Modeling Theory of Mind in Multimodal HCI

Yifan Zhu[1], Hannah VanderHoeven[2], Kenneth Lai[1], Mariah Bradford[2], Christopher Tam[1], Ibrahim Khebour[2], Richard Brutti[1], Nikhil Krishnaswamy[2], and James Pustejovsky[1(✉)]

[1] Brandeis University, Waltham, MA 02453, USA
{zhuyifan,jamesp}@brandeis.edu
[2] Colorado State University, Fort Collins, CO 80523, USA

Abstract. As multimodal interactions between humans and computers become more sophisticated, involving not only speech, but gestures, haptics, eye movement, and other input types, each modality introduces subtleties which can be misinterpreted without a deeper understanding of the agent's mental state. In this paper, we argue that Simulation Theory of Mind (SToM) [23], interpreted within a model of embodied HCI [41,42], can help model the capacity to attribute beliefs and intentions to oneself and others. We adopt a version of Dynamic Epistemic Logic that admits of degrees of belief, reflecting changing evidence available to an agent [5,6]. This model is able to address the complexities of mutual perception and belief, and how a dynamic common ground is constructed and changes [15]. To demonstrate this, we apply the SToM model to the problem of Common Ground Tracking (CGT) in multi-party dialogues, focusing here on a joint problem-solving task called the Weights Task, where participants cooperate to find the weights of a set of blocks.

Keywords: Theory of Mind · HCI · Epistemic Updating · Common ground tracking · multimodal dialogue · simulation · Embodiment

1 Introduction

Theory of Mind (ToM) refers to the cognitive capacity that humans have to attribute mental states such as beliefs (true or false), desires, and intentions to oneself and others, thereby predicting and explaining behavior [39,56]. Within the domain of Human-Computer Interaction (HCI), this concept has recently become more relevant for computational agents, especially in the context of multimodal communication [15]. As multimodal interactions involve not only speech, but gestures, actions, eye gaze, body posture and other input types (cf. Fig. 1), each modality introduces subtleties which can be misinterpreted without a deeper understanding of the agent's mental state. As a result, ToM's role becomes important, ensuring that agents grasp both the overt and covert nuances of human communication. For multimodal HCI, a computational agent needs to incorporate both the ability to reason about the beliefs of other agents

M. Kurosu and A. Hashizume (Eds.): HCII 2024, LNCS 14684, pp. 205–225, 2024.
https://doi.org/10.1007/978-3-031-60405-8_14

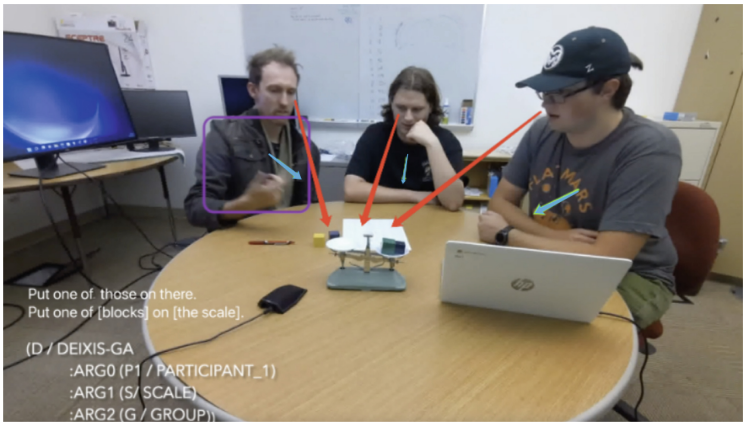

Fig. 1. Example of a multimodal annotated situation red arrows denote co-gazing blue arrows symbolize leaning towards the table. (Color figure online)

and their intentions, while also knowing what beliefs can be assumed or taken for granted in a specific context.

In this paper, we argue that Simulation Theory of Mind (SToM) [23,24], encoded as an evidence-based dynamic epistemic logic (EB-DEL), can help model these complexities [5,6]. We develop this view within a model of embodied HCI [41,42], where simulation theory is inherent in both the semantics of the model as well as the implementation of situated meaning and action. Specifically, we apply this model to the problem of Common Ground Tracking (CGT) in multi-party dialogues, focusing here on a joint problem-solving task called the Weights Task, where participants cooperate to find the weights of a set of blocks. In such task-oriented interactions, successful communication hinges not only on understanding the immediate intent but also on a shared context and knowledge of the actions and experiences of the agents. Theories of Common Ground posit that for effective communication, interlocutors must have a mutual understanding of the task at hand and the context in which it is situated [4,16]. In HCI, this translates to the system and the user operating with aligned expectations and shared knowledge about the ongoing task.

Unlike Dialogue State Tracking (DST), which is the ability to update the representations of the speaker's needs at each turn in the dialogue by taking into account the past dialogue moves and history [28], Common Ground Tracking (CGT) identifies the shared belief space held by all of the participants in a task-oriented dialogue. While ToM enables a system to "read" another person's mental states, Common Ground Tracking ensures that this reading is anchored in a shared context, making interactions more coherent and goal-directed.

Within the framework of SToM and Embodied HCI adopted here, we present a method for automatically identifying the current set of shared beliefs and questions under discussion (QUDs) of a group with a shared goal, the Weights Task. The task involves triads collaborating to determine the weights of five

blocks using a balance scale [29]. We track the shared knowledge of participants in a co-situated environment, which involves interpreting the communications over multiple modalities, and integrating these channels into a coherent model of the common ground.

We believe that models of belief and intent in multimodal HCI can be significantly enriched by integrating principles of ToM and Common Ground, allowing interactive systems to not only react to user inputs but interpret them in a shared context, making interactions more predictive, context-aware, and aligned with user expectations. By integrating ToM and common ground tracking into conversational agent architectures [19,51], we can better model the beliefs of participants by exposing unspoken assumptions of the participants or disagreements among them. Enhancing the epistemic modeling capabilities of multimodal HCI with ToM also has the potential to inform research in both Affective Computing, such as automatic Emotion Detection, by providing more contextualized interpretations of cognitive states and emotions in dialogue [46], as well as providing support for those with functional impairments [20].

In the final section, we provide detailed evidence assessing the contribution of each feature type from different channels toward successful construction of common ground relative to ground truth, and show how the combination of modalities results in a higher-fidelity prediction of both cognitive states of the participants and propositions implicitly or explicitly expressed.

2 Related Work

There is a significant tradition of research on Theory of Mind in philosophy and its application to questions of epistemic awareness within developmental psychology [25,39,55,56]. One view that is particularly relevant to the approach taken here is Simulation Theory [24], which models the process of understanding another agent's intentions as mental simulations from one's own perspective. Simulation Theory, as developed in philosophy of mind by Goldman and others, has focused on the role that "mind reading" plays in modeling the mental representations of other agents and the content of their communicative acts [24,26,27]. Simulation semantics as adopted within cognitive linguistics [17,36], argues that language comprehension is accomplished by means of such mind reading operations. Similarly, within psychology, there is an established body of work arguing for "mental simulations" of future or possible outcomes, as well as interpretations of perceptual input [3]. These simulation approaches can be referred to as *embodied theories of mind*. The goal here is to create a semantic interpretation of an expression or action by embodying it in a simulation.

Goldman's theory of mind [24], viewed from the perspective of simulation theory, provides a mechanism for how individuals understand and predict the behaviors of other agents. By constructing a simulation, an agent can generate hypotheses about others' mental states and intentions, anticipate possible future actions, and even empathize with their emotional states. SToM can be seen as consisting of three major components: (a) *Mental Simulation*, where an

agent simulates the mental processes of others; (b) *Perspective Taking*, where an agent adopts another person's Epistemic Frame of Reference; and (c) *Shared Mental Processes*, the view that others' mental states involve the same cognitive mechanisms as one's own.

Within HRI, the question of epistemic awareness and social appropriateness of robot behavior has been a concern since the foundation of the discipline [12]. Similarly, the modeling of first-order beliefs in HRI has been addressed within the modeling and reasoning community [49], while the application of Dynamic Epistemic Logic itself to planning has resulted in significant developments within the area of *epistemic planning* [8], and subsequent work [4,9].

But fundamental capabilities involving inferencing over others' beliefs as well as an agent's metacognitive abilities have been less studied. This problem is addressed in [15], where false-belief scenarios are encoded within the version of Dynamic Epistemic Logic introduced by Bolander [7]. This model accounts for an agent's false belief regarding a changing environment, as well as the ability of other agents to recognize and reason about this agent's incorrect epistemic state. While not adopting Bolander's specific model of DEL here, our research aligns squarely with their approach to modeling the dynamics of belief updating in HRI and HCI contexts. Since we are focusing on integrating the semantics associated with distinct channels of communication (speech and gesture) as well as actions and perceptions, we need to integrate semantic content derived from these distinct channels into a common format and data structure. This involves adopting a more expressive model for how common ground is constructed and updated, as discussed below in Sect. 4. We also account for epistemic content held individually and in common within a group, with an evidence-based model of DEL as introduced in [5,6].

There has been considerable work on simulating both physical and cognitive processes carried out by agents, human and computational [30,44,59]. Of particular relevance to the research reported here is the simulation framework, VoxWorld [34]. VoxWorld supports embodied HCI, where artificial agents consume different sensor inputs for awareness of not only their own virtual space but also the surrounding physical space. It brings together the notion of simulation systems from computer science as well as that mentioned here, in the context of Simulation ToM. VoxWorld is a collaborative creation of the VoxML modeling language [40] and its real-time Unity interpreter, VoxSim [32], culminating in an environment meticulously defined interaction semantics. Such an architectural framework readily facilitates the interpretation of action annotations, as descriptions converted to linguistic entities within VoxML exhibit an explicit correspondence within the simulation context. Within the VoxWorld ecosystem, the Diana agent emerges as a pivotal interface designed to discern user speech and gestures [31], setting itself apart from other Intelligent Virtual Agents (IVAs) through its capacity for reasoning and acting upon a diverse array of objects endowed with a well-defined affordance structure.

3 A Multimodal Dataset for Common Ground Tracking

The Weights Task, as described in [29], embodies a collaborative problem-solving task wherein groups of three participants engage in a concerted effort to infer the unknown weights of different blocks. This inference is achieved through comparative analyses of the blocks' weights with a balance scale. Each group is equipped with a scale and five blocks with different colors, sizes, and weights. Participants are informed of the weight of a singular block and are tasked with discerning the weights of the remaining blocks and the algebraic relation between them (the Fibonacci Sequence). Due to the co-situated nature of the task and its inclusion of physical objects and reasoning about their properties, the communication in this task can be annotated in several ways: speech with dense paraphrasing [52], gesture [10], as well as non-verbal behaviors that communicate intent such as gaze [35], body postures [43], facial expression [45]. Additionally, we label all actions performed [50], and collaborative problem solving (CPS) indicators according to the framework of [48]. Each group successfully deduces the accurate weights of the blocks, thereby establishing a uniform and reliable endpoint for evaluating our models.

First-order epistemic statements represent what an agent believes or knows about their environment, while second-order epistemic representations express what an agent believes or knows about other agents' beliefs and knowledge of the environment. One avenue for studying ToM in collaborative interactions is to track the propositional context expressed by verbal actions (explicitly or implicitly uttered), already embedded in common ground. Another avenue with not much exploration is tracking information results not only in speech but also gesture: non-verbal actions and gestures contained individual consenting an agreement including mere perception of the action itself [7], as well as false-belief scenarios within the version of Dynamic Epistemic Logic.

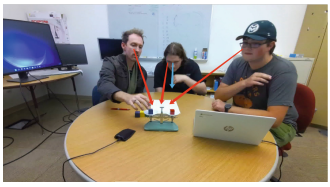

Fig. 2. Gaze and posture, P_1, P_2, P_3 co-attending, P_2 lean in

Dialogue in a Joint Activity

– P_3: *It's pretty balanced.*
– P_1: *Yeah.*

Fig. 3. Dialogue in 2.

Consider the joint activity shown in Fig. 2 among three participants; Participant 3 (P_3) has a public announcement and P_1 signals his consent. The figure has a blue arrow to highlight P_2 leaning in to have a closer look at the scale after he heard the dialogue between P_1 and P_3. Figure 3 presents their dialogue in the context of joint activity. In this situation of a multi-agent task interaction, there

are several elements constituting the common ground among the three participants. These elements include reference to: the agents, shared beliefs, shared goals, and shared perception of objects, including that the scale is balanced.

4 Constructing Multimodal Common Ground

4.1 Epistemic Modeling in ToM

The notion of common ground is a fundamental concept in HCI, growing out of a rich tradition in philosophy and cognitive science, focused on exploring how individuals coordinate and establish shared understanding to facilitate meaningful conversations [1,11,47]. For HCI, common ground is crucial for designing interactive systems and technologies that can effectively communicate with users by anticipating their needs, understanding their context, and responding to their inputs in an intuitive manner. With the presence of a common ground during shared experiences, embodied communication assumes agents can understand one another in a shared context, through the use of co-situational and co-perceptual anchors, and a means for identifying such anchors, such as gesture, gaze, intonation, and language. In this section, we develop a computational model of common ground for multimodal communication.

Within the context of multimodal interactions, the notion of common ground relies on identifying three key aspects of the interaction:

1. *Co-situatedness* of the agents, such that they can interpret the same situation from their respective frames of reference;
2. *Co-perception* and *co-attention* of a shared situated reference, which allows more expressiveness in referring to the environment (i.e., through language, gesture, visual presentation);
3. *Co-belief* of the agents regarding the goals as well as the steps involved towards accomplishing this goal.

Within this context, common ground emerges in one of the following ways during social interactions [16]:

- by public announcement, through either speech or gesture;
- by common witnessing of an event;
- by combinations of the above (indirect co-presence, cultural co-presence).

In order to characterize the many dimensions of human-computer interactions, we introduce an approach to evaluating interactions drawing on the most relevant parameters in co-situated communicative interactions. By introducing a formal model of shared context, we are able to track the intentions and utterances, as well as the perceptions and actions of the agents involved in a dialogue. The computer, either as an embodied agent distinct from the viewer, or as the totality of the rendered environment itself, presents an interpretation (*mind-reading*) of its internal model, down to specific parameter values, which are often assigned for the purposes of testing that model.

We assume a model of discourse semantics as proposed in [13], as it facilitates the adoption of a continuation-based semantics for discourse. In the present work, however, update functions will be limited to dialogue-based moves, and we will not focus on the sentence-level update semantics. We adopt the SToM model of VoxWorld (discussed in Sect. 2) [34], a framework for modeling HCI and human-human multimodal interactions as embodied simulations. In this model, participants are embodied agents endowed with intentions, goals, beliefs, and the knowledge to complete simple tasks involving multimodal interactions with co-participants. Each agent's state in the dialogue is continuously updated through encoding the changes in the environment shared by the interacting agents [41,42].

As mentioned in Sect. 1 above, our investigation involved a triad of co-situated students collaborating to solve a weights task for five blocks, using only a balance scale. The task is particularly relevant because the participants naturally engage in the different modalities that are so crucial for understanding multimodal HCI: namely, speech, gesture, gaze, pose, and of course joint actions. Hence, from a dialogue state tracking perspective, there are several distinct action types and their effects that need to be accounted for and tracked:

(1) a. **Ontic actions**; interactions with and movements of the objects in the shared space; i.e., blocks and the balance scale;
b. **Epistemic actions**; changes to the epistemic state of one or more of the participants in the interaction.

Before showing how this is done, we spell out the cognitive capabilities of the participants as computational agents performing these various actions.

We begin with the cognitive architecture adopted and developed in VoxWorld [41,42], and enrich the model capabilities to more systematically account for epistemic updates in the common ground. We assume an embodied computational agent has the following capabilities.

(2) a. *Perception*: perceptual sensors and interpreters.
b. *Action*: action effectors and planning.
c. *Belief*: a Dynamic Epistemic Logic with updating.

For the present discussion, we focus on those aspects of the architecture that are relevant to demonstrating the role ToM plays in tracking communicative content and intent. For this reason, we concentrate mainly on the role of perception and belief, and subsequent epistemic updating.

Let us first consider the role of an agent's perception of their environment. As discussed in [41], VoxWorld models an agent a's vision as sets of accessibility relations between situations (defined as S_a), where there are two kinds of perception reports: of a proposition, φ; or of an object, x, coerced to the propositional content of "x exists in the situation." The modal expression, $S_a\varphi$, is interpreted as a direct (veridical) perception of agent a of proposition, φ. Hence, modal axiom T holds, where $S_a\varphi \rightarrow \varphi$.

In VoxWorld, this impacts the way beliefs are updated, where it has the effect of introducing the axiom below (where the modal B_a represents belief of an agent a):

(3) **Seeing is Believing:** $S_a\varphi \rightarrow B_a\varphi$ (veridical perception)

For the multimodal interactions and experiments performed within VoxWorld to date, veridical perception was both required as well as a computational asset. However, in the context of the experimental configuration introduced by the Weights Task, we see a different role being contributed by an agent's perception, relative to the completion of the task: namely, an appreciation of the natural role that perception plays in arriving at evidence for a belief [18,53]. This requires distinguishing two types of "seeing", both direct and indirect, the latter which is implicated in forming belief, to which we now turn.

In terms of epistemic modeling, while the VoxWorld platform from [34] encodes an agent's belief for a dialogue state, the mechanisms used for updating epistemic values resulting from actions during the dialogue are linked to specific moves and transitions in the dialogue state machine [33]. As a result, identifying general axioms for epistemic updating across situations can be difficult.

To overcome both of these shortcomings, we adopt here an implementation of evidence-based Dynamic Epistemic Logic (DEL) as developed in [5] and [37]. Where epistemic attitudes toward propositions are graded, according to the evidence available to the agent. As the dialogue progresses, information becomes available, weakening or strengthening propositional content present in the situation. This affords a more nuanced encoding of how perception relates to belief, and a more general mechanism for belief updating, as we shall see.

Given a set of agents, engaged in a cooperative task with a specific goal, the scope of unknowns is delineated by how an answer to each one contributes to the task solution. As a result, cooperative and interactive engagement brings about evidence both for and against how to answer these unknowns, in the hope of solving a problem. To this end, the role of both direct and indirect evidence of a proposition is crucial to an agent being confident to believe it. Hence, following [37], we will assume a model for evidence-based belief as a tuple, $\mathcal{M} = (W, E, V)$, where:

(4) a. W is a non-empty *set of worlds*;
b. $E \subseteq W \times \wp(W)$ is an *evidence relation*;
c. V: $At \rightarrow \wp(W)$, is a *valuation function*.

We will distinguish two sources of evidence. Let $E(w)$ denote the set $\{X \mid wEX\}$, the worlds accessible to w through the general evidencing relation, E. Beyond this, we distinguish between two sources of evidentiality: E_P, a perceptually-sourced evidence; and E_I, evidence derived through an inference over current common ground data. Accordingly, let $E_P(w)$ denote the set $\{X \mid wE_PX\}$, the worlds accessible to w through the "evidence through seeing" relation, E_P; and let $E_I(w)$ denote the set $\{X \mid wE_IX\}$, the worlds accessible to w through the "evidence through inferencing" relation, E_I.

The evidence-based epistemic-perceptual language, $\mathcal{L}_p$, will be the set of formulas generated by the grammar below, for any arbitrary agent:

(5) a. $p \mid \neg\varphi \mid \varphi \wedge \psi \mid [E_I]\varphi \mid [E_P]\varphi \mid [B]\varphi \mid [K]\varphi \mid [S]\varphi \mid [CE_P]\varphi \mid [CS]\varphi \mid [CB]\varphi$
b. $E_I(w) \subseteq E(w)$ and $E_P(w) \subseteq E(w)$

We distinguish the situation where an agent has "evidence in favor of" a proposition φ, as $[E]\varphi$. Because an agent can have evidence for propositions that convey contradictory information, she can consider both $[E]\varphi$ and $[E]\neg\varphi$. This corresponds to an agent having multiple neighborhoods, X, that are each evidenced in their unique way by w. However, consider the set of non-contradictory worlds as a unique subset of X, one which has what [6] refer to as the *finite intersection property (fip)*. This property allows us to identify a neighborhood of accessible worlds with non-contradictory propositional content. When this occurs, we say an agent has *belief* in a proposition, $[B]\varphi$. Following [37], the universal modality is considered "knowledge" of a proposition, $[K]\varphi$. Finally, veridical perception of a situation φ, is expressed as $[S]\varphi$. In conjunction with individual modal relations, we incorporate the concept of joint activity to denote a modal relation that is jointly shared by two or more participants. This is already assumed in our definition for common belief, see below. For direct perception and evidence, this will be indicated by $[CS_{\{a,b\}}]\varphi$, and $[CE_{P_{\{a,b\}}}]\varphi$, where a pair of agents, a and b are jointly seeing or evidencing φ to be the case. We define the expression of shared belief and shared perception in φ by a group, g, as and $CB_g\varphi$ and $CE_{Pg}\varphi$ respectively.

(6) a. $\mathcal{M},w \models CB_g\varphi$ iff $\forall v \in W$, if $w(\bigcup_{j\in A} R_j)^* v$, then $\mathcal{M},w \models \varphi$.
b. $\mathcal{M},w \models CE_{Pg}\varphi$ iff $\forall v \in W$, if $w(\bigcup_{j\in A} E_{Pj})^* v$, then $\mathcal{M},w \models \varphi$.

With the model adopted here, we are able to distinguish between direct "veridical perception" and "evidencing through perception", where $\Box\varphi \rightarrow \varphi$ holds for S but not for E_P.

While we have a formal distinction in the modal force associated with each mode of seeing (direct or evidential), we have not identified the conditions under which they are applicable. In our current experimental setup, the distinction is brought out very clearly as a function of what propositions are under discussion for verification. For the Weights Task, these are any propositions relating to the weights of specific blocks, their relative weights, and then how they algebraically relate to each other. Such propositions contribute to the solution of the problem the participants are engaged in solving. Hence, they are both "under discussion" and subject to degrees of evidential reasoning. This is accounted for by our distinction between a direct perception of an object or an event and an evidential perception of a proposition under discussion.

4.2 Common Ground in Dialogue

Given the model of Dynamic Epistemic Logic presented above, together with the mechanisms for encoding perception and evidence-based belief, let us formalize our assumptions about the common ground, *cg*, within a dialogue. We define the minimal structure of a task-oriented interaction as a sequence, D, of dialogue steps, where each move in the dialogue takes it into another situation or state. When considering multiple modalities of communication, along with the modality of action itself, we can generalize D to a multimodal dialogue (D_M). We will

define the transitioning step from one situation to the next, as a generalization of a dialogue step. Let $Ag = \{p_1, p_2, p_3\}$, be the participants in our Weights Task triad-based dialogue. From any situation s_k, we define a D_M move, m_i, as $m_i = (p_j, C_j, s_{k+1})$: participant p_j performs a communicative act C_j, bringing the multimodal dialogue into situation s_{k+1}. The D_M can be defined as the sequence of these moves, $D_M = m_1, \ldots, m_n$, where m: $M \subseteq S \times A \times P \times S$ and A is the set of actions.

Here our interest is in tracking the situation content resulting from each move: the set of propositions that captures the current state of the world, the current progress towards a goal, or the status of a task. In addition, we are interested in capturing the current questions under discussion and beliefs in the dialogue.

Given these considerations, we identify three components for tracking common ground in dialogue: a minimal static model of degrees of belief; a data structure distinguishing the elements of the agents' common ground that are being tracked [41]; and a dynamic procedure which updates this structure, when new information and evidence is available to the agents. We adopt Ginzburg's [22] notion of *Dialogue Gameboard (DGB)*, the public information associated with a state in a dialogue or discourse, modeled as a state monad, modified to correspond to the following elements in the dialogue state: $DGB = (C_a, Ag, CG, \mathcal{E})$:

(7) a. The communicative act, C_a, performed by an agent, a: $\langle S, G, F, Z, P, A \rangle$, a tuple of expressions from the diverse modalities involved. This includes the modalities conveying propositional content (language S and gesture G); nonverbal modalities conveying emotional engagement (facial expressions F and posture P); nonverbal behaviors indicating perceptual attention (gaze Z); and an explicit action, A.
b. Ag: the agents engaged in communication;
c. CG: the common ground structure;
d. $\mathcal{E}$: The embedding space that all agents occupy in the interaction.

We will focus on how actions impact the common ground, CG, such that it is dynamically updated throughout the dialogue. Following [21,22], modified to reflect the varying degrees of evidence associated with propositions under discussion, the common ground, cg, is a triple, (QB, EB, FB), consisting of:

(8) a. QBank (QB): these are "questions under discussion", a set of topics or unknowns that need to be answered to solve the task [21];
b. EBank (EB): these are evidenced propositions, those for which there is some evidence they are true;
c. FBank (FB): the set of propositions believed as true by all participants.

4.3 Epistemic Updating

Let us now examine how the epistemic state of each agent and the group they form is updated throughout the task-oriented dialogue associated with the

Weights Task. These are the personal DGBs for each agent and the joint DGB for the group. The task is a triad joint activity, with agents, $Ag = \{p_1, p_2, p_3\}$, who are co-situated in the embedding space, $\mathcal{E}$. Our domain of objects contains five colored blocks and a balance scale: $\{r, y, b, g, p, s\}$.[1]

At the outset of the task, the block weights are unknown to the participants. Hence, both the EBank and the FBank are empty, since there is nothing evidenced or known. Because finding the value of each block weight is part of the goal, these unknowns constitute the propositions known as "Questions under Discussion" (QUDs), and what we also refer to here as QBank. For all objects in the domain relating to the task, questions are generated for each relation implicated in the task for that object. Because the weight of a block ranges between 10 and 50 g, in 10-gram intervals we have five possible values, expressed as yes/no questions. The initial value of the QBank results in the following set:

(9) QBank $= \{Eq(r, 10)?, \ldots, Eq(r, 50)?, \ldots\ldots Eq(p, 50)?\}$

As the task proceeds, the participants try weighing different blocks and discuss their relative weights. When they make observations through their actions, they discover evidence in favor of propositions that are marked as questions in the QBank. As mentioned above, the mechanism available for updating the agents' common ground are through either a public announcement or by witnessing an action. We consider each of these in turn.

Public Announcements. Following [38] and subsequent developments of Public Announcement Logic [2], the operator $!\phi$ is used to represent the action of announcing ϕ publicly. The effect of such an announcement is that all agents update their knowledge states by eliminating worlds (possible states of affairs) where ϕ does not hold. If $[!\varphi]$ represents the act of announcing φ, then $[!\varphi]\psi$ means "after φ is announced, then ψ is believed to be the case." When a participant, P_i, in a group activity makes a statement relating to an observation, we say that P_i publicly announces $\mathcal{M} \models [!\psi]\phi$ if and only if $\mathcal{M}, w \models \phi$, where $\mathcal{M}|_\psi$ is the model $\mathcal{M}$ restricted to the worlds where ψ is true.

Witnessing of Events. When a participant, P_i, performs an act resulting in φ, in the co-presence of another participant, P_j, we say that P_i performs a publicly perceived act and result, φ. If α_i is an act performed by P_i, then $[\alpha_i]\varphi$ means "after i performs α, φ is the result." Hence, if multiple agents are co-attentive (co-perceptive) to the act, α, then a public witnessing is brought about by an act being performed, where the co-perception is represented as $S_{i,j}[\alpha]\varphi$.

In order to distinguish perceptual evidence for φ from belief in φ, we relativize the impact of a statement to the perceptual or inferential context within which it is uttered. Let us interpret $[!\varphi]\psi$ as follows.

(10) a. *Update with Evidence*:
$[!\varphi][E]\psi$: Given the announcement of φ, there is evidence for ψ;

[1] The blocks are uniquely colored as: red (r), yellow (y), blue (b), green (g), and purple (p). The scale is denoted as s.

b. *Update with Belief*:
$[E]\varphi \rightarrow [!\varphi][B]\psi$: Belief in φ is conditionalized on φ's announcement in the prior context of evidence for φ.
c. *Seeing provides Evidence*: If a participating agent, i, perceives an action, a, occur, then i has some evidence, E_P, that a has occurred, $\langle a\rangle\varphi$.
$[S_i]\langle a\rangle\varphi \rightarrow [E_{P_i}]\varphi$
d. *Non-contradictory Evidence provides Belief*: If a participating agent, P_i, has multiple non-contradicting evidences, E, for an action, a, occurring, $\langle a\rangle\varphi$, then P_i believes φ.
$([E]\langle a\rangle\varphi \wedge \mathbf{fip}) \rightarrow B\phi_a$

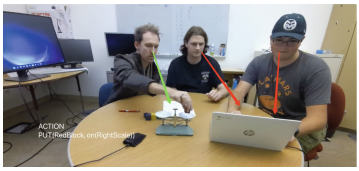

Fig. 4. P_2, P_3 co-perceive the laptop, P_1 perceives scale/blocks. Green arrow symbolizes gazing, red arrows denotes co-gazing. (Color figure online)

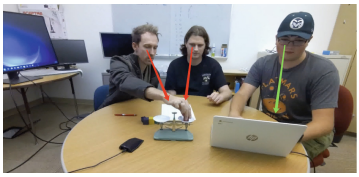

Fig. 5. P_1, P_2 co-perceive the scale/blocks, P_3 perceives the laptop. (Color figure online)

Semantically, an update represents the state of affairs after an announcement. This entails transforming the current model by removing all states where the announced formula is false. With evidence distinguished from belief/knowledge, we also update the evidence function, where $[!\varphi]$:

(11) a. Updates the worlds: $W' = W \cap \varphi$
b. Updates the Evidence function: $E'(w) = E(w) \cap \varphi$
c. $(M, w) \models \varphi$ implies $(M|_\varphi, w) \models [E]\psi$

This update actually changes the underlying evidence sets themselves. The announcement is taken as a piece of direct evidence. Hence, to capture that the announcement of φ becomes evidence and not just belief, the evidence sets for each agent get restricted (or updated) to reflect the worlds where φ is true. Subsequently, the belief function will then naturally adjust based on the new evidence sets.

Operationally, after (10a) is run, the model is relativized to evidencing neighborhoods, where φ is true. This corresponds to moving a proposition from the QBank to the EBank. Then, if the same proposition is "announced" again, as with an *ACCEPT* move, then (10b) promotes that proposition from the EBank to the FBank.

5 Common Ground Tracking

To illustrate the effect of the epistemic and evidential update functions outlined above, let us consider a joint activity scenario in Figs. 4, 5, 6, 7, 8 and 9. In Figs. 4 and 5, P_1 places the blue block on the left scale and the red block on the right scale. Subsequently, all three participants observe the equilibrium of the scales (Fig. 6). P_3 then issues a public declaration regarding the balanced state of the scales (Fig. 7), followed by a concurring public declaration from P_1 (Fig. 8).

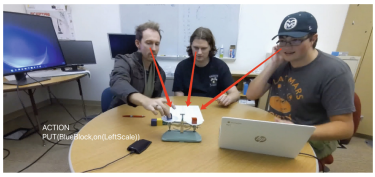

Fig. 6. P_1, P_2 and P_3 co-perceive scale/ blocks. (Color figure online)

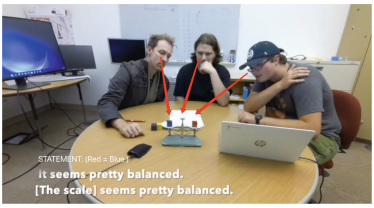

Fig. 7. P_1, P_2 and P_3 co-perceive scale/ blocks. (Color figure online)

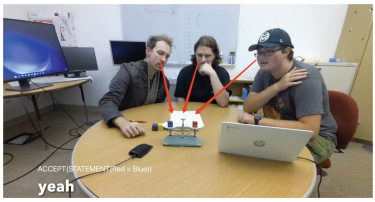

Fig. 8. P_1, P_2 and P_3 co-perceive scale/ blocks. (Color figure online)

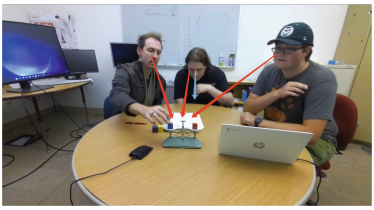

Fig. 9. P_1,P_2 and P_3 co-perceive scale/ blocks. P_2 leans towards table. (Color figure online)

Now assume that **b** refers to "the scales are balanced". Then, since $S_{p_1}\mathbf{b}$, $S_{p_2}\mathbf{b}$, $S_{p_3}\mathbf{b}$ we have a co-attention, CS **b**. Grounded in the axiom "Seeing provides Evidence," this observation serves as perceptual evidence for all three participants, manifested as $E_{P_{p_1}}\mathbf{b}$, $E_{P_{p_2}}\mathbf{b}$, $E_{P_{p_3}}\mathbf{b}$, thus constituting a collective evidence of perception, denoted as CE_P **b**. Subsequently, Participant P_3 publicly announces that the scales are balanced ([!**b**]), providing further for **b** being true. Given the interpretive nature of announcements as indicative of belief, this declaration also implies that P_3 believes **b**, expressed as $B_{p_3}\mathbf{b}$. This announcement is subsequently publicly affirmed by Participant P_1, indicating not only possession of evidence for this proposition, but also belief in it, articulated as $B_{p_1}\mathbf{b}$. P_2 abstains from expressing public concurrence or objection to P_3's announcement, yet his active engagement in the collaborative endeavor implies a lack of contradictory evidence against CE_P **b**. Relying on the axiom "Non-contradictory

Evidence provides Belief," this participatory posture, coupled with the absence of contradictory evidence, leads to the inference that P_2 also subscribes to the belief in **b**. Consequently, a shared belief among all participants ensues, denoted as CB_P **b**.

The entirety of this process also contributes to the updating of common ground banks. Prior to the placement of the blocks onto the scale by P_1, all three participants maintained a shared belief that the red block weighed 10 g, thereby establishing its inclusion within the Fbank. Meanwhile, the weight of the blue block remained a query within the Qbank. Upon their observation of the balanced scale, each participant acquires a new piece of evidence, which is subsequently updated within the Ebank. The public announcement by P_3 and the subsequent public agreement by P_1 signify not only the possession of evidence supporting their respective statements but also a belief in said statements. Furthermore, as elucidated previously, P_2 also subscribes to the belief in **b**. Through inference that the scale is balanced and that the red block occupies one side while the blue block occupies the other, a logical deduction emerges: the weight of the blue block equals that of the red block, thereby also amounting to 10 g. This inference constitutes evidence for each participant, denoted as $E_{I_{p_2}}(b = 10)$, $E_{I_{p_1}}(b = 10)$, and $E_{I_{p_3}}(b = 10)$, where $b = 10$ signifies the proposition that the weight of the blue block is 10 g. This shared belief among the three participants is subsequently updated within the Fbank.

6 Annotations

6.1 Annotating Multimodal Modalities

Gesture Annotation. Gesture AMR is employed for the detailed annotation of gestures within our study. Each instance of Gesture AMR systematically categorizes content-bearing gestures into four distinct "gesture acts": deictic, iconic, emblematic, and metaphoric. Additionally, it meticulously records the gesturer, addressee, and semantic content conveyed by each gesture. In our methodology, we utilize the ELAN [57] software platform[2] where distinct tracks are designated for each speaker, facilitating the systematic analysis of gesture interactions. For instance, in Fig. 1, P_1 is observed directing attention of P_2 and P_3 towards the blocks and scale by means of pointing. The representation of this gesture is also within the figure.

Speech Annotation. The process of speech annotation encompasses transcription, speaker diarization, and segmentation into discrete utterances. The utterances are systematically integrated into corresponding tracks within ELAN, with

[2] ELAN serves as an annotation tool designed for the enhancement of audio and video recordings. It facilitates users in incorporating an extensive array of textual annotations onto audio and/or video recordings. These annotations may encompass sentences, individual words or glosses, comments, translations, or descriptions of observed features within the media.

distinct tracks designated for each speaker. Additionally, these speech transcripts underwent further refinement through dense paraphrasing [52], wherein sentence information and action annotations are amalgamated to enhance clarity by substituting pronouns with more explicit references within the original sentences (see Fig. 7).

Action Annotation. The actions executed by participants are systematically annotated within our study. We operate within a distinctly limited set of predicates suitable for modeling participant actions, primarily encompassing "putting" and "lifting." Additionally, we employ prepositions such as "on," "in," or "at," which delineate specific spatial relations within the VoxML framework [50]. For instance, in Fig. 4, P_1 is observed placing the red block on the right scale, and the annotation is `put(Red, on(RightScale))`.

Gaze Annotation. The orientation of eye gaze represents a pivotal marker demarcating the focal point of an individual's attentional engagement [14]. Consequently, we incorporate eye gaze direction as an additional source of evidence to discern whether a participant is attentively engaged in the experimental procedure or experiencing distraction. In Fig. 2, we use red arrows to indicate that P_1, P_2 and P_3 are all gazing at the scale/blocks.

Body Posture Annotation. Body postures represents a fundamental component of nonverbal communication, serving as conduits through which profound insights into people's internal states, such as, their engagement towards a joint activity [58]. Consequently, we integrate body posture within our multimodal annotation framework to discern whether participants are actively engaged in the experimental task, or experiencing boredom or agitation. In Fig. 2, the utilization of a blue arrow symbolizes P_2's inclination forward, indicative of a deliberate effort to scrutinize the scale closely. This behavior suggests an increased level of engagement with the experiment of the participant.

6.2 Common Ground Annotation

Building upon the multimodal annotations gathered in the preceding section, we conducted move-by-move tracking of the group's collective view of evidence and acceptance of task-relevant facts by introducing an additional layer of "common ground annotations" (CGA). The annotation process within the dialogue entails the identification of categories pertaining to participants' cognitive states, actions, and beliefs pertaining to the task at hand. These categories encompass: (a) OBSERVATION: participant P_i has perceived an action, a; (b) INFERENCE: deduction from φ; (c) STATEMENT: announcement of evidence φ; (d) QUESTION: introducing an interrogative role relating to φ; (e) ANSWER; supplying a filler to question about φ; (f) ACCEPT: agree with evidence φ; (g) DOUBT: negative evidence for φ.

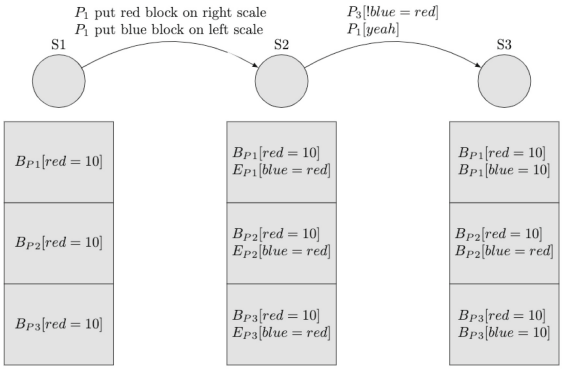

Fig. 10. Scenario: 3 participants engaged in determining weight of blue block.

Figure 10 illustrates an interaction depicted in Figs. 4, 5, 6, 7, 8 and 9, where s_1, s_2, and s_3 denote discrete states within the real world, with corresponding frames below each state, depicting the mental states of the three participants. P_1 places the blue and red blocks on opposing sides of the scale, resulting in a collective observation by all three participants of the scale balancing. Subsequently, they collectively extract a singular piece of information from this observation in the second mental state. Following this, the public announcements made by both P_3 and P_1 reflect their beliefs in the balanced state of the scale. As described in Sect. 4, in the absence of overt verbal or nonverbal cues indicating dissent towards P_3's announcement, P_2 conforms to the beliefs posited by both P_1 and P_3. Consequently, a common ground is established. The annotation of the common ground within the scenario is depicted in Fig. 11:

Common Ground Annotation

```
O1: OBSERVATION: on(RedBlock,  LeftScale) and on(BlueBlock,
 ↪  RightScale) and Scale: lean(Zero)
I1: INFERENCE-FROM (O1 and red = 10 )
S1:  STATEMENT( red = blue = 10)
ACCEPT(S1)
```

Fig. 11. Common Ground Annotation of Fig. 10

Common ground annotation focuses on propositions describing situations or facts contributed to the formation of shared belief among participants of a joint activity and the evidence for them as the dialogue progresses. Hence, in Fig. 11, action annotation is excluded from the common ground annotation. Instead, observations, inferences providing evidence for updating common ground, and statements incorporating propositions describing situations are all encompassed within the common ground annotation framework.

7 Classification of Belief in Multimodal HCI

In the example sequence given in Figs. 4, 5, 6, 7, 8 and 9, we can see how different modalities contribute to reviewing of pieces of evidence and the implicit and explicit construction of common ground.

1. In Fig. 4, P_1 puts the red block on the right side of the scale (from the perspective of the camera). At this time, both P_2 and P_3 are co-attending to the laptop, and may not have seen this action take place, or its result (the tilt of the scale).
2. Fig. 5. P_1 concludes the action from above ($put(Red, on(RightScale))$, but now P_2's gaze direction has shifted to the scale and the blocks. Under the assumption that gaze direction automatically equates to observing all visual content within the field of view, all three participants are now able to make the observation that the red block is on the right side of the scale ($on(Red, RightScale)$).
3. Fig. 6. P_1 puts the blue block on the left side of the scale. All three participants are now co-attending to the blocks and scale, and so all make the observation ($on(Blue, RightScale)$). It is important to note that at this step, according to our model, no evidence has been reviewed and no inference has yet been made.
4. In Fig. 7 all three participants, still co-attending to the scale, are able to observe that it is not leaning substantially to either side. P_3 makes this explicit with his utterance "It [the scale] seems pretty balanced", which is considered to be a statement of the proposition $red = blue$, and elevates this proposition into the EBank, as something that is evidenced but not yet agreed upon.
5. Subsequently (Fig. 8), P_1 says "Yeah", which is taken to be agreement with the above statement, thus elevating $red = blue$ to the FBank.
6. No one says anything in Fig. 9, but under a model where transitive closure takes place, an inference can be made that *blue=10*, even though the numerical value is never explicitly stated in the dialogue. This is in fact confirmed by the next utterance in the sequence (not shown), in which P_1 says, "Okay, so now we know that this [blue block] is also ten."

A couple of points should be noted regarding parameters of the model that affect when and how different kinds of evidencing is conducted. Many philosophical schools debate the level of epistemic validity to be assigned to direct perception vs. inference. Here we assume both to be equally valid (see Sect. 4), meaning that the inference in Fig. 9 would be directly elevated to FBank, but under other specifications of the model, this may not be the case. Additionally, we make an assumption here that gaze direction automatically means observation of content under that gaze, but under certain other assumptions (e.g., such as one in which all participants are not assumed to be paying close attention unless otherwise indicated), this would be softened. Finally, in Fig. 8, P_2 could have disagreed but didn't, and subsequently leans in toward the experimental apparatus. This is taken to be implicit agreement with P_1/P_3's positioning, but

this need not always be the case, and other models may require explicit acceptance by all parties to elevate a proposition from EBank to FBank.

8 Conclusion and Future Work

In this paper we have argued for the importance of Simulation Theory of Mind (SToM), encoded as an evidence-based dynamic epistemic logic (EB-DEL) for HCI particularly in the context of a multimodal task-oriented joint activity. We outlined a theory of perceptually-driven belief updating for multi-agent cooperative task completion. We extended the evidence-based dynamic epistemic logic from [6] to account for how perceptual evidence and inference interact and can cascade into strengthening an agent (or group) epistemic attitude towards a proposition, for updating the common ground. This subsequently provides situation-based epistemic data for tracking the common ground through a dialogue, by integrating the contributions of different modalities toward modeling the cognitive states of the group. Namely, by extracting the propositions expressed, and building common ground structures as the group proceeds through the task, our model holds potential for deployment in the creation of artificial agents proficient in simulating real-world situational settings. These agents would be adept at recording, adhering to, and comprehending common ground within collaborative activities. Such agents could find application in environments such as classrooms, where they can effectively monitor the collective knowledge of a group and foster productive collaborations [54].

By integrating ToM and common ground tracking into conversational agent architectures, we can better model the beliefs of participants by exposing unspoken assumptions of the participants or disagreements among them. Enhancing the epistemic modeling capabilities of multimodal HCI with ToM may also inform research in both Affective Computing, e.g., automatic emotion detection, by providing more contextualized interpretations of cognitive states and emotions in dialogue, and in providing support for those with functional impairments.

Acknowledgements. This work was supported in part by NSF grant DRL 2019805, to Dr. Pustejovsky at Brandeis University, and Dr. Krishnaswamy at Colorado State University. It was also supported in part by NSF grant CNS 2033932 to Dr. Pustejovsky. We would like to thank the reviewers for their comments and suggestions. The views expressed herein are ours alone.

References

1. Asher, N.: Common ground, corrections and coordination. J. Semant. **15**, 239–299 (1998)
2. Baltag, A., Moss, L.S., Solecki, S.: The logic of public announcements, common knowledge, and private suspicions. In: Arló-Costa, H., Hendricks, V.F., van Benthem, J. (eds.) Readings in Formal Epistemology. SGTP, vol. 1, pp. 773–812. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-20451-2_38

3. Barsalou, L.W.: Perceptions of perceptual symbols. Behav. Brain Sci. **22**(4), 637–660 (1999)
4. Belle, V., Bolander, T., Herzig, A., Nebel, B.: Epistemic planning: perspectives on the special issue. Artif. Intell. **316**, 103842 (2023)
5. van Benthem, J., Fernández-Duque, D., Pacuit, E.: Evidence and plausibility in neighborhood structures. Ann. Pure Appl. Logic **165**(1), 106–133 (2014)
6. van Benthem, J., Pacuit, E.: Dynamic logics of evidence-based beliefs. Stud. Logica. **99**, 61–92 (2011)
7. Bolander, T.: Seeing is believing: formalising false-belief tasks in dynamic epistemic logic. In: Jaakko Hintikka on Knowledge and Game-theoretical Semantics, pp. 207–236 (2018)
8. Bolander, T., Andersen, M.B.: Epistemic planning for single-and multi-agent systems. J. Appl. Non-Classical Logics **21**(1), 9–34 (2011)
9. Bolander, T., Jensen, M.H., Schwarzentruber, F.: Complexity results in epistemic planning. In: IJCAI, pp. 2791–2797 (2015)
10. Brutti, R., Donatelli, L., Lai, K., Pustejovsky, J.: Abstract meaning Representation for gesture. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 1576–1583. European Language Resources Association, Marseille, France, June 2022
11. Clark, H.H., Brennan, S.E.: Grounding in communication. Perspect. Socially Shared Cogn. **13**(1991), 127–149 (1991)
12. Dautenhahn, K.: Socially intelligent robots: dimensions of human-robot interaction. Philos. Trans. R. Soc. B: Biol. Sci. **362**(1480), 679–704 (2007)
13. De Groote, P.: Type raising, continuations, and classical logic. In: Proceedings of the Thirteenth Amsterdam Colloquium, pp. 97–101 (2001)
14. Dey, I., Puntambekar, S.: Examining nonverbal interactions to better understand collaborative learning. In: Proceedings of the 16th International Conference on Computer-Supported Collaborative Learning-CSCL 2023, pp. 273–276. International Society of the Learning Sciences (2023)
15. Dissing, L., Bolander, T.: Implementing theory of mind on a robot using dynamic epistemic logic. In: IJCAI, pp. 1615–1621 (2020)
16. Eijck, J.: Perception and change in update logic. In: van Eijck, J., Verbrugge, R. (eds.) Games, Actions and Social Software. LNCS, vol. 7010, pp. 119–140. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29326-9_7
17. Feldman, J.: Embodied language, best-fit analysis, and formal compositionality. Phys. Life Rev. **7**(4), 385–410 (2010)
18. Feldman, R.: Respecting the evidence. Philos. Perspect. **19**, 95–119 (2005)
19. Geib, C., George, D., Khalid, B., Magnotti, R., Stone, M.: An integrated architecture for common ground in collaboration (2022)
20. Gianotti, M., Patti, A., Vona, F., Pentimalli, F., Barbieri, J., Garzotto, F.: Multimodal interaction for persons with autism: the 5A case study. In: Antona, M., Stephanidis, C. (eds.) Universal Access in Human-Computer Interaction, HCII 2023. LNCS, vol. 14020, pp. 581–600. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-35681-0_38
21. Ginzburg, J.: Interrogatives: Questions, Facts and Dialogue. The Handbook of Contemporary Semantic Theory, pp. 359–423. Blackwell, Oxford (1996)
22. Ginzburg, J.: The Interactive Stance: Meaning for Conversation. OUP, Oxford (2012)
23. Goldman, A.I.: In defense of the simulation theory. Mind Lang. **7**(1–2), 104–119 (1992)

24. Goldman, A.I.: Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading. Oxford University Press, Oxford (2006)
25. Gopnik, A.: How we know our minds: the illusion of first-person knowledge of intentionality. Behav. Brain Sci. **16**(1), 1–14 (1993)
26. Gordon, R.M.: Folk psychology as simulation. Mind Lang. **1**(2), 158–171 (1986)
27. Heal, J.: Simulation, Theory, and Content. Theories of Theories of Mind, pp. 75–89 (1996)
28. Henderson, M., Thomson, B., Williams, J.D.: The second dialog state tracking challenge. In: Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pp. 263–272 (2014)
29. Khebour, I., et al.: The weights task dataset: a multimodal dataset of collaboration in a situated task. J. Open Humanities Data **10** (2024)
30. Kolve, E., et al.: AI2-THOR: an interactive 3D environment for visual AI. arXiv preprint arXiv:1712.05474 (2017)
31. Krishnaswamy, N., et al.: Diana's World: a situated multimodal interactive agent. In: AAAI Conference on Artificial Intelligence (AAAI): Demos Program. AAAI (2020)
32. Krishnaswamy, N., Pustejovsky, J.: VoxSim: a visual platform for modeling motion language. In: Proceedings of COLING 2016, The 26th International Conference on Computational Linguistics: Technical Papers. ACL (2016)
33. Krishnaswamy, N., Pustejovsky, J.: Multimodal continuation-style architectures for human-robot interaction. arXiv preprint arXiv:1909.08161 (2019)
34. Krshnaswamy, N., Pickard, W., Cates, B., Blanchard, N., Pustejovsky, J.: VoxWorld platform for multimodal embodied agents. In: LREC Proceedings, vol. 13 (2022)
35. Miller, P.W.: Body language in the classroom. Tech. Connecting Educ. Careers **80**(8), 28–30 (2005)
36. Narayanan, S.: Mind changes: a simulation semantics account of counterfactuals. Cognitive Science (2010)
37. Pacuit, E.: Neighborhood Semantics for Modal Logic. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67149-9
38. Plaza, J.: Logics of public communications. In: Proceedings 4th International Symposium on Methodologies for Intelligent Systems, pp. 201–216 (1989)
39. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behav. Brain Sci. **1**(4), 515–526 (1978)
40. Pustejovsky, J., Krishnaswamy, N.: VoxML: a visualization modeling language. arXiv preprint arXiv:1610.01508 (2016)
41. Pustejovsky, J., Krishnaswamy, N.: Embodied human computer interaction. KI-Künstliche Intelligenz **35**(3–4), 307–327 (2021)
42. Pustejovsky, J., Krishnaswamy, N.: The role of embodiment and simulation in evaluating HCI: theory and framework. In: Duffy, V.G. (ed.) HCII 2021. LNCS, vol. 12777, pp. 288–303. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-77817-0_21
43. Radu, I., Tu, E., Schneider, B.: Relationships between body postures and collaborative learning states in an augmented reality study. In: Bittencourt, I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, 6–10 July 2020, Proceedings, Part II 21, pp. 257–262. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52240-7_47

44. Savva, M., et al.: Habitat: a platform for embodied AI research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9339–9347 (2019)
45. Schneider, B., Pea, R.: Does seeing one another's gaze affect group dialogue? A computational approach. J. Learn. Analytics **2**(2), 107–133 (2015)
46. Sousa, A., Young, K., D'aquin, M., Zarrouk, M., Holloway, J.: Introducing CALMED: multimodal annotated dataset for emotion detection in children with autism. In: Antona, M., Stephanidis, C. (eds.) International Conference on Human-Computer Interaction, pp. 657–677. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-35681-0_43
47. Stalnaker, R.: Common ground. Linguist. Philos. **25**(5–6), 701–721 (2002)
48. Sun, C., Shute, V.J., Stewart, A., Yonehiro, J., Duran, N., D'Mello, S.: Towards a generalized competency model of collaborative problem solving. Comput. Educ. **143**, 103672 (2020)
49. Suzuki, R., Karim, A., Xia, T., Hedayati, H., Marquardt, N.: Augmented reality and robotics: a survey and taxonomy for AR-enhanced human-robot interaction and robotic interfaces. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1–33 (2022)
50. Tam, C., Brutti, R., Lai, K., Pustejovsky, J.: Annotating situated actions in dialogue. In: Proceedings of the 4th International Workshop on Designing Meaning Representation (2023)
51. Tolzin, A., Körner, A., Dickhaut, E., Janson, A., Rummer, R., Leimeister, J.M.: Designing pedagogical conversational agents for achieving common ground. In: Gerber, A., Baskerville, R. (eds.) International Conference on Design Science Research in Information Systems and Technology, pp. 345–359. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-32808-4_22
52. Tu, J., Rim, K., Pustejovsky, J.: Competence-based question generation. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 1521–1533 (2022)
53. Van Fraassen, C.: Belief and the will. J. Philos. **81**(5), 235–256 (1984)
54. VanderHoeven, H., et al.: Multimodal design for interactive collaborative problem-solving support. In: HCII 2024. Springer, Cham (2024)
55. Wellman, H.M., Carey, S., Gleitman, L., Newport, E.L., Spelke, E.S.: The Child's Theory of Mind. The MIT Press, Cambridge (1990)
56. Wimmer, H., Perner, J.: Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition **13**(1), 103–128 (1983)
57. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: ELAN: a professional framework for multimodality research. In: 5th LREC 2006, pp. 1556–1559 (2006)
58. Won, A.S., Bailenson, J.N., Janssen, J.H.: Automatic detection of nonverbal behavior predicts learning in dyadic interactions. IEEE Trans. Affect. Comput. **5**(2), 112–125 (2014)
59. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson ENV: real-world perception for embodied agents. In: Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, pp. 9068–9079 (2018)