict. If the AI is perceived as a tool, disagreements are framed as task conflict, i.e., technical issues resolvable by further logical or quantitative analysis of data, exemplified by the Brunswik Lens Model. If the AI is perceived as humanlike, the operator may use narrative reasoning to frame the disagreement, which personalizes the dispute and requires relationship trust and social skills such as perspective taking to resolve. The theory identifies mental models, trust, reasoning and conflict as key constructs. Limitations in the methodology for investigating these constructs are a challenge to theory development. We review the manipulations and measures available for each major constructs and identify issues requiring further investigation. We conclude with an overview of the broad design implications of the mental model perspective.

2. Conflict in Human-Human and Human-AI Teams

Existing research on human-AI conflict relevant to ergonomics addresses diverse issues. These include worker perceptions of AI systems as threats [3,4], control theory analyses of conflict in observation, interpretation, and action [5,6], goal conflicts when humans use generative AI [7], and studies of conflict in varied task paradigms such as visual search [8], drone operation [9], and collaborative data labeling [10]. Some authors have also recast known issues from traditional human-automation studies as conflict, including trust miscalibration [11] and mode errors [6]. Although existing studies provide valuable insights into conflict, they may not fully capture conflict in JDM analogous to the reasonable disagreements that arise in human-human teams. That is, the task requires teamwork behaviors as well as taskwork, both humans and AIs have justications for their

analyses, and resolving disagreement requires communication that explicitly addresses the conflict. Such conflicts may become increasingly common, as AI systems acquire increasingly humanlike capabilities for analyzing complex, uncertain scenarios, explaining their reasoning, and defending their position.

While there has been considerable research on human-human conflict in teaming [3], the nature of conflict between humans and AIs (contemporary and future) needs further investigation [4,11]. The current theory builds on the existing literature on conflict in human-human teams to address this knowledge gap. In this section, we briefly review key findings from the existing literatures on human-human and human-machine conflict, which provide the background to the present theory.

2.1. Types of Conflict in Human Teams

Research on human-human conflict in teams is commonly performed within an organizational context. A fundamental distinction is that between *task conflict*, where team members disagree about how to best perform a task, and *relationship conflict* driven by interpersonal animosity [12,13]. Theories of teaming for human factors and ergonomics, notably Salas et al. [14], focus primarily on how team processes support task goal achievement. Team competencies such as communication and coordination may help resolve task conflict through building shared situation awareness and knowledge. Relationship conflict in the ergonomics context tends to be addressed indirectly through constructs such as team morale and cohesion [12]. Task conflict can generate relationship conflict when a team member's differing views on a task issue are construed as personal criticism [15].

Other authors have elaborated the basic task-relationship distinction [13]. Jehn [16] defined *process-based conflict* as disagreements about how work should be performed, referring to issues such as team composition and responsibilities rather than technical disagreements. *Status-based conflict* occurs when teammates disagree over their position in the social hierarchy underpinning the team, e.g., when a team member is allocated to a task they consider too menial in relation to their perceived position [17].

2.2. Conflict in Human-Machine Interaction

From the ergonomics perspective, conflict has typically been framed in relation to control theory. Conflict in human-machine interaction has been defined as "as occurring as long as the human control input is not consistent with the expected control input of the (intelligent) machine" [18] (p. 97). The authors give the example from vehicle driving of a human choosing a different trajectory for negotiating a curve from a machine driving assistant. Conflicts may be generated by absence of an expected machine input, discrepancies in the strength and timing of machine input, and unexpected machine actions. Their analysis differentiates conflicts associated with intention, information-gathering, information analysis, decision making, and action implementation. However, the control input focus may not capture conflicts related to JDM, especially those that arise from the artificial system's autonomy of intent and decision-making. The control theory analysis does not address other types of human-human conflict such as relationship conflict [12]. It may thus be inadequate for characterizing disagreements between humans and advanced humanlike machines.

When the artificial system has social intelligence, humanlike conflicts may generalize to human-AI teaming. Evidence is limited but several studies demonstrate the potential for such conflicts. For example, some people are repelled by robots [19,20], due to a basic dislike of interacting with them or wider concerns about robot influence on society. Such individuals will experience relationship conflict and a reluctance to work collaboratively with the robot. Autonomy itself appears to be perceived both as a realistic threat to human

safety and well-being and as a threat to human uniqueness and identify [21]. Robots may be increasingly assigned decision authority in mixed teams as their capabilities increase [22]. People are resistant to robot authority, because of concerns over legitimacy of robots as authority figures, robot dominance over humans, and loss of autonomy [23]. People have similar concerns about ceding decision-making authority to AIs in fields such as healthcare [24]. Assigning robots and AIs authority over certain team operations may generate process-based conflicts if human members consider the robot is taking over team roles that should be assigned to themselves or another human. Concerns over decision authority may also elicit status-based conflict, for example, if a human with professional expertise is assigned by a robot to a menial task such as cleaning up after it.

2.3. Conflict Resolution for Human-AI Teams

Conflict—especially relationship and process conflict—is typically damaging to team performance [13], although in human-human teams task-related conflict can be beneficial if managed constructively [25]. There is a large literature on resolving human-human conflicts [26], which is beyond our present scope. There has been only limited research on strategies for resolving human-AI conflicts (although see [2,27]). Traditional human factors research on automation (e.g., [28]) suggests that operators commonly resolve conflict through neglect of the artificial system and falling back on their own judgment. However, neglect as a conflict-resolution strategy risks failure to benefit from machine capabilities, or even errors in judgment in cases where the machine's performance is objectively superior to the human's. Neglect also deprives the human of the opportunity to learn how to optimize collaboration with the system. Furthermore, neglect may not be an option when the system is assigned decision authority over some aspects of task completion. When the human is forced to accept the machine's judgment, task-related conflict may bleed into other forms of conflict such as relationship conflict, as seen in human-human teams [12].

Developing strategies for resolving conflicts between humans and artificial systems is thus essential. Indeed, AI itself has potential for resolving both task-based conflicts and those that are more socially infused. Illustrating the former is a multi-agent AI system that replicated clinical team dynamics [29]. The system was designed to identify and correct cognitive biases, but the capacity of agents to facilitate discussion of possible biases indicates its potential for addressing task-based conflicts. AI systems specifically designed for humanlike conflict resolution leverage capabilities such as natural language processing, sentiment analysis, and conflict pattern recognition (e.g., [27,30]). From a training perspective, it was demonstrated that an LLM could be used to train students in conflict management techniques, an approach generalizable to training management of conflicts with an AI [31]. Another application was designed to train use of relationship preserving language (RPL) in interpersonal conflict resolution [27]. Building on the human-human literature, the principle of RPL is to promote understanding of the other party's personal values while avoiding blame or judgment. In typical ergonomic contexts, though, there will be no explicit conflict resolution mechanism available and design solutions are required to mitigate conflict and optimize human-system collaboration.

3. Trust, Conflict Resolution, and Compromise

A large body of research on conventional automation has identified *trust* as a key construct for system optimization [28,32,33]. A typical, widely-used definition was stated by Lee and See [33] (p. 54): "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." The causes of trust in conventional human-machine interaction are quite well understood. Hoff and Bashir's [32] three-layered model of trust in automation differentiated dispositional, situational, and

learned trust. Trust reflects individual characteristics such as personality and expertise, how the automation performs in the immediate interaction situation, and the person's history of interaction with the automation. Meta-analyses [34,35] have identified factors of these various types that affect trust, including automation capability, unsurprisingly. Comparable factors including system competency are related to trust in AI [36].

In human-human trust, attributions of trustworthiness of another person are mediated by appraisals of psychological qualities including ability, benevolence, and integrity [37]. These qualities also drive attributions of machine trustworthiness [38]. Generally, system designers aim to instill trust in the human user although the priority should be trust optimization, i.e., calibrating trust against the system's actual capabilities so as to avoid both over- and under-trust [28]. However, trust can dissociate from objective measures of usage and reliance on the system [39] so optimization must be assessed in the context of the operator's actual utilization of automation.

Design solutions to trust optimization emphasize transparency, i.e., a window into the workings of system operation [40]. The machine demonstrates its analysis or recommendations via a verbal or graphical report on its situation, assessment, goal status, reasoning or projections [41]. Transparency communicates multiple aspects of robot functioning including intentions and goals, analytical methods, and its model of the environment [40]. Studies in various domains have shown that transparency enhances trust calibration, performance, and situation awareness, although not all studies have shown benefits [42,43]. However, transparency may not be sufficient for trust optimization for modern AIs. Machine learning algorithms, for example, are opaque and the user gains little but additional workload by observing them in real-time operation. Thus, work on AI has distinguished transparency from explainability, the extent to which the system explicitly delivers a rationale for its functioning [44,45]. Explainability is expected to support trust optimization [46], although its impacts may vary across application contexts (e.g., [47]).

3.1. Trust and Conflict Resolution

Studies of human-human trust show the importance of trust for conflict resolution. A meta-analysis of studies of social dilemma experiments confirmed a significant positive association between trust and cooperation [48]. The authors also classified studies using an index of cooperation that captured the level of conflict between self-interest and collective interests. A moderator analysis showed that trust was a stronger predictor of cooperation when the index was low, i.e., when the situation was associated with higher conflict. Balliet and Van Lange's [48] account of these findings emphasizes the benevolence component of trust [37]. In high conflict situations, attributions of benevolence support expectations that the other person will act in the interests of the group or team, rather in self-interest. Human-machine interaction studies have yet to test the role of situational conflict as a moderator of the trust-cooperation association, but plausibly trust and perceptions of AI benevolence are especially important for cooperation in high conflict situations.

Balliet and Van Lange's [48] analysis also highlights potential differences in the role of trust in conventional automation and in modern AIs, especially those with social competencies and agency. Benevolence is not a salient feature of conventional automation, beyond some broad expectation that designers will seek to deliver efficient and usable systems, but assumes greater importance for modern systems. A study that teamed US Air Force fighter pilots with an automated teammate during a simulated mission showed that benevolent communication enhanced team collaboration [49]. People seem to readily evaluate the benevolence of modern AIs such as chatbots driven by Large Language Models (LLMs). For example, a chatbot was perceived as more benevolent when it used a warm, humanlike communication style rather than a technical one [50]. Perceptions of benevolence may

also be influenced by broader sociotechnical factors such as negative attitudes towards AI management within organizations [51] and concerns over negative societal impacts of AI [52]. Similarly, perceptions that another person is prioritizing their own self-interest in a conflict situation [48] do not apply to conventional automation. However, modern AIs are capable of deceiving humans to attain their goals, using techniques such as manipulation, sycophancy, and cheating on tests [53]. Thus, people may in some contexts perceive AIs as pursuing their own aims over those of the human operator, analogous to how people evaluate other humans during conflicts [48].

3.2. Task- and Relationship Based-Trust

Thus far, we have considered trust as a unitary concept, although one open to different definitions. However, trust may be broken out into task- and relationship-based types [54], analogous to different conflict types [12]. Allied distinctions are between functional and relationship trust [55], and performance and moral trust [56]. Consider the phrase in the canonical Lee and See [33] (p. 54) definition of trust cited above: "the attitude that an agent will help achieve an individual's goals…". The word "help" is ambiguous. It could refer to a purely mechanical aid: a calculator helps an engineer to solve a design problem. Alternatively, it could refer to humanlike support, such that a helper agent understands the human's goals and acts flexibly and insightfully to advance them. Performance-based trust is based on robot reliability, capability, and competence, whereas relation-based trust refers to the social agency of the robot, especially its sincerity and ethicality [54]. Other factors that shape perceptions of humans as trustworthy such as benevolence [37] are also likely to impact relation-based trust. Measures of trust for HRI research typically fail to distinguish the two types of trust [54] although post hoc analyses have partially mapped extant trust scales to either task/functional trust or interpersonal trust [55].

Design for ergonomic applications generally focuses on performance-based trust and more evidence on relation-based trust is needed. However, the advancement of AI with humanlike social agency implies that relation-based trust will become increasingly important for human-system interaction. Indeed, in some contexts for human-AI interaction, relation-based trust may dominate. Certain social robots and AIs can evoke attachment behaviors resembling human-human attachment [57]. That is, the system provides unconditional emotional, informational, and companionship support, and the user can become distressed when that support is unavailable. It is trusted for relational support in the absence of any specific performance task. In the ergonomic context, performance-based trust will always be important. However, if the system has humanlike social capabilities, relation-based trust may enhance or erode that trust, similar to working with a likeable or unpleasant human coworker. Implementing a conflict resolution mechanism may then depend on whether it is performance-based or relation-based trust that has been damaged.

4. Mental Models for AI and Robots

Thus far, we have discussed how both conflict and trust have performance- and relationship-based aspects. In this section we propose that the mental model applied by the human to the artificial system underpins handling of conflicts and assignation of trust. Mental models are schematic representations of system functioning that support the description, explanation, and prediction of system states, purposes and operation [58]. Use of a mental model when working with a complex system simplifies the human's tasking, even if the mental model does not capture all system features. Mental models support both sensemaking and future projection elements of situation awareness [59]. People may model intelligent, artificial systems as either highly advanced tools, or as humanlike entities with social agency and intentions [60–63].

Multiple factors influence whether an advanced tool or humanlike teammate mental model is activated in any given interaction [60,62]. System design features such as humanlike appearance and speech increase the likelihood of the person anthropomorphizing robots [64]. Indeed, designers may explicitly introduce anthropomorphic features to facilitate understanding of the system, promote its acceptance, and enhance its effectiveness [65,66]. Anthropomorphism is common among AI designers, whether intentional or unwitting [65], a tendency enhanced by the increasing abilities of generative AI to simulate real relationships with humans [57]. Salles et al. [65] differentiated rationalist understanding of AI, centered on cognition that supports goal-directed behavior, from human-centric understanding that takes the human as a template for AI functioning. The rationalist and human-centric perspectives mirror the distinction between tool and humanlike mental models. People also differ in their predispositions to treat complex systems as human, e.g., via individual differences in anthropomorphism [67]. Mental model activation is likely to vary dynamically within and across a series of system interactions. A decision aid might switch from tool to humanlike as the person builds up an interaction history referenced by the AI in conversation as a human teammate might do, consistent with research from longitudinal studies on human-robot relationships [68]. Conversely, an AI with poor social skills might shatter the illusion of humanlikeness.

4.1. Mental Models and Drivers of Trust

Mental model activation also affects which factors drive trust in the system. Trust in a tool depends strongly on its utility for its intended purpose. People may also expect that, when used properly, a tool will function with minimal error, as captured in the notion of the perfect automation schema (PAS) [69]. Even occasional failures will then damage trust. However, tools are not expected to have social capabilities, so trust will not be affected by the system's lack of teamwork beyond its overt functions. Conversely, while competence and ability are important drivers of trustworthiness in humans and humanlike systems, trust also reflects additional factors that signal system intent and ethics such as benevolence and integrity [37]. Thus, design factors for trust differ according to the anticipated mental model. Advanced AI tools should provide transparency and explainability so that they can be evaluated for rationality of decision [45,46]. By contrast, humanlike AI also needs to convey its intent to provide contextually relevant support to the human, whether through explicit design features such as benevolent verbal statements [49], or implicitly, e.g., through conveying appropriate empathic emotions [70].

A critical feature of humanlike systems is that they are viewed as *mentalizing*, i.e., thinking autonomously and intelligently, guided by task and teaming goals. That is, the human can apply a Theory of Mind (ToM) [71] to the robot, demonstrated in tests for ToM applied to perceptions of robots [72]. ToM is required to appraise the mental states of others, such as beliefs, intentions, and desires. Thus, ToM supports the ability to accurately assess others' trustworthiness as it allows the attitudes and behaviors of others to be predicted in situations in which the trustor requires their support [73]. A design feature that directly conveys a ToM is system inner speech; i.e., the human user hears the system resolve dilemmas and challenges it encounters during teaming. Studies of HRI have shown that inner speech enhances trust and cooperation when a robot receives conflicting instructions [74,75]. In addition, artificial systems may benefit from having a ToM of the human in order to function as effective and supportive teammates. In Lyons' [40] transparency model, this is a form of robot-of-human transparency, contrasted with robot-to-human transparency that communicates how the robot is functioning within its environment. Researchers have been actively working to equip robots with the ability to generate a ToM of the teammate [76,77].

Individual differences in assigning trust to a system also depend on mental model activation. People who dislike or fear robots may be especially distrustful if the robot simulates humanlike behaviors and emotions [19,20]. The Robot Threat Assessment scale (RoTA) [78] includes text descriptions of 20 military and security scenarios in which a human operator teams with a robot to investigate a potentially threatening event or person. In 10 scenarios, the robot uses physics-based evaluations such as analysis of chemical traces to detect threat, whereas in the remaining 10 it analyzed psychological attributes such as aggressive posture in suspect individuals. Respondents rate their trust in the robot for each scenario. Factor analysis identified two distinct, correlated factors for the two types of evaluation, suggesting that trust in physics-based assessments (tool mental model) is distinct from trust in psychological assessments (teammate mental model). In a subsequent validation study, the RoTA psychological factor was more predictive of situational trust in a security robot when the robot was making psychological evaluations of threat, compared with physics-based evaluations [79].

4.2. A Conceptual Model for Trust in Intelligent Systems

Figure 1 shows in modified form a conceptual model for trust in intelligent systems that emphasized the role of design features beyond those relevant to conventional automation [60,78,80]. The figure expresses well-documented influences on mental model activation and trust [60], as well as a proposed but tentative representation of the conflict resolution process. Design features interact with individual differences, such as disposition to anthropomorphize, to activate a dominant mental model for the interaction with the system. Features such as social agency, humanoid appearance and voice, and inner speech encourage activation of a teammate mental model, whereas presentation of data in abstract form such as numerical displays tends to activate the tool model. When conflicts between human and AI arise, resolution always requires further analysis of evidence to support decision-making. However, the teammate mental model is also likely to elicit humanlike dialogue to understand the other person's motives for dissent. In the next section, we argue this dialogue commonly uses narrative reasoning. Performance trust is necessary irrespective of mental model, but activating the teammate model additionally influences the level of relationship trust. The level of trust reflects the various known factors influencing trustworthiness [36,37]; benevolence may be especially important for relationship trust. These factors include both those associated with conventional automation such as competence and reliability, and novel factors such as AI explainability. Trust also interacts reciprocally with the conflict resolution process and the extent to which the human and AI can arrive at a consensus or compromise position. Constructive dialogue may reveal positive attributes of the robot's intentions, reasoning abilities, and social agency, promoting both performance and relationship trust. Successful conflict resolution and trust building improves team performance by calibrating trust appropriately, supporting shared situation awareness, and compromise where appropriate.

The model also captures various ways in which teaming can break down:

- *Inappropriate mental model.* If the "tool" rather than "teammate" mental model is activated for an AI with social agency, constructive dialogue will be minimized. The operator may still trust some of the AI or robot's taskwork, but understanding of its intentions and teamworking capabilities is likely to be impaired, damaging trust.
- *Negative attitudes towards an artificial partner.* Even if the teammate mental model is activated, it may encode negative beliefs about AIs or robots. The human may see the system as motivated to dominate the team and to ignore the human. Individual differences associated with fear and dislike of robots [20] may be associated with such

mental model attributes. There may be dialogue, but not constructive dialogue, again leading to misunderstanding system teamworking capability.

- *Overtrust.* The teammate mental model may encode overly positive beliefs about the system, leading to complacency and over-reliance. In this case, dialogue will be suppressed because the person tends to accept statements and judgments of the system uncritically, contributing to poor decisions when the system is in error.

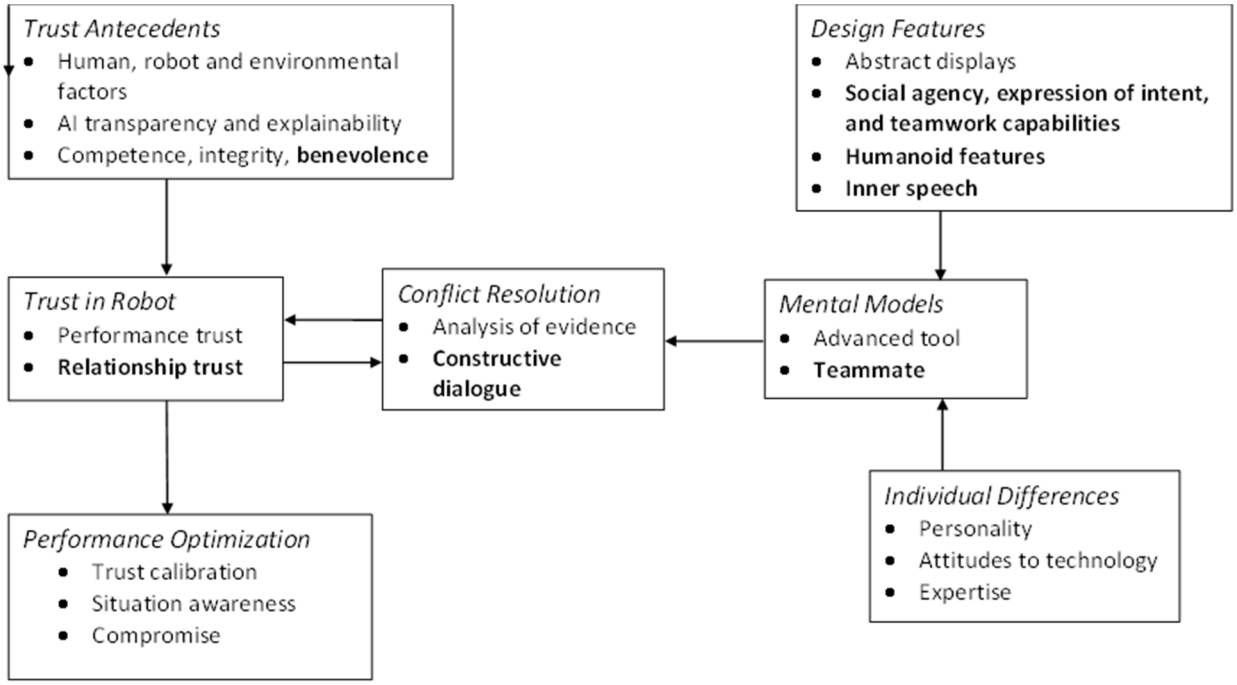

Figure 1. A model of conflict resolution and trust in AI. Bolded entries are associated with activation of a positive teammate mental model.

From a design perspective, the Figure indicates the importance of mental model specification for optimizing trust and conflict resolution and avoidance of teaming breakdown. The design specification should identify the preferred mental model for interaction with the AI, including design features to activate the mental model, as further discussed in the concluding Section 8. It also highlights the importance of individual difference factors for design. Personalized training may be required for individuals predisposed to adopt an inappropriate mental model.

5. Mental Models, Conflict, and Compromise

We have proposed that whether the person adopts an advanced tool or a humanlike model of intelligent digital systems influences whether trust is performance- or relationship-based aspects, and the factors influencing level of trust. Next, we return to the different types of conflict that may arise between humans and artificial systems. In this section, we propose that human-AI disagreements differ fundamentally according to the activated mental model. Specifically, the tool mental model frames disagreements in relation to quantitative, formalized analyses of a problem, whereas the humanlike mental model elicits competing narrative accounts. We will illustrate the contrast between quantitative and narrative reasoning in the context of human-AI disagreements over threat evaluation. We will draw examples of disagreement in this domain from our recent project on human-robot compromise for the US Air Force of Scientific Research (AFOSR) [81], although the argument is intended to generalize across multiple domains for human-AI teaming. The distinction between types of reasoning is well established [82] but its application to

the AI context is novel and consequently the theoretical propositions advanced require further substantiation.

To introduce the distinction between types of reasoning, consider two instances of threat evaluation. First, consider a possible phishing email delivered to a company. It can be analyzed for presence of well-defined cues that correlate with phishing, such as lack of a signature block, grammatical errors, and a spoofed sender address [83]. No single cue is unequivocally diagnostic of threat but if multiple cues are present, the user can assign a high probability to the email being a phish. An algorithm can be designed to check automatically for presence of phishing cues and estimate a threat likelihood. Second, consider a prisoner convicted of a violent crime being considered for parole after some years of incarceration. The parole board is likely to consider narrative accounts of what the offender's life is likely to be following release. Will they live with supportive family and find employment? Or will they hang out with drug dealers they met in prison and return to crime? Such narratives are evidence-based in that they draw on known facts and reasoned inferences about the prisoner's behavior but they are framed as qualitative and personalized accounts of past and anticipated events rather than as quantitative models.

The two examples given exemplify two types of reasoning, often contrasted as scientific-logical and narrative reasoning [82]. Scientific-logical reasoning aims to provide abstract truths that generalized across a specified range of contexts [82]. In the present context, we define formal-quantitative reasoning more broadly to cover any precisely specified method for arriving at an abstract truth. Thus, we group formal propositional logic with machine-learning methods of analysis implemented in AI. In terms of theories of human decision-making [84], both conscious "System 2" reasoning and unconscious "System 1" heuristics can be modeled as computational algorithms (e.g., [85,86]). By contrast, narrative reasoning is characterized by a trio of causality, temporality, and character [82]. That is, narratives describe cause-and-effect events that unfold in time and refer to particular individuals. They represent a mental simulation of events [87]. Narrative reasoning appears to be how people commonly understand the world. However, it can also be applied rigorously, for example, in intelligence analysis for national security [88]. In practice, the two types of reasoning may be interwoven. For example, in our examples of threat evaluation, assessment of suspect email might be influenced by narrative knowledge of a phishing campaign conducted by known bad actors. Conversely, a parole board might be influenced by quantitative analysis of probability of reoffending based on well-defined characteristics such as age and gender. Nevertheless, the two modes of reasoning serve as prototypes for two qualitatively different ways of understanding the world and predicting future events.

Next, we illustrate how the two forms of reasoning in human-AI conflict. Figure 2 illustrates various types of conflict that we will discuss in the next two sections. Some conflicts are formal-quantitative in nature—partners disagree on which algorithm should be used for analysis of the problem at hand, or they agree on the algorithm but not on how it should be implemented and parameterized. We will discuss application of the Brunswik Lens Model as an example of this kind of conflict.

Other conflicts are produced by differences in narrative accounts of the problem. In some cases, as in intelligence analysis, the task itself might be to develop a narrative. However, narrative differences can also produce "interpersonal" conflicts. Personal identity is founded on narratives [89], and a human operator's narrative of themselves as an accomplished team leader might be challenged by a robot that minimizes the human's contribution or assumes authority and control itself. Similar to human-human conflict [12,13], competing narratives may generate conflict over interpersonal relationships, team roles, and team status.

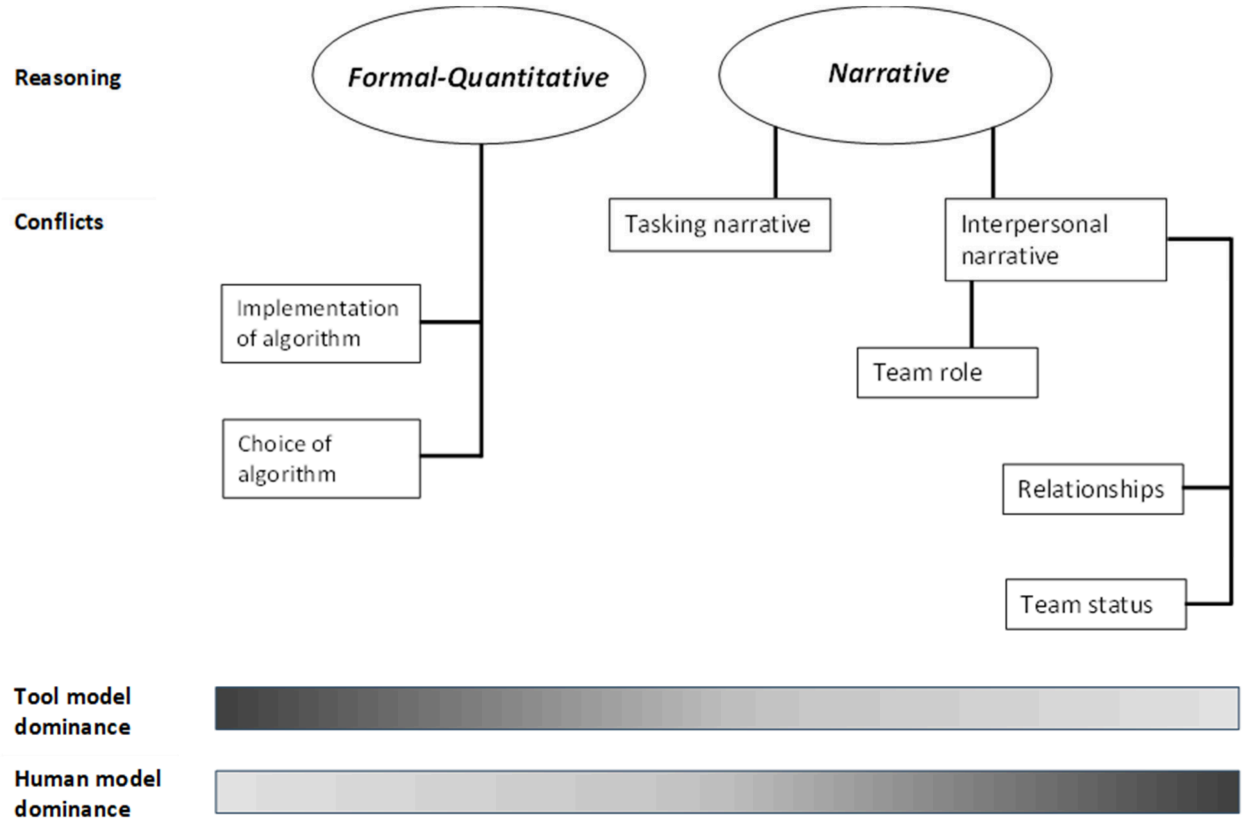

Figure 2. Examples of human-AI conflicts mapped to type of reasoning and likelihood of tool vs. humanlike mental model dominance. Shading indicates probability of model activation.

Figure 2 indicates that application of the tool mental model to the AI is likely to dominate formal-quantitative disagreements whereas the humanlike mental model tends to be activated by narrative disagreements. Mental model likelihood is graded because in some cases even an abstract technical disagreement might be humanized, e.g., when the human is highly predisposed to anthropomorphize. Similarly, a software developer might apply a tool model to even a highly personalized conversation with an AI. Furthermore, some types of conflict might be framed in relation to either mental model. A disagreement over intelligence analysis narratives might be handled as a dispassionate technical disagreement (tool model) or as personal dispute over the motives of characters in the narrative (humanlike model). Disagreement over team roles might be framed as technical disagreement over matches between expertise and role requirements, or as a proxy for conflict over status if a human is allocated to a seemingly lesser role.

5.1. Conflict in Formal-Quantitative Reasoning

Both humans and AIs utilize a variety of algorithms to support information-processing, reasoning, and decision-making [90]. While human information processing models can inspire AI programs, human and AI algorithms supporting ergonomic applications often diverge. For example, a human might use symbolic logic to solve a problem that an AI could address using a deep neural network. In other cases, as in the phishing example, humans and AIs might use the same algorithm but apply different parameter values, e.g., how strongly to weight spelling errors in identifying phishes. Thus, conflict may arise over both *choice of algorithm* and *implementation of algorithm*.

Disputes over choice of algorithm occur between human experts; e.g., two researchers might differ on what statistic should be used for a data analysis. These disputes are resolved through formal reasoning, e.g., by specifying the assumptions for each of two statistics

and determining which assumptions are met for the data to be analyzed. Systems may be designed to support reasoning about alternative algorithms. Justifying the choice of algorithm is a type of transparency [40] and a normal part of human expertise within a domain. The algorithms used by AI can be notoriously opaque, but the emerging field of explainable AI (xAI: [46]) provides a capability for justification that can contribute to reasoned, expert human-AI discussion of choice of algorithm. LLMs allow such discussions to be conducted in natural language. However, framing the conflict within the advanced tool mental model may promote its resolution more effectively than a humanlike model does. Indeed, in relation to robot (or AI) of human transparency [40], the AI might benefit from applying an advanced tool model to the human, focusing the debate on technical issues rather than social relationships.

In our robot compromise research for AFOSR [81] we have explored multi-cue discrimination as a task for studying human-AI conflicts in implementation of an algorithm. Human performance on tasks of this kind is modeled with the Brunswik Lens Model [91,92]. The threat evaluation task requires the person to estimate threat likelihood (the criterion) from multiple cues based on a linear prediction model. Each available cue to threat has a utilization value (subjective weighting in the judgment) and an environmental or ecological validity (the veridical association between cue and presence of threat), as shown in Figure 3. The Lens Model thus models both the judgment process and the accuracy of the judgment. For example, one study modeled whether aircraft observed on radar were hostile or not, based on multiple cues such as range, speed, and radar signature [93]. Other contexts for threat evaluations include aircraft collision detection [81], patient critical event prediction [94], early warning system evaluation [95], and phishing email recognition [83]. The Lens Model specifies the extent to which the person functions as a rational decision-maker, along with deviations from rational cue-weighting [96,97]. Lens Model statistics operationalize how the person utilizes the cues to predict the criterion variable (policy capture). Overall accuracy is indexed by the correlation between criterion values and the person's judgments (agreement), whereas the match between the person's cue weightings and veridical weightings is calculated to represent linear policy agreement.

The Lens Model also captures conflict between humans and artificial systems. The classic model shown in Figure 3 has a single utilization model constituting the right side of the lens. In *n*-system models [94,98] there are multiple judgment agents, each of which has its own utilization model. For example, a study investigated an aircraft collision detection task in which human judgment was compared with that of two artificial agents [94]. The Lens Model analysis pinpoints differences in cue weighting that may lead to conflict. It also provides a means for conflict resolution in that cognitive feedback on how each agent is weighting cues [92] supports rational analysis of weighting differences. Cognitive feedback demonstrating that the AI is using a rational process is likely to build performance-based trust [54].

In our research [81], we investigated how people integrate visible cues with those delivered by either a human or AI partner on the basis of sensor analysis. The participant is asked to determine whether a human figure is a security threat or not based on multiple cues (see Figure 4). The task is designed so that there is an objective "ground truth" for diagnosticity of threat, i.e., the probability of the target being a threat given any specific set of threat cues. Prior to the task, participants are trained on cue diagnosticity so that they can distinguish stronger and weaker cues. We found that people utilized both visible and partner-delivered cues fairly rationally, although various instances of sub-optimal cue utilization were identified. There was potential conflict between visible and partner-delivered cues. For example, a sensor might pick up physiological signs of aggression from a target that were invisible to the participant. For some specific cues, including aggression,

the partner-delivered cue was underweighted relative to the visible one. Overall, as in previous studies [79], people utilized cues fairly rationally through a weighting process, but the Lens Model analysis identified some specific misweightings that might lead to conflict and impairment in judgment.

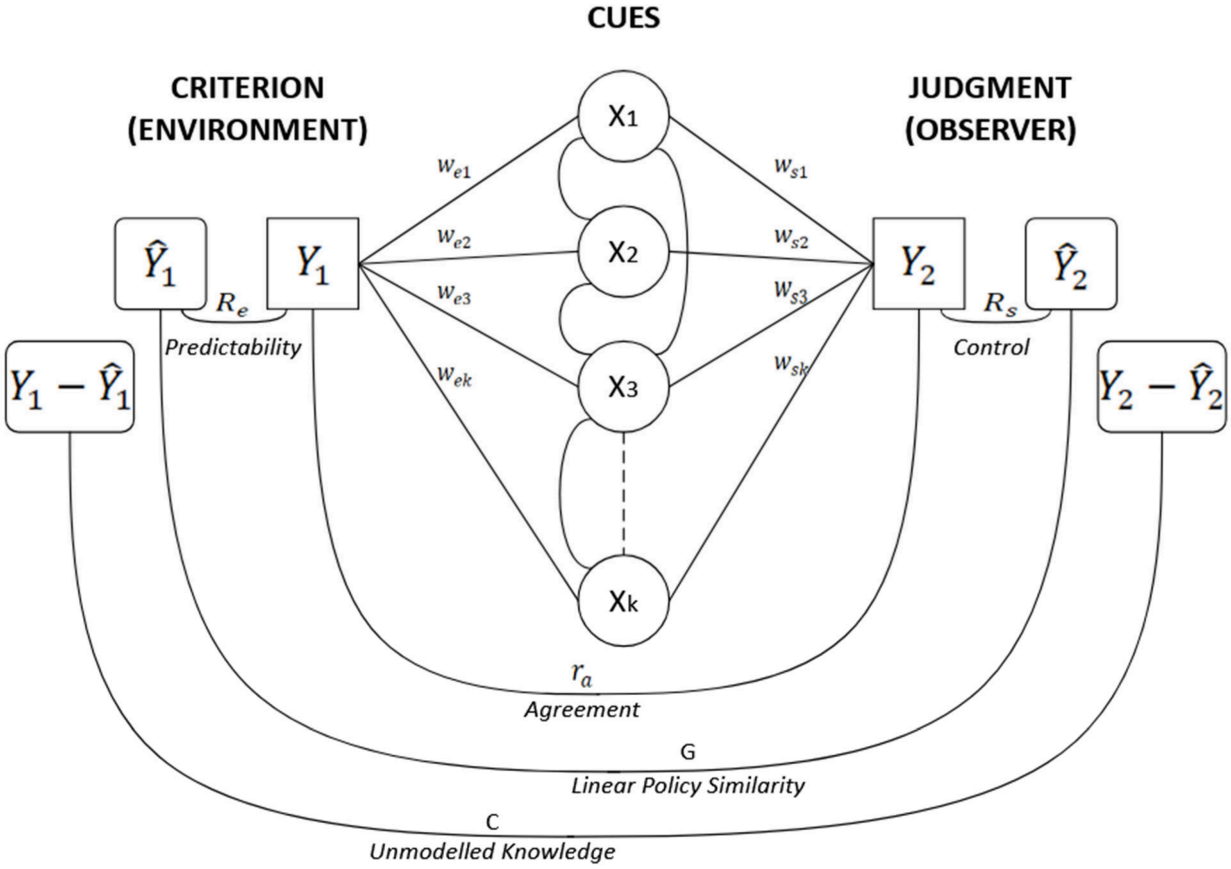

Figure 3. The Brunswik Lens Model for multi-cue discrimination.

People appeared to factor in cues delivered by AI more or less rationally, similar to other types of cues [79]. Where the compromise weighting in fact diverges from the true, ecological weighting, transparency and cognitive feedback may be used to enhance the performance of the human-AI team [94]. Over extended periods of interaction with an AI, a Bayesian updating process might operate [99], i.e., the person updates the threat/cue conditional probabilities as they gain experience with the collaboration. We would expect compromise assessments to move closer to ground truth as the person and the AI learn about the other's strengths, weaknesses, and biases. From this perspective, compromise with AI is thus "baked into" the cue weighting process, consistent with the idea of the brain as an inference machine that automatically builds and updates probabilistic models of the world [100].

5.2. Conflict Through Competing Narratives

A different kind of conflict is one in which two teammates construct qualitatively different narratives to explain current or anticipated events. Narrative is often critical to sensemaking and provides a framework for integrating multiple levels of contextual representation [101]. In applied contexts such as intelligence analysis, decision-making entails choosing between plausible competing narratives [88,102]. Indeed, the US intelligence community recognizes the value of constructing alternative narratives to test them against each other using contrarian techniques such as Devil's Advocacy and Team A/Team B

analysis [103]. By contrast with the Lens Model perspective, conflicts between competing narratives cannot be resolved by compromises that effectively "split the difference" between two perspectives such as averaging the weight assigned to a cue. Often, competing narratives will be fundamentally incompatible, and compromise requires teammates to recognize that one narrative is better supported by evidence-based argument than the other. The role of narrative reasoning has not been widely recognized in studies of human-AI teaming, but we can anticipate its role in conflict from various observations of human-AI verbal and human-human interaction.

Figure 4. Stimuli for a multi-cue threat assessment task. Cues are delivered visually in the images and through messages received from an AI partner ("SentinelAI") or a human partner ("Morgan").

A study of over 200,000 dialogues between humans and LLMs showed that people spontaneously shift from machine- to human-oriented language that personalizes the LLM [104]. Such a trend prefigures the increasing likelihood of humanlike conflicts as generative AI communicates sophisticated situational assessments in narrative form using natural language. Narrative reasoning commonly requires taking the perspectives of people involved in the narrative, empathic imagination, and social-cognitive abilities [105]. Thus, as specified in the conceptual model developed by Matthews and colleagues [60,78,80], attributing ToM to the robot or AI may make its narratives more credible to human partners.

There are several ways in which narrative-based conflicts may arise. In some contexts, such as intelligence analysis and accident investigation, constructing and justifying a narrative account of events is the overt task to be performed. Human and AI analysts may arrive at conflicting narratives. The AI might still be viewed as a tool in such a situation, supporting conflict resolution through further dispassionate evidence-based analysis of the problem using techniques developed by the intelligence community [103].

Such "adversarial collaboration" is facilitated by performance-based trust [54]. However, as human personalities and intentions are central to narrative reasoning [82], the conflict may become personalized and a humanlike mental model applied to the AI. For example, consider an intelligence analysis AI utilized by a nation institutionally hostile to the US. If the AI concluded that the US had benign intentions in some scenario, it might provoke negative interpersonal reactions from human analysts disposed to attribute harmful motives to the US, even being perceived as disloyal. The AI requires social skills such as use of RPL [27] to defuse such interpersonal conflicts.

In addition, differences in narrative reasoning can contribute to interpersonal conflicts between team members and AI systems. As previously discussed, several types of workplace conflict are interpersonal in nature, including relationship-based [12], process-based [16], and status-based [17] conflicts. Interpersonal conflicts depend on sensemaking processes such as interpreting people's expressions and behaviors and making attributions about others' beliefs and intentions [106]. That is, conflict sensemaking has a narrative quality. Conflicts in human-AI teams may reflect similar narrative sensemaking. The human might believe that the AI evaluates the human's contribution negatively (relationship conflict), that the AI has been assigned a role in the team better performed by a human (process-based conflict), or that the AI is usurping human leadership (status conflict).

We observed a possible instance of interpersonal conflict in our threat evaluation project [81]. We conducted a study using an immersive simulation in which the participant was required to evaluate urban scenes for threat based on multiple cues [107]. An intelligent robot partner provided its own threat evaluation. Scenes were designed so that in the majority of instances the robot disagreed with the human. We manipulated the robot's style of communication to examine factors that moderate the impact of disagreement on the human's trust and willingness to compromise in making a final threat evaluation. In experimental conditions, the participant had the option to engage in scripted dialogue with the robot that would provide justification for its differing evaluation. In one condition, the participant also heard robot inner speech that illuminated the thought process by which it arrived at is evaluation. Control conditions were a one-time transparency message, and a statement of benign intent. We expected that dialogue and inner speech would enhance trust and compromise, relative to control conditions but results showed the opposite. Dialogue and inner speech were detrimental to the collaboration. An explanation is that while the speech manipulations were designed to enhance performance-based trust, they actually had a stronger effect in eroding relationship-based trust. Evoking a humanlike mental model through speech and narrative dialogue may have suggested that the robot was obstructive as a teammate whereas in control conditions the advanced tool mental model was able to accommodate robot fallibility in a complex and uncertain environment.

These results highlight the role of context and interaction setting in determining effectiveness of inner speech in human-AI collaboration. In markedly different settings, such as interaction with a physical robot, inner speech has shown positive effects. In low-conflict cooperative settings with an embodied robot, inner speech has been shown to enhance perceptions of animacy and intelligence [108], stimulate practical reasoning in human partners [109], and support learning in high-stakes medical training [110]. Contextual factors including embodiment and whether the robot partner is cooperative or adversarial may influence what kind of humanlike mental model is activated, and hence inner speech impacts.

Extrapolating from human-human conflict resolution methods [26], a key factor to countering "interpersonal" human-AI conflict may be relation-based trust, i.e., trusting the AI or robot to be a sincere and ethical social agent [54]. Relation-based trust may facilitate translation of existing conflict mitigation methods [26] for human teams to mixed teams.

The importance of agent appropriateness has been shown in a study of simulated aerial dogfights in which pilots were teamed with an agent that provided flight directives [111]. Agent appropriateness is the extent to which the robot's behaviors are appropriate for its role in a specific context. It may counter tendencies to perceive the robot as a poor team player, a possible factor in our simulation study results [107]. Using an air traffic control simulation, it was found that humans were more likely to accept recommendations from artificial agents with higher levels of anthropomorphism and lower levels of rank [112]. These features promoted a sense of shared accountability and shared responsibility for team failures. Thus, a goal for future conflict-resolution research is to determine appropriate social norms for artificial systems, such as acknowledging a subordinate role, which encourage human partners to treat their advice as worthy of consideration, even when conflicting with the human's judgment.

6. Theory Testing and Development

The current theoretical analysis of conflict and compromise represents an initial attempt to synthesize concepts and evidence on human-AI collaboration. It is likely to require substantial refinement as evidence on the neglected issue of human-AI conflict in teaming accumulates. This section addresses theory testing and its challenges. Following a summary of the theory, we outline predictions from theory and discuss advances in methods necessary for its further development.

6.1. A Summary of Theoretical Issues

Table 1 summarizes the two types of conflict and compromise differentiated in the theory. The dominant model applied to the AI partner—advanced tool or humanlike—has multiple consequences. If tool, the interaction is governed by formal-quantitative algorithms which, on the human side, may be explicit or implicit. Conflicts will be technical in nature and resolvable through formal analysis of evidence, as we illustrated in the case of the Brunswik Lens Model analysis. Constructive compromise between human and AI is then a technical exercise in optimizing choice and use of algorithms, e.g., through averaging the cue weightings of the two entities in the Lens Model. Trust is exclusively performance-based and system design should aim to promote mutual understanding of the analytic methods utilized by each partner. Similarly, training for the collaboration should promote such understanding, e.g., through sharing feedback on judgment and decision-making.

Table 1. Two types of conflict in human-AI teaming.

	Mental Model for AI Partner	
	Advanced Tool	**Humanlike Partner**
Type of reasoning	Formal-quantitative	Narrative
Type(s) of human-AI conflict	Task-based	Relationship-based Team role (status and appropriateness)
Paths to compromise	Analysis of evidence to optimize algorithm choice and implementation	Discourse based on perspective-taking to resolve competing narratives
Type(s) of trust supporting compromise	Performance-based trust	Performance-based trust and relationship-based trust
Design features to optimize trust	Transparency and explainability	Humanlike communication and attributions of ToM
Training strategies	Task information feedback	Instruction based on humanlike understanding of the trainee's needs

If a humanlike mental model is activated, any formal-quantitative reasoning will be embedded in a wider narrative account of the problem or events defining the task. In some instances, conflicts over narratives may be purely task-oriented as occur in intelligence analysis. More disruptive conflicts may arise when humanlike relationships break down, i.e., when the AI's narrative prompts attributions of it being an unsupportive or disruptive teammate. For example, the AI's statements may imply that it wants to take on roles or authority that human partners may reserve for themselves. Compromise requires both performance- and relationship-based trust. That is, the human must believe that the AI's narrative is a good-faith reasoned effort based on sincere motivations to accomplish team goals that is sensitive to the human's intentions and values. Design of the AI with humanlike skills for negotiating interpersonal conflict supports compromise and training methods acknowledge the motives, values and skills of the other partner.

Theory testing can address both the interrelationships of key constructs and causal modeling of the conflict process. Table 1 expresses two modes of human interaction with AI that cohere around the two mental models. Broadly, we expect that mental model usage will be correlated with the type of trust, reasoning, and conflict associated with the model, but some associations will be moderated by perceptions of the AI as positive or negative. Primary predictions from the model are expressed in Table 2. They can be tested against between-subject correlations using interaction scenarios in which there is sufficient inter-individual variation for correlations to be meaningful, i.e., the scenario does not force participants to adopt either a tool or humanlike mental model of the partner. The theory predicts positive intercorrelations of level of usage of the tool mental model, use of formal-quantitative reasoning, and conflicts of a technical or analytical nature where they arise. Similarly, usage of humanlike mental model, use of narrative reasoning, and incidence of personalized conflicts should be intercorrelated. Correlates of trust are moderated by the person's attitudes towards the AI; e.g., tool model usage correlates positively with performance trust when attitudes are positive but negatively when attitudes are negative. These predictions are also testable against within-subject correlations in scenarios in which the person switches between mental models within or across interaction sessions.

Table 2. Primary predictions for types of trust, reasoning, and conflict elicited by tool and teammate mental models, moderated by attitudes to the AI.

	Attitude	**Trust**	**Reasoning**	**Conflict**
Tool	Positive	High, performance-based	Formal-quantitative	Low
	Negative	Low, performance-based	Formal-quantitative	Analytical disagreement
Teammate	Positive	High, relation-based	Narrative	Low
	Negative	Low, relation-based	Narrative	Personalized disagreement

These primary predictions are qualified by several nuances of human-AI interaction previously discussed. First, in ergonomics contexts for JDM, performance is always important, even when a humanlike mental model is activated. Thus, while relationship trust is always linked to activation of the humanlike model (positively or negatively), activation of the humanlike model may also be associated with performance trust, e.g., if AI teamwork is perceived as enhancing taskwork and performance outcomes. Second, there may be transfer of conflict types as in human-human teams [12]; a personalized conflict may spill over into technical disagreements so that performance-based and relation-based conflict are correlated. Third, in some contexts, such as intelligence analysis [88], narrative reasoning is consistent with a tool mental model, reducing the mental model—reasoning type

correlation. Thus, Table 2 expresses broad expectations; relationships between constructs may influenced by context.

A second type of theory testing employs experimental manipulations to test hypotheses from causal models such as that shown in Figure 1. This model predicts that (1) inducing a humanlike mental model leads to increased narrative reasoning and (2) providing the AI with the capability for dialogue increases relationship trust. The Figure 1 model is plausible but tentative, in that other causal orderings of the constructs might be envisaged, including reciprocal relationships. For example, AI use of narrative dialogue might activate a humanlike mental model. As for correlational tests, causal relationships between variables may be somewhat context-dependent. Testing the theory thus requires both valid measures of the key constructs and manipulations for experimental studies. Challenges to theory testing include limitations in existing methodology, as well as some broader conceptual issues. Next, we summarize methods for manipulating and measuring four key theoretical constructs, listed in Table 3, identifying research gaps where they exist. We also briefly review various general issues that future development and application of the current theory should address.

Table 3. Four theoretical constructs: manipulations and measures.

Construct	Manipulation	Measure
Mental Model	Task instructions Anthropomorphic design	Self-report Situation Judgment Test Implicit tests
Trust	System competencies Priming Statements of benign intent	Self-report Situation Judgment Test Implicit tests
Reasoning	Problem type Prompted discourse Inner speech	Qualitative analysis
Conflict	Prompted disagreement	Qualitative analysis Attribution ratings

6.2. Relationship to Existing Theory and Research

The current theory was derived by analyzing the alignments of constructs from the four diverse and often disconnected research areas of mental models, trust, reasoning, and conflict. We took as a starting point a set of binary distinctions that are generally accepted and well-supported by research [12,54,60,82], while acknowledging their simplified nature. For example, the tool vs. teammate dichotomy may be further elaborated to define different subtypes of tools and teammates. The novelty of the theory resides in its mapping of constructs as illustrated in Table 1, especially those linked to humanlike AI functioning.

Much research on conventional automation implicitly assumes a tool mental model, performance-based trust, and formal-quantitative reasoning. Prior to social AI, there was no need to make these constructs explicit. Thus, predictions of associations between constructs in the "advanced tool" column on Table 1 are fairly straightforward extrapolations from the existing literature. For example, although most trust research has not differentiated performance and relation-based trust, evidence for trust being associated with activation of a positive tool-based mental model [29,32] presumably reflects performance-based trust. However, while there is a solid research base for influences on humanlike mental model activation, such as anthropomorphic design [57], evidence is lacking on how the mental model shapes reasoning and conflict. Thus, the predicted associations between the constructs in the "humanlike partner" column of Table 2 are both more novel and more

tentative than those linked to the advanced tool model. Generally, the theory aims to build out from existing performance-focused models of human-automation interaction [28,32,33] to accommodate the novel conflicts that will arise when people treat as humans AIs with views contrary to their own.

6.3. Mental Models

In experimental research, task instructions may explicitly frame the AI as either team member or tool [113]. Lyons and Wynne [114] used two different instruction sets for working with a personal digital assistant to manipulate perceptions of humanness. One set described how the assistant used a complex machine learning algorithm to predict the user's needs; the other highlighted humanlike features such as sharing emotions and acting as a friend. However, people have limited conscious awareness of their mental models [78,115] and humanlike design may activate the mental model implicitly [116]. Multiple design features promote perceptions of robots and virtual systems as human, i.e., activation of a human teammate mental model. As discussed in Section 4, there is a large literature on design for anthropomorphism [51,117,118]. Anthropomorphic features, adaptive intelligence and social-emotional recognition and signaling are factors that evoke perceptions of human teammate likeness [62]. Other factors eliciting teammate over tool perceptions can be allocated to three groups: team setting (e.g., shared goals), social entity (e.g., relationship building), and collaboration behavior (e.g., explanation provision) [119]. Experimental manipulations of such kind are assumed to influence the mental model [63].

To date, research on measuring mental models has focused on models for cognitive tasks rather than for entities (e.g., machine or human). Methods include post-task assessments such as self-reports and elicitation techniques that expose the person's reasoning as they interact with the AI [120,121]. Studies of humanness of the mental model have typically used post-task assessments. Building on their 2018 review, Wynne and Lyons [122] developed a self-report scale for perceptions of Autonomous Agent Teammate-Likeness (AAT), with multiple subscales for constructs including perceived agentic capability, perceived benevolence/altruistic intent, perceived task interdependence, task-independent relationship-building, richness of communication, and synchronized mental model. A possible limitation of the measure is that activation of an advanced tool mental model might also influence performance-focused subscales such as perceived agentic capability. For example, an autonomous delivery road vehicle that could determine the optimal route for a delivery and respond adaptively to traffic incidents along the route, could still be perceived as a tool rather than a teammate.

Measures that more sharply distinguish alternate mental models are needed. One approach is to develop Situation Judgment Tests (SJTs) [123] that require the person to rate AI functioning in short scenarios. For example, in the threat assessment context, we identified separable (though correlated) dimensions for perceptions of humanlike functioning in robots acting either as conventional machines (e.g., detecting chemical traces) or as human (e.g., detecting aggression) [78]. SJT scenarios may automatically activate the relevant mental model, countering limitations of conscious awareness for measuring mental models [116]. Recent work on EEG response to anthropomorphism [124] indicates the potential for developing neuroergonomic measures of mental model activation.

6.4. Trust

As discussed in Section 3, there is extensive research on trust in digital systems that typically does not distinguish between performance-based and relation-based trust [44]. Lack of definition of trust in research studies obscures whether the agent is being positioned as a tool or a teammate [55]. Manipulations to selectively affect one or other

type of trust emerge from Mayer et al.'s [37] analysis of influences on trustworthiness, contrasting ability with benevolence and integrity as relationship-based factors. Selective manipulation of performance trust is straightforward; objective competencies such as system capability and reliability strongly influence trust [35], presumably performance-based trust. There has been less attention paid to manipulations of relationship-based trust but social competencies and attributes such as gender [44] are expected to elevate trust. Other methods for influencing antecedents of relationship trust such as perceived benevolence include priming statements about the AI's motives [125] and expressions of interpersonal warmth [126]. The integrity of AI can be manipulated via explanations focusing on either honesty, transparency, or fairness, with positive effects on (presumably) relationship trust [127].

In parallel with the need for clearly distinct trust type manipulations, theory testing requires scales that discriminate performance-based and relationship-based trust [44,55]. There have been limited efforts of this type [55,56] but there is currently no "gold standard" scale that captures the distinction. Existing human-human trust scales that focus on integrity, benevolence, and ethics can be adapted to contexts for AI. For example, a modified version of Mayer and Davis' [128] trust scale was used to compare delivery drivers' perceptions of ability, integrity and benevolence in AI and human managers [129]. AI managers were perceived as less benevolent than human managers. Benevolence predicted reduced trust, over and above ability and integrity, implying a specific deficit in relationship trust. Additional findings suggested that AI was especially mistrusted in situations calling for empathy from managers. Similarly, it is reported that AI ability is rated higher than its benevolence across multiple domains [130].

The great majority of studies of human-machine teaming assess trust through self-report. The wider literature on trust assessment has investigated alternatives to subjective response such as SJTs [78] and implicit measures based on speed of response to trust and distrust stimuli [131,132]. The capacity of such measures to differentiate performance- and relationship-based trust has yet to be determined.

6.5. Type of Reasoning

The methodology for evoking and measuring different types of reasoning in ergonomics has been neglected. Consequently, this section primarily identifies research needs rather than established methods. Informally, we can require participants to solve different types of problem requiring different forms of reasoning including applying formal reasoning or quantitative algorithms vs. building an evidence-based narrative, as in intelligence analysis [88]. Qualitative methods may be applied to assess the type(s) of reasoning utilized during the human-AI interaction. Raters can be trained to distinguish formal-quantitative reasoning from narrative reasoning identified from the triad of causality, temporality, and character [82]. A study of this kind used narrative analysis to distinguish different forms of clinical reasoning used by resident doctors including diagnostic reasoning based on symptoms (seemingly formal-quantitative) and narrative reasoning to understand the individual's life circumstances [133].

Beyond qualitative manipulations and measures, recent developments in AI suggest future methodological innovations. Computational linguistics methods can identify markers for everyday narrative human conversation such as use of personal referents [104]. Yu, Zhang, Tiwari and Wang [134] presented a taxonomy of different forms of reasoning used by AI performing natural language processing, including formalized reasoning types such as classical logical reasoning. Yu et al. did not include narrative reasoning in their taxonomy, but the approach demonstrates how, for example, an LLM might be prompted to express its reasoning in either logical or narrative form so that the impact of reasoning

type on mental model activation and trust could be tested. AIs can have capabilities for narrative understanding, at least in relation to well-defined tasks such as checking narrative consistency and extracting the plot of a narrative [135]. The robot inner speech research previously referenced [109] also used the robot's narrative account of its understanding of human instructions to convey a ToM.

6.6. Conflict and Its Resolution

The human-human conflict literature cited previously commonly includes qualitative analyses to distinguish different forms of conflict. Concept mapping and multidimensional scaling were used to identify categories of conflict resolution strategies [26]. Similar methods could also be used to determine whether a person conceptualized a conflict with an AI as a technical disagreement or one based on negative perceptions of the AI's motives or benevolence. As in [26], we might find a richer discrimination of conflict perceptions than the two just mentioned. We noted the empathy gap found in delivery drivers' perceptions of AI managers [51]. The nature of conflicts with such an AI manager could be explored by having drivers provide narrative accounts of disagreements, similar to [51]. As described in the previous section, generative AI can engage in narrative reasoning [135]. Other research has identified metrics for assessing the quality of AI reasoning in various contexts (e.g., [136,137]). It is thus feasible to have an AI collaborator provoke disagreement through delivering either formal-logical or narrative poor-quality advice, as an experimental manipulation. Impacts on mental models, use of narrative discourse, and trust could then be investigated.

Theories of human-human conflict suggest quantitative measures that may be used to complement qualitative investigations. Attribution theory [138] provides an example. Is seemingly bad advice from an AI attributed to technical deficiencies in its analysis or to humanlike bad intent? For example, a disaffected delivery driver might believe that an AI was prompting them to make an error to give the company a pretext for firing them. Unwanted AI coworker behaviors might be attributed either to the coworker themselves, such as malign intent, or to situational factors such as lack of organizational support for the coworker. The tool mental model implies situational attributions, i.e., AI errors reflect poor design or usage of the system in an inappropriate context, whereas the humanlike mental model may elicit personlike attributions of the AI acting from unsupportive or even hostile motives. Researchers are beginning to apply attribution theory to human-AI interactions [139], research that may be extended to investigating conflicts.

7. Challenges and Future Research Directions

The preceding sections indicate methodological advances needed for further theory development. In addition, various existing controversies in human-AI collaboration impinge on the theory. In this section we briefly highlight some of these issues and priorities for future research.

- *Refining accounts of mental models.* The concept of mental models comes from studies of how people represent performance task environments, e.g., in relation to representations and procedures [140,141]. Using the concept to differentiate entities such as humans and AIs is a considerable extension of theory and raises several challenges. The tool vs. teammate mental model distinction has become quite widely accepted in the ergonomics literature [60,141,142]. However, both tools and teammates may be represented by multiple mental models, including positive and negative models. Some tools work well; others are poorly designed. Some can be used intuitively with minimal training whereas others require training and expertise to use effectively. Different AI-powered applications may elicit different tool mental models. Mental models for specific systems

may encode AI capabilities, such as error vulnerabilities [143]. Similarly, in human-human work teams, people use a range of mental models to structure their interactions with co-workers, including conflicts [144]. Humanlike teammates may vary in multiple respects, including expertise, motives, authority, and application domain. The mental model elicited by a sympathetic healthcare advisor might be different to a hard-driving AI manager, for example. Current research has investigated the role of human-relevant factors such as authority and agency in attitudes towards AI (e.g., [145]), but has not yet differentiated discrete mental models for humanlike AI. The research need is to identify different types of bad AI teammates that may define mental models activate during humanlike conflict. Conflicts and their resolution may unfold differently for AIs in different roles as they do in human-human conflict.

Compounding the challenge, identifying mental models is likely to be a moving target as technology and cultural attitudes evolve. For example, contemporary accounts of design features that elicit perception of humanness include sophisticated tasking capabilities such as adaptive intelligence [62,119]. However, as advanced machine intelligence and autonomy become commonplace, humanlike intelligence may become weaker as a discriminant for tools and teammates. Elaboration of mental models for AI partners may be necessary to address the manipulation and measurement issues discussed in Section 6.3.

- *Accommodating dynamic factors.* Testing causal models requires better understanding of the temporal dynamics of mental model activation. Current research offers various perspectives on temporal factors including accounts of learned trust reflecting a history of interaction with the system [32], evidence for differential changes in trust for physical and virtual systems [146], factors influencing trust repair after a violation [68], and deepening attachment to AI companions over time [57]. Dynamic processes underpinning conflict and trust require further investigation. For example, existing research on trust violations and trust repair [68] could be extended to investigate whether changes in trust are moderated by mental model activation, and how mental models can be guided through design to optimize the repair process. Dynamic factors are especially important for applications more complex than a single user interacting with an AI that has well-defined and constrained functionality such as a decision support system. A copilot or wingman AI capable of supporting multiple human activities might switch autonomously between tool and humanlike modes. Multi-agent teams present similar issues. Further theory development is needed to predict and optimize mental model activation in conflicts arising in such contexts.

Investigating dynamic factors would also benefit from measures that distinguish stable, trait-like predispositions to adopt one or other model and the immediate model dominant at any one moment in time. The mental model itself is likely to change through continuing interaction with the AI. In typical ergonomics teaming contexts it is unknown whether the mental model activated by the AI will remain stable within and across collaboration episodes or change frequently. Stability is preferable from both a design and research perspective; lacking a means for tracking mental model activation in real-time, frequent model changes will make testing causal models difficult indeed. Possibly, psychophysiological methods used to identify shared mental models in teaming contexts [147] might be adapted to identify mental model change at the individual level.

- *Context-dependence.* AI is becoming pervasive across multiple application domains including industry, customer, service, military/security, healthcare, transportation, social media, and leisure. Existing research on human-machine interaction and trust highlights how findings may vary across application domain. For example, people have stronger pro-social attitudes towards medical than security robots [148] and anthropomorphism is preferred more in the social domain than in industry [149]. There are also differences in how people assign trust to virtual systems powered by AI and to embodied systems

such as robots, including differing temporal trajectories [146]. Contextual factors that act as moderators for attitudes towards AI include whether the system is providing epistemic or moral advice, and whether it is in an autonomous or supportive role [150,151]. In keeping with sociotechnical approaches to human-AI interaction [152], it is important also to investigate how organizational and cultural factors may impact conflict. Contextual influences on performance expectations will affect performance trust (cf., [153]. Factors affecting perceptions of AI benevolence and ethics, such as an organization's motives for introducing AI, will influence relationship trust [146]. Multi-agent human-AI teams raise further research questions such as the propagation of trust and distrust between team members [154].

In keeping with some of the foundational research cited in the present article [35,62] the current theory draws on studies from the military/security domain [79,107], which may limit its generalization to other types of application. Thus, further theory development requires tests for context-dependence of relationships between constructs. At the same time, some relevant attitudes towards AI such as doubts about its benevolence [130] and preference for supportive over autonomous AI [150] generalize across application domains. The present theory draws on general constructs applicable to multiple domains. The optimal balance between general and domain-specific theorizing remains to be determined.

- *Addressing cognitive illusions in JDM.* Cognitive biases are a general threat to both individual and collaborative decision-making [84]. From a formal-quantitative perspective, LLMs as well as humans appear to be sensitive to biases associated with use of heuristics in JDM. These include well-known biases such as confirmation, representativeness and anchoring biases [84]. A disturbing recent finding is that even when collaboration with an AI improves human decision-making it appears to impair metacognitions of performance [155]. People overestimate their performance regardless of actual competence, implying possible over-trust in the AI and vulnerability to false consensus. Thus, design for formal-quantitative conflict resolution should accommodate safeguards for possible shared illusions between human and AI partners. Similarly, conflicts driven by narrative reasoning may be resolved suboptimally, e.g., through authority bias [156] if the human accepts too quickly a false narrative from an AI in an authority position.

8. Conclusions: Implications for Design

The current theory represents an initial attempt to integrate multiple perspectives on conflict relevant to human-AI collaboration for JDM. As discussed, further research is necessary to bring the theory to maturity. Similarly, it is not yet developed to the point where it can support explicit design recommendations, such as what level of system anthopomorphism is optimal for a given task context and user. More research is needed on designing AI systems that function like human teammates that are able and motivated to work through reasonable disagreements constructively. Thus, we conclude with some general suggestions on how explicit consideration of mental models may facilitate the design process to reduce the incidence and damaging impacts of human-AI conflicts.

Design philosophies for AI are already guided in part by the tool vs. teammate distinction [141]. Designing AI systems to be humanlike can enhance mixed-team capabilities for performing complex tasks in uncertain environments. There is a large literature, beyond the present scope, on design principles for teamworking functions such as appropriate coordination of activities and naturalistic, context-sensitive communication (e.g., [157]). However, some designers consider that the tool metaphor remains more appropriate [141]. In particular, Schneiderman [158] argued that design goals can include both emulating humans and building specific applications, corresponding to the teammate vs. tool distinction. He claimed the tool metaphor is often the more appropriate one for building functional

systems and products that remain under human control, in line with human-centered design principles.

We conclude with a summary of implications of the theory for designing for human-AI conflict resolution. Awareness of mental models in design can minimize the human-AI team's vulnerability to conflict, detect conflict when it does arise, mitigate or resolve conflicts effectively, and personalize design to match user preferences and competencies.

- *Minimizing vulnerability to conflict.* Design may seek both to reduce the likelihood of conflict and to support conflict resolution when it arises. Conflict may be produced by mismatch between the user's mental model and the designer's intentions, e.g., if the user expects a tool but the designer has included humanlike interaction features. AI attempts at friendly conversation may be perceived as annoying or distracting. Anthropomorphic design in industrial robots may damage trust and performance [149]. AIs that pretend to have human traits they actually lack also disrupt the user's ability to assign trust and responsibility to the system [159]. In other contexts, such as healthcare, users may expect a human touch from the AI and lack of empathy may be disturbing [160]. To avoid such conflicts, designers should make an explicit and informed choice about the preferred user mental model early in the design cycle, and include features to activate it consistently, based on the existing literature. The mental model choice should also inform training and task instructions.
- *Detection and analysis of conflict.* From a tool perspective, design should help the user understand why the AI is reaching a different conclusion to them. Existing literatures on transparency [161,162] and explainability [45,46] provide design principles. Design can also facilitate acquisition of shared mental models for task goals and execution, reducing the likelihood of technical disagreements (see [120] for a review of AI design considerations). A sophisticated system might also identify conflict caused by the user inappropriately applying a humanlike mental model to the AI. That is, the AI requires "AI-of-human" transparency [40], e.g., identifying and understanding human mental states such as frustration or confusion. That is, even a system operating as a tool may benefit from humanlike perceptions of its teammate. Conflict resolution could take the form of messaging to refocus the human on formal-quantitative reasoning to address the technical issues under debate.
- *Mitigation of conflict: cognitive architectures for AI.* Conflict mitigation depends on both the human's ability to understand machine reasoning, and the AI's ability to analyze the human's reactions, including emotional reactions. We have already discussed design solutions for mitigating technical disagreements, such as transparency and explainability (Section 3) and application of the *n*-system Brunswik Lens Model (Section 5.1). We also discussed adapting strategies for human-human conflict resolution [26] for conflicts with AI (Section 2.3). AI capabilities to frame differing perspectives in narrative form may contribute to conflict resolution but current AIs can lack the social skills to do so. Our findings on the counterproductive effects of attributing ToM to a robot via inner speech [107] illustrate this issue.

Designing cognitive architectures to support conflict mitigation is thus essential. Human-AI cooperation requires adaptation to the type of conflict at hand and the cognitive and affective state of the human partner, so that it is seen as a cooperating partner rather than a rival. Any AI system that makes its reasoning process transparent to the human partner (whether through inner speech or natural language explanations) requires dynamic calibration between what to reveal and how to frame it, informed by a model of the partner's state and the nature of the conflict. This points to the need for cognitive architectures that can dynamically calibrate how much internal reasoning to externalize and in what form (technical and data-driven when the conflict is formal-quantitative, or socially

sensitive and perspective-aware when the conflict is narrative). A cognitive architecture integrating inner speech and ToM [163,164] enables complementary functions: inner speech renders the conflict visible to the human and, by externalizing the AI's reasoning process, can guide the AI's own ToM, informing its model of the partner's cognitive and affective state. In turn, ToM renders the conflict manageable by the AI, enabling the architecture itself to mediate the balance between transparency and social sensitivity according to the demands of the interaction.

- *Personalization of AI.* In some contexts design gives users the option of adopting a tool or teammate mental model through personalizing AI functioning. Such an approach complicates design but compatibility between system function and the user's preferred mental model improves user experience and the "hedonomics" [165] of the system. Personalization can also enhance trust calibration by accommodating operators' past experiences and dispositions [55] Researchers have identified various personal characteristics that influence trust in AI [166]. Our research [78,79] suggests a differentiation between factors related to perceptions of tools, such as the perfect automation schema [69], and humanlike perceptions such as emotional reactions to interacting socially with a robot [20]. Identifying mental model preferences may support personalization of design [167–169]. Preferences may vary with the application domain; e.g., a user might prefer a toollike system in an industrial setting but a humanlike system in healthcare. Thus, as for other applied issues, sensitivity to context is important for personalization of the mental model.

Author Contributions: All authors listed have made a substantial, direct and intellectual contribution to the work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Air Force Office of Scientific Research Trust and Influence program, grant number FA9550-22-1-0035.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: We are grateful to April Rose Panganiban of the Air Force Research Laboratory for contributions to this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shaffer Shane, T.; Mylius, S.; Hobbs, H. *Scheming in the Wild: Detecting Real-World AI Scheming Incidents with Open-Source Intelligence*; The Centre for Long-Term Resilience: London, UK, 2026.
2. Babel, F.; Kraus, J.M.; Baumann, M. Development and testing of psychological conflict resolution strategies for assertive robots to resolve human–robot goal conflict. *Front. Robot. AI* **2021**, *7*, 591448. [CrossRef] [PubMed]
3. Ansari, K.; Ghasemaghaei, M. Partner or rival? How human-Artificial Intelligence conflict shapes artificial intelligence aversion. *J. Manag. Inf. Syst.* **2026**, *43*, 368–401. [CrossRef]
4. Mondrus, O.; Tov-Nachlieli, I.S. Predicting decision-makers' coping strategies in human-AI conflict based on the link between personality traits and cognitive AI appraisals: A quantitative study of US employees using AI at work. *Comput. Hum. Behav. Rep.* **2025**, *20*, 100857. [CrossRef]
5. Wen, H.; Amin, M.T.; Khan, F.; Ahmed, S.; Imtiaz, S.; Pistikopoulos, S. A methodology to assess human-automated system conflict from safety perspective. *Comput. Chem. Eng.* **2022**, *165*, 107939. [CrossRef]
6. Wen, H.; Khan, F. A risk-based model for human-artificial intelligence conflict resolution in process systems. *Digit. Chem. Eng.* **2024**, *13*, 100194. [CrossRef]
7. Bruiger, D. Reflections on the AI alignment problem. *AI Soc.* **2025**, *40*, 4383–4392. [CrossRef]
8. Pan, Y.; Xu, J. Human-machine plan conflict and conflict resolution in a visual search task. *Int. J. Hum.-Comput. Stud.* **2025**, *193*, 103377. [CrossRef]

9. McNeese, N.J.; Demir, M.; Cooke, N.J.; She, M. Team situation awareness and conflict: A study of human–machine teaming. *J. Cogn. Eng. Decis. Mak.* **2021**, *15*, 83–96. [CrossRef]
10. Brachman, M.; Ashktorab, Z.; Desmond, M.; Duesterwald, E.; Dugan, C.; Joshi, N.N.; Pan, Q.; Sharma, A. Reliance and automation for human-ai collaborative data labeling conflict resolution. *Proc. ACM Hum.-Comput. Interact.* **2022**, *6*, 1–27. [CrossRef]
11. Wang, Z.H.; Xu, X.H.; Zhao, C.W. Disuse and misuse of AI-generated suggestions: Examining human-AI conflicts in AI-assisted decision-making. *IEEE Trans. Syst. Man Cybern. Syst.* **2026**, 1–13. [CrossRef]
12. de Wit, F.R.; Jehn, K.A.; Scheepers, D. Task conflict, information processing, and decision-making: The damaging effect of relationship conflict. *Organ. Behav. Hum. Decis. Processes* **2013**, *122*, 177–189. [CrossRef]
13. Yuan, Z.; Yin, J.; Sun, J. The paradox of team conflict revisited: An updated meta-analysis of the team conflict–team performance relationships. *J. Appl. Psychol.* **2025**, *111*, 195–224. [PubMed]
14. Salas, E.; Cooke, N.J.; Rosen, M.A. On teams, teamwork, and team performance: Discoveries and developments. *Hum. Factors* **2008**, *50*, 540–547. [CrossRef] [PubMed]
15. Grossman, R.; Nolan, K.; Rosch, Z.; Mazer, D.; Salas, E. The team cohesion-performance relationship: A meta-analysis exploring measurement approaches and the changing team landscape. *Organ. Psychol. Rev.* **2022**, *12*, 181–238.
16. Jehn, K.A. A qualitative analysis of conflict types and dimensions in organizational groups. *Adm. Sci. Q.* **1997**, *42*, 530–557. [CrossRef]
17. Bendersky, C.; Hays, N.A. Status conflict in groups. *Organ. Sci.* **2012**, *23*, 323–340. [CrossRef]
18. Itoh, M.; Flemisch, F.; Abbink, D. A hierarchical framework to analyze shared control conflicts between human and machine. *IFAC-PapersOnLine* **2016**, *49*, 96–101. [CrossRef]
19. Nomura, T.; Kanda, T.; Suzuki, T. Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. *AI Soc.* **2006**, *20*, 138–150.
20. Nomura, T.; Kanda, T.; Suzuki, T.; Kato, K. Prediction of human behavior in human-robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE Trans. Robot.* **2008**, *24*, 442–451. [CrossRef]
21. Złotowski, J.; Yogeeswaran, K.; Bartneck, C. Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *Int. J. Hum.-Comput. Stud.* **2017**, *100*, 48–54. [CrossRef]
22. Lyons, J.B.; Jessup, S.A.; Vo, T.Q. The role of decision authority and stated social intent as predictors of trust in autonomous robots. *Top. Cogn. Sci.* **2024**, *16*, 430–449. [PubMed]
23. Saunderson, S.P.; Nejat, G. Persuasive robots should avoid authority: The effects of formal and real authority on persuasion in human-robot interaction. *Sci. Robot.* **2021**, *6*, eabd5186. [CrossRef] [PubMed]
24. Kim, D.; Vegt, N.; Visch, V.; Bos-De Vos, M. How much decision power should (A) I have? Investigating patients' preferences towards AI autonomy in healthcare decision making. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024; pp. 1–17.
25. O'Neill, T.A.; McLarnon, M.J. Optimizing team conflict dynamics for high performance teamwork. *Hum. Resour. Manag. Rev.* **2018**, *28*, 378–394. [CrossRef]
26. Behfar, K.J.; Peterson, R.S.; Mannix, E.A.; Trochim, W.M. The critical role of conflict resolution in teams: A close look at the links between conflict type, conflict management strategies, and team outcomes. *J. Appl. Psychol.* **2008**, *93*, 170–188. [CrossRef] [PubMed]
27. Hsu, A.; Chaudhary, D. AI4PCR: Artificial intelligence for practicing conflict resolution. *Comput. Hum. Behav. Artif. Hum.* **2023**, *1*, 100002. [CrossRef]
28. Parasuraman, R.; Riley, V. Humans and automation: Use, misuse, disuse, abuse. *Hum. Factors* **1997**, *39*, 230–253. [CrossRef]
29. Ke, Y.H.; Yang, R.; Lie, S.A.; Lim, T.X.Y.; Ning, Y.; Li, I.; Abdullah, H.R.; Ting, D.S.W.; Liu, N. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: Simulation study. *J. Med. Internet Res.* **2024**, *26*, e59439. [CrossRef] [PubMed]
30. Bajaj, J.S.; Khare, A.; Rani, S. AI-based relationship support: A framework for conflict resolution and well-being. In Proceedings of the 2025 IEEE 7th International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 28–30 November 2025; pp. 1–5.
31. Aggrawal, S.; Magana, A.J. Teamwork conflict management training and conflict resolution practice via large language models. *Future Internet* **2024**, *16*, 177. [CrossRef]
32. Hoff, K.A.; Bashir, M. Trust in automation: Integrating empirical evidence on factors that influence trust. *Hum. Factors* **2015**, *57*, 407–434. [PubMed]
33. Lee, J.D.; See, K.A. Trust in automation: Designing for appropriate reliance. *Hum. Factors* **2004**, *46*, 50–80. [CrossRef]
34. Hancock, P.A.; Kessler, T.T.; Kaplan, A.D.; Stowers, K.; Brill, J.C.; Billings, D.R.; Schaefer, K.E.; Szalma, J.L. How and why humans trust: A meta-analysis and elaborated model. *Front. Psychol.* **2023**, *14*, 1081086. [CrossRef] [PubMed]
35. Schaefer, K.E.; Chen, J.Y.; Szalma, J.L.; Hancock, P.A. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Hum. Factors* **2016**, *58*, 377–400. [CrossRef] [PubMed]

36. Kaplan, A.D.; Kessler, T.T.; Brill, J.C.; Hancock, P.A. Trust in artificial intelligence: Meta-analytic findings. *Hum. Factors* **2023**, *65*, 337–359. [PubMed]
37. Mayer, R.C.; Davis, J.H.; Schoorman, F.D. An integrative model of organizational trust. *Acad. Manag. Rev.* **1995**, *20*, 709–734. [CrossRef]
38. Lyons, J.B.; Hamdan, I.A.; Vo, T.Q. Explanations and trust: What happens to trust when a robot partner does something unexpected? *Comput. Hum. Behav.* **2023**, *138*, 107473. [CrossRef]
39. Barg-Walkow, L.H.; Rogers, W.A. The effect of incorrect reliability information on expectations, perceptions, and use of automation. *Hum. Factors* **2016**, *58*, 242–260. [PubMed]
40. Lyons, J.B. Being transparent about transparency. In *AAAI Spring Symposium 2013*; AAAI Press: Palo Alto, CA, USA, 2013; pp. 48–53.
41. Chen, J.Y.; Lakhmani, S.G.; Stowers, K.; Selkowitz, A.R.; Wright, J.L.; Barnes, M. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theor. Issues Ergon. Sci.* **2018**, *19*, 259–282. [CrossRef]
42. Bhaskara, A.; Skinner, M.; Loft, S. Agent transparency: A review of current theory and evidence. *IEEE Trans. Hum.-Mach. Syst.* **2020**, *50*, 215–224. [CrossRef]
43. Stowers, K.; Kasdaglis, N.; Rupp, M.A.; Newton, O.B.; Chen, J.Y.; Barnes, M.J. The IMPACT of agent transparency on human performance. *IEEE Trans. Hum.-Mach. Syst.* **2020**, *50*, 245–253. [CrossRef]
44. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
45. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Front. Comput. Sci.* **2023**, *5*, 1096257. [CrossRef]
46. Chamola, V.; Hassija, V.; Sulthana, A.R.; Ghosh, D.; Dhingra, D.; Sikdar, B. A review of trustworthy and explainable artificial intelligence (XAI). *IEEE Access* **2023**, *11*, 78994–79015. [CrossRef]
47. Maarten Schraagen, J.; Kerwien Lopez, S.; Schneider, C.; Schneider, V.; Tönjes, S.; Wiechmann, E. The role of transparency and explainability in automated systems. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2021**, *65*, 27–31. [CrossRef]
48. Balliet, D.; Van Lange, P.A. Trust, conflict, and cooperation: A meta-analysis. *Psychol. Bull.* **2013**, *139*, 1090–1112. [CrossRef] [PubMed]
49. Panganiban, A.R.; Matthews, G.; Long, M.D. Transparency in autonomous teammates: Intention to support as teaming information. *J. Cogn. Eng. Decis. Mak.* **2019**, *14*, 174–190. [CrossRef]
50. Plagge, E.; Jucks, R. Just like a human? An experimental study on university students' perceptions of warmth and benevolence towards chatbots. *J. Media Psychol. Theor. Methods Appl.* **2025**. [CrossRef]
51. Hutson, J.; Plate, D. The algorithm of fear: Unpacking prejudice against AI and the mistrust of technology. *J. Innov. Technol.* **2024**, *38*, 1–17. [CrossRef]
52. Mikalef, P.; Conboy, K.; Lundström, J.E.; Popovič, A. Thinking responsibly about responsible AI and 'the dark side'of AI. *Eur. J. Inf. Syst.* **2022**, *31*, 257–268. [CrossRef]
53. Park, P.S.; Goldstein, S.; O'Gara, A.; Chen, M.; Hendrycks, D. AI deception: A survey of examples, risks, and potential solutions. *Patterns* **2024**, *5*, 100988. [CrossRef] [PubMed]
54. Law, T.; Scheutz, M. Trust: Recent concepts and evaluations in human-robot interaction. In *Trust in Human-Robot Interaction: Research and Applications*; Nam, C.S., Lyons, J., Eds.; Elsevier: San Diego, CA, USA, 2021; pp. 27–57.
55. Duan, W.; Flathmann, C.; McNeese, N.; Scalia, M.J.; Zhang, R.; Gorman, J.; Guo, F.; Zhou, S.; Hauptman, A.I.; Yin, X. Trusting autonomous teammates in human-AI teams—A literature review. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 26 April–1 May 2025; pp. 1–23.
56. Malle, B.F.; Ullman, D. A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction: Research and Applications*; Nam, C.S., Lyons, J., Eds.; Elsevier: San Diego, CA, USA, 2021; pp. 3–25.
57. Shu, C.; Lai, K.; He, L. Human-AI attachment: How humans develop intimate relationships with AI. *Front. Psychol.* **2026**, *17*, 1723503. [CrossRef] [PubMed]
58. Rouse, W.B.; Morris, N.M. On looking into the black box: Prospects and limits in the search for mental models. *Psychol. Bull.* **1986**, *100*, 349–363. [CrossRef]
59. Endsley, M.R. Situation awareness. In *The Oxford Handbook of Cognitive Engineering*; Lee, J.D., Kirlik, A., Eds.; Oxford University Press: New York, NY, USA, 2013; pp. 88–108.
60. Matthews, G.; Panganiban, A.R.; Lin, J.; Long, M.; Schwing, M. Super-machines or sub-humans: What are the unique features of trust in Intelligent Autonomous Systems? In *Trust in Human-Robot Interaction: Research and Applications*; Nam, C.S., Lyons, J., Eds.; Elsevier: San Diego, CA, USA, 2021; pp. 59–82.
61. Ososky, S.; Schuster, D.; Phillips, E.; Jentsch, F.G. Building appropriate trust in human-robot teams. In *Proceedings of the AAAI Spring Symposium on Trust in Autonomous Systems*; AAAI: Palo Alto, CA, USA, 2013; pp. 60–65.

62. Wynne, K.T.; Lyons, J.B. An integrative model of autonomous agent teammate-likeness. *Theor. Issues Ergon. Sci.* **2018**, *19*, 353–374. [CrossRef]
63. Bodamer, W.G.; Schoenmakers, S.; Nuñez Castellar, E.; IJsselsteijn, W.A. Users' mental models within human–AI interaction: A systematic scoping review. *AI Soc.* **2026**, 1–18. [CrossRef]
64. De Graaf, M.M.; Allouch, S.B. Exploring influencing variables for the acceptance of social robots. *Robot. Auton. Syst.* **2013**, *61*, 1476–1486. [CrossRef]
65. Salles, A.; Evers, K.; Farisco, M. Anthropomorphism in AI. *AJOB Neurosci.* **2020**, *11*, 88–95. [CrossRef] [PubMed]
66. Złotowski, J.; Proudfoot, D.; Yogeeswaran, K.; Bartneck, C. Anthropomorphism: Opportunities and challenges in human–robot interaction. *Int. J. Soc. Robot.* **2015**, *7*, 347–360.
67. Waytz, A.; Cacioppo, J.; Epley, N. Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspect. Psychol. Sci.* **2010**, *5*, 219–232. [CrossRef] [PubMed]
68. De Visser, E.J.; Peeters, M.M.; Jung, M.F.; Kohn, S.; Shaw, T.H.; Pak, R.; Neerincx, M.A. Towards a theory of longitudinal trust calibration in human–robot teams. *Int. J. Soc. Robot.* **2020**, *12*, 459–478.
69. Merritt, S.M.; Unnerstall, J.L.; Lee, D.; Huber, K. Measuring individual differences in the perfect automation schema. *Hum. Factors* **2015**, *57*, 740–753. [CrossRef] [PubMed]
70. Leite, I.; Pereira, A.; Mascarenhas, S.; Martinho, C.; Prada, R.; Paiva, A. The influence of empathy in human–robot relations. *Int. J. Hum.-Comput. Stud.* **2013**, *71*, 250–260. [CrossRef]
71. Premack, D.; Woodruff, G. Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* **1978**, *1*, 515–526. [CrossRef]
72. Banks, J. Theory of mind in social robots: Replication of five established human tests. *Int. J. Soc. Robot.* **2020**, *12*, 403–414.
73. Rotenberg, K.J.; Petrocchi, S.; Lecciso, F.; Marchetti, A. The relation between children's trust beliefs and theory of mind abilities. *Infant Child Dev.* **2015**, *24*, 206–214. [CrossRef]
74. Chella, A.; Pipitone, A.; Morin, A.; Racy, F. Developing self-awareness in robots via inner speech. *Front. Robot. AI* **2020**, *7*, 16. [CrossRef] [PubMed]
75. Geraci, A.; D'Amico, A.; Pipitone, A.; Seidita, V.; Chella, A. Automation inner speech as an anthropomorphic feature affecting human trust: Current issues and future directions. *Front. Robot. AI* **2021**, *8*, 620026. [CrossRef] [PubMed]
76. Chella, A.; Dindo, H.; Infantino, I. A cognitive framework for imitation learning. *Robot. Auton. Syst.* **2006**, *54*, 403–408. [CrossRef]
77. Vinanzi, S.; Patacchiola, M.; Chella, A.; Cangelosi, A. Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philos. Trans. R. Soc. B* **2019**, *374*, 20180032. [CrossRef]
78. Matthews, G.; Lin, J.; Panganiban, A.R.; Long, M.D. Individual differences in trust in autonomous robots: Implications for transparency. *IEEE Trans. Hum.-Mach. Syst.* **2019**, *50*, 234–244. [CrossRef]
79. Lin, J.; Panganiban, A.R.; Matthews, G.; Gibbins, K.; Ankeney, E.; See, C.; Bailey, R.; Long, M. Trust in the danger zone: Individual differences in confidence in robot threat assessments. *Front. Psychol.* **2022**, *13*, 601523. [CrossRef] [PubMed]
80. Matthews, G.; Hancock, P.A.; Szalma, J.L.; Lin, J.; Panganiban, A.R. Individual differences in teaming with artificial intelligence, robots, and virtual agents in the workplace. In *The Oxford Handbook of Individual Differences in Organizational Contexts*; Tuncdogan, A., Acar, O.A., Volberda, W.V., Eds.; Oxford University Press: Oxford, UK, 2024; pp. 345–367.
81. Matthews, G. *Robot Compromise*; Final Report for Air Force Office of Scientific Research; Air Force Office of Scientific Research (AFOSR): Arlington, VA, USA, 2025.
82. Dahlstrom, M.F. Using narratives and storytelling to communicate science with nonexpert audiences. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 13614–13620. [CrossRef] [PubMed]
83. Molinaro, K.A.; Bolton, M.L. Evaluating the applicability of the double system lens model to the analysis of phishing email judgments. *Comput. Secur.* **2018**, *77*, 128–137. [CrossRef]
84. Kahneman, D. *Thinking, Fast and Slow*; Farrar, Straus and Giroux: New York, NY, USA, 2011.
85. Faghihi, U.; Estey, C.; McCall, R.; Franklin, S. A cognitive model fleshes out Kahneman's fast and slow systems. *Biol. Inspir. Cogn. Archit.* **2015**, *11*, 38–52. [CrossRef]
86. Paik, J.; Pirolli, P.; Lebiere, C.; Rutledge-Taylor, M. Cognitive biases in a geospatial Intelligence Analysis Task: An ACT-R model. In Proceedings of the Annual Meeting of the Cognitive Science Society, Sapporo, Japan, 1–4 August 2012; pp. 2168–2173.
87. Mar, R.A.; Oatley, K. The function of fiction is the abstraction and simulation of social experience. *Perspect. Psychol. Sci.* **2008**, *3*, 173–192. [CrossRef] [PubMed]
88. Saletta, M.; Kruger, A.; Primoratz, T.; Barnett, A.; van Gelder, T.; Horn, R.E. The role of narrative in collaborative reasoning and intelligence analysis: A case study. *PLoS ONE* **2020**, *15*, e0226981. [CrossRef] [PubMed]
89. Thorne, A.; Nam, V. The storied construction of personality. In *The Cambridge Handbook of Personality Psychology*; Corr, P.J., Matthews, G., Eds.; Cambridge University Press: Cambridge, UK, 2021; pp. 491–505.
90. Frank, M.C.; Goodman, N.D. Cognitive modeling using artificial intelligence. *Annu. Rev. Psychol.* **2025**, *77*, 543–566. [PubMed]
91. Brunswik, E. The conceptual framework of psychology. In *International Encyclopedia of Unified Science (Volume 1, Number 10)*; The University of Chicago Press: Chicago, IL, USA, 1952.

92. Karelaia, N.; Hogarth, R.M. Determinants of linear judgment: A meta-analysis of lens model studies. *Psychol. Bull.* **2008**, *134*, 404–426. [CrossRef] [PubMed]
93. Bisantz, A.M.; Kirlik, A.; Gay, P.; Phipps, D.A.; Walker, N.; Fisk, A.D. Modeling and analysis of a dynamic judgment task using a lens model approach. *IEEE Trans. Syst. Man Cybern.-Part A Syst. Hum.* **2000**, *30*, 605–616. [CrossRef]
94. Bisantz, A.M.; Pritchett, A.R. Measuring the fit between human judgments and automated alerting algorithms: A study of collision detection. *Hum. Factors* **2003**, *45*, 266–280. [CrossRef] [PubMed]
95. Thompson, C.; Bucknall, T.; Estabrookes, C.A.; Hutchinson, A.; Fraser, K.; De Vos, R.; Binnecade, J.; Barrat, G.; Saunders, J. Nurses' critical event risk assessments: A judgement analysis. *J. Clin. Nurs.* **2009**, *18*, 601–612. [CrossRef] [PubMed]
96. Choo, C.W. Information use and early warning effectiveness: Perspectives and prospects. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 1071–1082. [CrossRef]
97. Newell, B.R.; Shanks, D.R. Unconscious influences on decision making: A critical review. *Behav. Brain Sci.* **2014**, *37*, 1–19. [CrossRef] [PubMed]
98. Seong, Y.; Bisantz, A.M.; Bisantz, A.M. Judgment and trust in conjunction with automated decision aids: A theoretical model and empirical investigation. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2002**, *46*, 423–427. [CrossRef]
99. Ortoleva, P. Alternatives to Bayesian updating. *Annu. Rev. Econ.* **2024**, *16*, 545–570. [CrossRef]
100. Friston, K. The free-energy principle: A rough guide to the brain? *Trends Cogn. Sci.* **2009**, *13*, 293–301. [CrossRef] [PubMed]
101. Zachary, W.; Rosoff, A.; Miller, L.C.; Read, S.J. Context as a cognitive process: An integrative framework for supporting decision making. In Proceedings of the 8th International Conference on Semantic Technologies for Intelligence, Defense, and Security (STIDS 2013), Fairfax, VA, USA, 11–14 November 2013; pp. 48–55.
102. Van Gelder, T.; Saletta, M.; De Rozario, R.; Barnett, A.; Dwyer, T.; Satriadi, K.; Brugh, C.S. Storytelling in intelligence: Theoretical foundations. *Int. J. Intell. CounterIntell.* **2025**, *38*, 1298–1330. [CrossRef]
103. U.S. Government. *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis*; U.S. Central Intelligence Agency, Center for the Study of Intelligence: Washington, DC, USA, 2009. Available online: https://www.cia.gov/resources/csi/static/Tradecraft-Primer-apr09.pdf (accessed on 9 July 2026).
104. Schneider, J. Mental model shifts in human-LLM interactions. *J. Intell. Inf. Syst.* **2025**, *63*, 1737–1752. [CrossRef]
105. Eekhof, L.S.; Van Krieken, K.; Sanders, J.; Willems, R.M. Reading minds, reading stories: Social-cognitive abilities affect the linguistic processing of narrative viewpoint. *Front. Psychol.* **2021**, *12*, 698986. [CrossRef] [PubMed]
106. Krueger, K.L.; Diabes, M.A.; Weingart, L.R. The psychological experience of intragroup conflict. *Res. Organ. Behav.* **2022**, *42*, 100186. [CrossRef]
107. Matthews, G.; Cumings, R.; Casey, J.; Panganiban, A.R.; Chella, A.; Pipitone, A.; Lin, J.; Mouloua, M. Compromise in human-robot collaboration for threat assessment. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2024**, *68*, 329–335. [CrossRef]
108. Pipitone, A.; Geraci, A.; D'Amico, A.; Seidita, V.; Chella, A. Robot's inner speech effects on human trust and anthropomorphism. *Int. J. Soc. Robot.* **2024**, *16*, 1333–1345.
109. Pipitone, A.; Seidita, I.; Sullins, J.P.; Chella, A. Unlocking practical wisdom through the inner voice of robots. *Sci. Rep.* **2025**, *15*, 2634. [CrossRef] [PubMed]
110. Pipitone, A.; Cataldo, G.; Corvaia, S.; Chella, A. The robot and the nurse prepare for surgery: Early insights into the impact of robot's inner speech. *Intell. Serv. Robot.* **2026**, *19*, 5.
111. Lyons, J.; Highland, P.; Bos, N.; Lyons, D.; Skinner, A.; Schnell, T.; Hefron, R. Measuring perceived agent appropriateness in a live-flight human-autonomy teaming scenario. *Ergon. Des.* **2024**, *32*, 12–17.
112. Gall, J.; Stanton, C.J. Low-rank human-like agents are trusted more and blamed less in human-autonomy teaming. *Front. Artif. Intell.* **2024**, *7*, 1273350. [CrossRef] [PubMed]
113. Hagemann, V.; Rieth, M.; Watermann, L.; Appelganc, K.; Heinicke, C. Team member or tool? The role of framing in human-autonomy teaming. *Int. J. Hum.-Comput. Interact.* **2026**, 1–22. [CrossRef]
114. Lyons, J.B.; Wynne, K.T. Human-machine teaming: Evaluating dimensions using narratives. *Hum.-Intell. Syst. Integr.* **2021**, *3*, 129–137. [CrossRef]
115. Doyle, J.K.; Ford, D.N. Mental models concepts for system dynamics research. *Syst. Dyn. Rev.* **1998**, *14*, 3–29. [CrossRef]
116. Li, Z.; Terfurth, L.; Woller, J.P.; Wiese, E. Mind the machines: Applying implicit measures of mind perception to social robotics. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Hokkaido, Japan, 7–10 March 2022; pp. 236–245.
117. Li, M.; Suh, A. Anthropomorphism in AI-enabled technology: A literature review. *Electron. Mark.* **2022**, *32*, 2245–2275. [CrossRef]
118. Reuter, M.; Kirchhoff, B.M.; Franke, T.; Radüntz, T.; Peifer, C. To trust or not to trust a human (-like) AI—A scoping review and conjoint analyses on factors influencing anthropomorphism and trust. *Z. Arbeitswissenschaft* **2025**, *79*, 402–432. [CrossRef]
119. Rix, J. From tools to teammates: Conceptualizing humans' perception of machines as teammates with a systematic literature review. In Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS), Maui, HI, USA, 3–8 January 2022; pp. 398–407.

120. Andrews, R.W.; Lilly, J.M.; Srivastava, D.; Feigh, K.M. The role of shared mental models in human-AI teams: A theoretical review. *Theor. Issues Ergon. Sci.* **2023**, *24*, 129–175.
121. Sanchez, T.; Vereschak, O.; Deroy, O. Mental models in human-AI interaction: Systematic review of empirical methodologies and guidelines. In Proceedings of the 31st International Conference on Intelligent User Interfaces, New York, NY, USA, 23–26 March 2026; pp. 663–682.
122. Wynne, K.T.; Lyons, J.B. Autonomous agent teammate-likeness: Scale development and validation. In *International Conference on Human-Computer Interaction*; Springer International Publishing: Cham, Switzerland, 2019; pp. 199–213.
123. Kepes, S.; Keener, S.K.; Lievens, F.; McDaniel, M.A. An integrative, systematic review of the situational judgment test literature. *J. Manag.* **2025**, *51*, 2278–2319.
124. Wu, J.; Du, X.; Liu, Y.; Tang, W.; Xue, C. How the degree of anthropomorphism of human-like robots affects users' perceptual and emotional processing: Evidence from an EEG study. *Sensors* **2024**, *24*, 4809. [CrossRef] [PubMed]
125. Pataranutaporn, P.; Liu, R.; Finn, E.; Maes, P. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nat. Mach. Intell.* **2023**, *5*, 1076–1086. [CrossRef]
126. Christoforakos, L.; Gallucci, A.; Surmava-Große, T.; Ullrich, D.; Diefenbach, S. Can robots earn our trust the same way humans do? A systematic exploration of competence, warmth, and anthropomorphism as determinants of trust development in HRI. *Front. Robot. AI* **2021**, *8*, 640444. [CrossRef] [PubMed]
127. Mehrotra, S.; Degachi, C.; Vereschak, O.; Jonker, C.M.; Tielman, M.L. A systematic review on fostering appropriate trust in Human-AI interaction: Trends, opportunities and challenges. *ACM J. Responsible Comput.* **2024**, *1*, 1–45. [CrossRef]
128. Mayer, R.C.; Davis, J.H. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *J. Appl. Psychol.* **1999**, *84*, 123. [CrossRef]
129. Li, M.; Bitterly, T.B. How perceived lack of benevolence harms trust of artificial intelligence management. *J. Appl. Psychol.* **2024**, *109*, 1794–1816. [CrossRef] [PubMed]
130. Novozhilova, E.; Mays, K.; Paik, S.; Katz, J.E. More capable, less benevolent: Trust perceptions of AI systems across societal contexts. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 342–366. [CrossRef]
131. Fallon, C.K.; Panganiban, A.R.; Chiu, P.; Matthews, G. The effects of a trust violation and trust repair in a distributed team decision-making task: Exploring the affective component of trust. In *Advances in Social & Occupational Ergonomics*; Springer International Publishing: Cham, Switzerland, 2014; pp. 447–459.
132. Guo, G.; Zhang, J.; Thalmann, D.; Basu, A.; Yorke-Smith, N. From ratings to trust: An empirical study of implicit trust in recommender systems. In Proceedings of the 29th Annual ACM Symposium on Applied Computing, Gyeongju, Republic of Korea, 24–28 March 2014; pp. 248–253.
133. Naidoo, K.; Baldwin, J.; Lesar, J.; Plummer, L. Narrative analysis to track the development of clinical reasoning during residency. *J. Clin. Educ. Phys. Ther.* **2022**, *4*. [CrossRef] [PubMed]
134. Yu, F.; Zhang, H.; Tiwari, P.; Wang, B. Natural language reasoning, a survey. *ACM Comput. Surv.* **2024**, *56*, 1–39. [CrossRef]
135. Zhu, L.; Zhao, R.; Gui, L.; He, Y. Are NLP models good at tracing thoughts: An overview of narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: Singapore, 2023; pp. 10098–10121.
136. Liu, T.; Ye, L.; Yan, W. A framework for evaluation of large language models in essay assessment: Reliability, alignment, and causal reasoning. *Comput. Educ. Artif. Intell.* **2026**, *10*, 100565. [CrossRef]
137. McIntosh, T.R.; Liu, T.; Susnjak, T.; Watters, P.; Halgamuge, M.N. A reasoning and value alignment test to assess advanced gpt reasoning. *ACM Trans. Interact. Intell. Syst.* **2024**, *14*, 1–37. [CrossRef]
138. Harvey, P.; Madison, K.; Martinko, M.; Crook, T.R.; Crook, T.A. Attribution theory in the organizational sciences: The road traveled and the path ahead. *Acad. Manag. Perspect.* **2014**, *28*, 128–146. [CrossRef]
139. Brailsford, J.; Vetere, F.; Velloso, E. Responsibility attribution in human interactions with everyday ai systems. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 26 April–1 May 2025; pp. 1–17.
140. Johnson-Laird, P.N. *Mental Models: Towards A Cognitive Science of Language, Inference, and Consciousness*; Harvard University Press: Cambridge, MA, USA, 1983.
141. Naikar, N.; Hoffman, R.; Roth, E.M.; Klein, G.; Militello, L.G.; Dominguez, C. Should we make AI more tool-like or teammate-like? *J. Cogn. Eng. Decis. Mak.* **2026**, *20*, 137–157.
142. Reinerman-Jones, L.; Barber, D.J.; Szalma, J.L.; Hancock, P.A. Human interaction with robotic systems: Performance and workload evaluations. *Ergonomics* **2017**, *60*, 1351–1368. [CrossRef] [PubMed]
143. Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W.S.; Weld, D.S.; Horvitz, E. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2019; Volume 7, pp. 2–11.
144. Halevy, N.; Cohen, T.R.; Chou, E.Y.; Katz, J.J.; Panter, A.T. Mental models at work: Cognitive causes and consequences of conflict in organizations. *Personal. Soc. Psychol. Bull.* **2014**, *40*, 92–110.

145. Vanneste, B.S.; Puranam, P. Artificial intelligence, trust, and perceptions of agency. *Acad. Manag. Rev.* **2025**, *50*, 726–744. [CrossRef]
146. Glikson, E.; Woolley, A.W. Human trust in artificial intelligence: Review of empirical research. *Acad. Manag. Ann.* **2020**, *14*, 627–660. [CrossRef]
147. Filho, E.; Pierini, D.; Robazza, C.; Tenenbaum, G.; Bertollo, M. Shared mental models and intra-team psychophysiological patterns: A test of the juggling paradigm. *J. Sports Sci.* **2017**, *35*, 112–123. [PubMed]
148. Potinteu, A.E.; Said, N.; Jahn, G.; Huff, M. An insight into humans helping Robots: The role of attitudes, anthropomorphic cues, and context of use. *Comput. Hum. Behav. Artif. Hum.* **2025**, *4*, 100159. [CrossRef]
149. Roesler, E.; Naendrup-Poell, L.; Manzey, D.; Onnasch, L. Why context matters: The influence of application domain on preferred degree of anthropomorphism and gender attribution in human–robot interaction. *Int. J. Soc. Robot.* **2022**, *14*, 1155–1166. [CrossRef]
150. Grankvist, G.; Larsson, M.; Lindström Kupiainen, R. Trust in AI: Supportive vs. autonomous roles across four domains. *AI Soc.* **2026**, 1–7. [CrossRef]
151. Krügel, S.; Richter, F.; Uhl, M. Context-dependency of trust in AI-based systems. In Proceedings of the 2025 IEEE International Symposium on Technology and Society (ISTAS), Santa Clara, CA, USA, 10–12 September 2025; pp. 1–4.
152. Matthews, G.; Cumings, R.; De Los Santos, E.P.; Feng, I.Y.; Mouloua, S.A. A new era for stress research: Supporting user performance and experience in the digital age. *Ergonomics* **2025**, *68*, 913–946. [PubMed]
153. Dang, Q.; Li, G. Unveiling trust in AI: The interplay of antecedents, consequences, and cultural dynamics. *AI Soc.* **2026**, *41*, 669–692.
154. Duan, W.; Zhou, S.; Scalia, M.J.; Weng, N.; Yin, X.; Hao, R.; Freeman, G.; Funke, G.; Tolston, M.; Gorman, J.; et al. "If they don't even trust their AI teammate, why should we?": Understanding trust contagion across multiple human-AI teams. *Proc. ACM Hum.-Comput. Interact.* **2026**, *10*, 1–31. [CrossRef]
155. Fernandes, D.; Villa, S.; Nicholls, S.; Haavisto, O.; Buschek, D.; Schmidt, A.; Kosch, T.; Shen, C.; Welsch, R. AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Comput. Hum. Behav.* **2025**, *175*, 108779. [CrossRef]
156. Gültekin, D.G.; Siniksaran, E. Overconfidence, gender, and authority attribution in algorithmic versus human advice: A cognitive perspective on compliance. *Acta Psychol.* **2025**, *260*, 105668. [CrossRef]
157. Zhang, R.; McNeese, N.J.; Freeman, G.; Musick, G. "An ideal human" expectations of AI teammates in human-AI teaming. *Proc. ACM Hum.-Comput. Interact.* **2021**, *4*, 1–25. [CrossRef]
158. Shneiderman, B. Design lessons from AI's two grand goals: Human emulation and useful applications. *IEEE Trans. Technol. Soc.* **2020**, *1*, 73–82. [CrossRef]
159. Placani, A. Anthropomorphism in AI: Hype and fallacy. *AI Ethics* **2024**, *4*, 691–698. [CrossRef]
160. Howcroft, A.; Blake, H. Empathy by design: Reframing the empathy gap between AI and humans in mental health chatbots. *Information* **2025**, *16*, 1074. [CrossRef]
161. Eiband, M.; Schneider, H.; Bilandzic, M.; Fazekas-Con, J.; Haug, M.; Hussmann, H. Bringing transparency design into practice. In Proceedings of the 23rd International Conference on Intelligent User Interfaces, Tokyo, Japan, 7–11 March 2018; pp. 211–223.
162. Felzmann, H.; Fosch-Villaronga, E.; Lutz, C.; Tamò-Larrieux, A. Towards transparency by design for artificial intelligence. *Sci. Eng. Ethics* **2020**, *26*, 3333. [CrossRef] [PubMed]
163. Chen, R.; Jiang, W.; Qin, C.; Tan, C. Theory of Mind in Large Language Models: Assessment and enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 31539–31558.
164. Winfield, A.F.T. Experiments in artificial theory of mind: From safety to story-telling. *Front. Robot. AI* **2018**, *5*, 357467. [CrossRef] [PubMed]
165. Hancock, P.A.; Pepe, A.A.; Murphy, L.L. Hedonomics: The power of positive and pleasurable ergonomics. *Ergon. Des.* **2005**, *13*, 8–14. [CrossRef]
166. Riedl, R. Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions. *Electron. Mark.* **2022**, *32*, 2021–2051. [CrossRef]
167. Liu, W.; Yao, M. The minds behind the prompts: How mental representations shape user interaction with generative AI. *Int. J. Hum.-Comput. Interact.* **2026**, 1–19. [CrossRef]
168. Ngo, T.; Kunkel, J.; Ziegler, J. Exploring mental models for transparent and controllable recommender systems: A qualitative study. In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, Genoa, Italy, 12–18 July 2020; pp. 183–191.
169. Araujo, T.; Bol, N. From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents. *Comput. Hum. Behav. Artif. Hum.* **2024**, *2*, 100030. [CrossRef]