Action understanding as a social-cognitive benchmark for human-aligned robotics

Author Names Omitted for Anonymous Review.

***Abstract*—As robots increasingly operate in everyday human environments, their success depends not only on task completion but also on how their actions are perceived and understood. This paper introduces an *action understanding framework* as a social-cognitive benchmark for evaluating whether robot behavior supports interpretability, transparency, and trust. We identify ten psychologically grounded action types and propose benchmark tasks that reflect core human cognitive mechanisms. By prioritizing cognitive legibility over performance metrics, the framework provides a principled and inclusive basis for designing and evaluating robots that are socially aligned, interpretable, and trustworthy.**

Action understanding; human-aligned robotics; cognitive science; human-robot interaction; trust and transparency in robots

I. Introduction

As robots increasingly enter human environments, their success depends not only on task execution but also on how people perceive and understand their actions. While current benchmarks emphasize task performance, robustness, and generalization, real-world employment also hinges on cognitive and social alignment with human users. A robot may complete a task flawlessly yet fail in collaborative settings if its behavior is ambiguous, unintuitive, or socially inappropriate—a challenge amplified when interacting with non-experts, such as children, older adults, or first-time users.

To address this, we propose an ***action understanding framework*** as a social-cognitive benchmark for evaluating human alignment in robotics. Drawing from cognitive science, developmental psychology, and human-robot interaction, the framework emphasizes **interpretability**, **transparency**, and **social resonance**. It assesses whether a robot's actions support shared understanding, enable context-based inference, and align with the diverse expectations of various users. By foregrounding action understanding, we provide a developmentally grounded approach to designing robots that are not only effective but also intelligible and trustworthy.

Importantly, our framework does not depend on anthropomorphism or human mimicry. While anthropomorphic features may trigger social responses [1], our approach prioritizes the functionality of understanding. That is, we center action understanding on whether robot behavior is interpretable through the same cognitive mechanisms humans use to understand others. This shift from superficial cues to cognitive resonance offers a more principled and inclusive model of alignment, rooted in the developmental foundations of social understanding.

II. Theoretical foundations

To build robots that align with human expectations, we must understand how people interpret the actions of others. Research in cognitive science and developmental psychology indicates that social understanding develops through embodied interaction, symbolic learning, causal reasoning, and the development of a theory of mind. These mechanisms enable humans to infer the goals and intentions of others from their observable behavior. Grounding our action understanding framework in these processes enables us to evaluate not only task success but also whether robot behavior aligns with human interpretive strategies. This section outlines the cognitive foundations of socially aligned robot action.

A. Embodied cognition

Embodied cognition posits that cognitive processes are rooted in the body's physical interactions with the world [2]. From this view, action understanding involves simulating observed behavior via the sensorimotor system. Research with infants shows that motor experience enhances perception: 3-month-olds trained to reach better interpreted others' goal-directed action than untrained peers [3], suggesting that humans use motor representations to predict others' behavior [4]. Embodied cognition thus provides a powerful account that explains how people infer meaning from motion without relying on language or symbolic reasoning.

In robotics, embodied cognition implies that action should evoke sensorimotor resonance to be cognitively legible. Studies in human-robot interaction show that people better interpret robot behavior when its motion (e.g., speed, trajectory, timing) mirrors human-like kinematics [5], suggesting that designing robot motion to align with embodied cognitive mechanisms enhances understanding of actions.

B. Symbolic learning

Symbolic learning is the capacity to understand that one entity (e.g., gesture, image, or word) can represent another. This ability underpins abstract reasoning, communication, and cultural development [6]. Children often struggle with dual representation, such as treating a scale model as an object rather than a symbol, underscoring that symbolic understanding develops gradually with experience or cognitive maturation [7]. In social contexts, symbolic competence enables individuals to

interpret indirect cues and shared meaning across time and space [8].

Robots frequently use symbolic cues (e.g., lights, gestures, screens, or speech) to communicate. For these to be effective, users must perceive them as meaningful, not arbitrary. Research indicates that symbolic congruence, or alignment between cue and intent, influences trust and comprehension. Mismatched gestures reduce perceived competence [9], whereas human-like deictic cues enhance coordination and intention inference [10]. The developmental trajectory of symbolic reasoning also suggests that robot cues should be tailored to users' cognitive abilities: What adults interpret easily may be opaque to children. Designing symbolically meaningful, context-sensitive actions is therefore essential for fostering shared understanding.

C. *Causal reasoning*

Causal reasoning is the ability to infer the mechanisms behind observed events. It underlies our ability to distinguish between deliberate and accidental actions and to predict outcomes [11]. Even infants demonstrate a basic understanding of causality, for example, expecting one object to move another upon contact [12]. By preschool age, children begin using causal cues to evaluate whether actions are intentional, effective, or erroneous [13]. As they develop, children employ epistemic actions, such as poking or rotating, to test causal hypotheses and gain a deeper understanding of how things work [14]. These shifts support more nuanced interpretations of others' behavior by linking actions to internal mental states.

In robotics, causal interpretability is crucial for establishing trust and fostering collaboration. Robots that fail without explanation may be perceived as unreliable, unless users can infer whether the failure stems from mechanical fault, environmental constraint, or miscommunication [15]. Robots that signal causal reasoning, for example, by retrying a failed grasp, pausing to signal uncertainty, or offering feedback, promote user trust. Motions that invite epistemic interaction (e.g., slowing down when uncertain or gesturing toward relevant causes) support causal understanding and social resilience in the face of error [16].

D. *Theory of mind*

Theory of mind refers to the capacity to attribute mental states, such as beliefs, desires, intentions, and knowledge, to others and to use those attributions to explain, predict, and coordinate social behavior [17]. A significant developmental milestone is understanding false beliefs, which typically emerges around age four, allowing children to recognize that others can act based on inaccurate information [18]. Theory of mind enables children to infer what others see, know, or expect, supporting more sophisticated forms of communication, teaching, and joint attention [8]. In action understanding, the theory of mind enables observers to interpret behavior based on beliefs and intentions, rather than just visible outcomes.

In robotics, theory of mind is critical for enabling socially aligned behavior. Robots in collaborative settings must signal awareness of human mental states. Users judge robots' social intelligence based on cues like gaze-following, error correction, and perspective-taking [19]. Robots perceived as "mind-aware" are rated as more competent and trustworthy [20]. Benchmarks may include false-belief tasks (e.g., the robot adapts its behavior based on what the human partner falsely believes) or information asymmetry scenarios (e.g., the robot helps only when the user lacks necessary information), where robots adapt their behavior to users' mental states. Such capabilities enhance interpretability, adaptability, and trust in real-world interactions.

III. BENCHMARKING DIMENSIONS AND APPLICATIONS

To assess whether robot behavior aligns with human cognition, we propose benchmarking dimensions based on psychological classifications of action. These dimensions capture how people interpret behavior, not only goal-directed tasks, but also in social, exploratory, and communicative contexts. Structured as benchmark tasks, they assess whether robot actions promote cognitive legibility and social alignment. The following sections present ten core action types, along with evaluation scenarios, to guide the design of interpretable, relatable, and trustworthy robots.

A. *Epistemic actions: Acting to learn*

Epistemic actions are behaviors performed not to accomplish an immediate external goal, but to gain information, reduce uncertainty, or simplify internal computation [14]. Examples include touching or rotating objects to clarify input or test hypotheses. These actions emerge early in development as children explore to learn about causal and spatial relationships [21].

In robotics, epistemic actions serve as a benchmark for visible information-seeking and self-monitoring. Benchmark tasks might assess whether robots perform diagnostic behaviors, such as rotating an object before placement or pausing to inspect after a failure. Such actions signal adaptive reasoning and promote user trust by making cognitive effort visible. Robots that visibly "check" or "investigate" promote user trust by signaling transparency and cognitive effort, helping users interpret errors as part of an active problem-solving process rather than a passive failure and supporting trust repair.

B. *Exploratory actions: Acting to discover*

Exploratory actions are curiosity-driven behaviors that aim to uncover the unknown properties of objects, environments, or causal relationships. Unlike epistemic actions, which address specific ambiguities, exploratory actions reflect a general drive to gather information without a clear problem. Infants and young children often exhibit such behaviors in response to novelty, which helps them learn about affordances and build internal models of the world [21, 22].

To benchmark exploratory capacity, robots can be placed in unfamiliar environments or presented with novel objects, such as twistable or compressible ones. Tasks assess whether the robot engages in structured exploration, such as trying different actions, observing outcomes, or adapting based on feedback. Observers may also rate whether the behavior appears "curious" or "investigative." Research suggests that legible exploration fosters user trust and encourages human-robot co-learning, particularly in autonomous and adaptive systems [23].

C. Pragmatic actions: Action to do

Pragmatic actions are goal-directed behaviors that aim to alter the physical environment, such as grasping or pressing. They are typically deliberate, efficient, and task-focused, forming the basis of sensorimotor cognition and problem-solving [24]. In both humans and robots, these actions are expected to follow predictable patterns that reflect planning and intentionality.

To benchmark the legibility of pragmatic actions, robots can be placed in a goal-oriented task, such as object retrieval, navigation, or tool use. For example, in a cluttered scene, does the robot reach around obstacles or adjust its posture to signal which item it intends to grasp? Actions that visibly align with task goals help users make accurate predictions, build confidence in the robot's competence, and support smooth coordination. This dimension is especially critical in collaborative, time-sensitive, or high-stakes settings, where ambiguity may lead to safety risks or task failure. Legible pragmatic behavior promotes both functional success and human trust in shared environments.

D. Goal-directed actions: Action with intention

Goal-directed actions are intentional behaviors aimed at achieving an internally represented outcome. They reflect an agent's capacity for planning, anticipation, and persistence in the face of obstacles. In human development, sensitivity to goal-directedness emerges early: by six months, infants can infer purposeful action [25], and by the end of the first year, they can distinguish between intentional and accidental acts [13]. This ability is foundational to social cognition, enabling observers to attribute agency, infer hidden intentions, and adjust their behavior accordingly. Unlike pragmatic actions, which emphasize the physical completion of a task, goal-directed actions highlight the internal intention driving behavior, regardless of whether the task is completed.

Robots can be benchmarked for goal-directed legibility through tasks involving competing affordances or distractors. For instance, when presented with two identical cups but only one of which contains an object, does the robot act in a way that clearly reveals its intended choice? Does it adjust its trajectory, gaze, or posture early enough for human observers to predict its goal? Benchmark success can be measured by both task completion and the extent to which human observers can infer the robot's intention before the action concludes. Alignment between internal goals and external cues enhances predictability, trust, and collaborative fluency in human-robot interaction [5].

E. Habitual and reflective actions: Acting automatically

Habitual actions are behaviors triggered automatically by familiar environmental cues, often with minimal cognitive effort or conscious deliberation. These routines emerge through repetition and reinforcement, offering efficient strategies for navigating stable contexts [26]. While they conserve cognitive resources, habitual actions can become maladaptive in dynamic environments that require flexibility.

To benchmark habitual action alignment, robots can be assigned learned routines, such as refilling a cup when a red light blinks. After the behavior is established, the context is altered (e.g., the cup is missing or the light changes color), and observers assess whether the robot adapts or continues to behave rigidly. Success is evaluated based on behavioral flexibility and whether observers interpret the action as habitual or deliberate. This benchmark reflects real-world situations where users must discern whether a robot is operating on "autopilot" and decide when to intervene, highlighting the trade-off between efficiency and adaptability in socially aligned robotics.

F. Reflective actions: Acting with deliberation

Reflective actions are deliberative behaviors guided by conscious evaluation, internal goals, and social reasoning. Unlike habitual actions, they involve metacognition, self-monitoring, and the capacity to inhibit automatic responses when flexibility or ethical judgment [27]. Reflective control is particularly important in contexts involving uncertainty, moral ambiguity, or multi-agent interactions, such as resolving conflicting priorities or adapting to shifting social norms. In developmental trajectories, these abilities mature later than habitual responses and are linked to executive function and social-emotional regulation.

Robots can be benchmarked for reflective behavior through scenarios that involve conflict resolution, perspective-taking, or value-sensitive decision-making. For instance, a robot may receive conflicting commands from two users or need to decide whether interrupting a speaker is appropriate. Observers assess whether the robot shows signs of deliberation, such as hesitating, seeking clarification, or adapting to social cues. Robots that offer context-aware justifications (e.g., "I paused because I wasn't sure who to follow") are more likely to be perceived as reflective and socially intelligent. This benchmark is especially relevant in domains such as healthcare, education, and collaborative decision-making, where thoughtful and deliberate responses are often preferred over speed or automation.

G. Affordance-driven actions: Acting with the environment

Affordance-driven actions are guided by the perceived possibilities for action provided by the environment. Based on Gibson's ecological theory of perception, affordances refer to the actionable properties of objects or settings that emerge through interaction between the agent and its surroundings [28]. Infants gradually learn to perceive and act upon affordances, such as grasping a handle, through trial and error, refining their understanding of object functionality over time [29]. This perception-action coupling enables adaptive and fluent behavior without requiring abstract or symbolic reasoning in every situation.

To benchmark affordance sensitivity in robots, tasks can present objects with distinctive physical features (e.g., handles, squishy surfaces, or buttons). The robot's actions should demonstrate appropriate responses to these properties (e.g., pulling, compressing, or pressing), and its orientation, posture, or trajectory should make its intention interpretable to human observers. When robots act in ways that ignore affordance constraints, such as pushing an object that should be pulled, they may appear less adaptable or intelligent. This benchmark

is crucial for assessing whether a robot can behave in physically intuitive, context-sensitive ways, especially in shared human environments like homes, classrooms, and collaborative workplaces.

H. Social actions: Acting with others in mind

Social actions are behaviors intentionally performed in response to the presence, attention, or inferred expectations of others. These actions are inherently communicative, relying on cues such as gaze, timing, body orientation, and turn-taking to coordinate joint activity. A key developmental milestone is joint attention—the ability to direct or allow another's gaze, which arises in infancy and serves as a foundation for language acquisition, collaboration, and shared intentionality [30]. In adulthood, social actions impact rapport, cooperation, and mutual understanding.

Robots can be benchmarked for social action through tasks that require initiating or responding to joint attention or coordinating shared activities. For instance, a robot might point to an object and alternate its gaze between the object and the user. Observers assess whether the behavior appears referential, engaging, or socially attuned. This dimension is particularly relevant in child-robot interaction, where users rely heavily on nonverbal cues to infer the robot's intent. Studies show that behaviors such as gaze-following and gesture timing shape perceptions of a robot's social intelligence [10, 31]. This benchmark is crucial for collaborative, assistive, and educational contexts, where mutual attention, responsiveness, and social alignment are essential for effective interaction.

I. Expressive actions: Acting to communicate emotion

Expressive actions convey internal emotion or mental states through nonverbal cues such as body movements, gestures, vocalizations, or facial expressions. These actions can be voluntary or involuntary and serve communicative functions even without speech. From early infancy, expressive behaviors (e.g., smiling or crying) signal needs, regulate social interaction, and foster emotional bonding [32]. As children develop, they learn to interpret such cues as reflections of others' intentions and feelings, supporting empathy, prosocial behavior, and social learning.

Robots can be evaluated for expressive competence in emotionally charged scenarios, such as succeeding or failing at a task (e.g., assembling a puzzle vs. dropping a block). In one condition, the robot responds using expressive modalities, such as exaggerated motion, vocal prosody, light signals, or animated facial expressions, while in another, it reacts neutrally. Benchmark tasks assess whether observers interpret the robot's timing, posture, and multimodal signals as emotionally appropriate or socially meaningful. Expressive behavior enhances emotional transparency and user engagement, which are particularly crucial in educational, care, and therapeutic settings [33], making robots more emotionally intelligible and relatable partners in human-robot interaction.

J. Repair actions: Acting to restore understanding

Repair actions are behaviors aimed at re-establishing mutual understanding after breakdowns in communication, coordination, or task execution. These behaviors may include repetition, clarification, rephrasing, apology, or adjustments in timing and delivery. Even toddlers begin using basic verbal and gestural repairs (e.g., repeating or modifying an utterance) when they recognize a communication failure [34]. In adult interaction, repair is a hallmark of conversational alignment and social resilience, helping interlocutors maintain shared meaning and trust despite occasional missteps [35].

To benchmark repair behaviors in robots, evaluators can embed them in intentional failures or ambiguities, such as providing the wrong object or misinterpreting a command. Observers assess whether the robot repeats a gesture, pauses for clarification, or acknowledges the error. The effectiveness of a repair can be measured by how well it restores task flow, rapport, and clarity, or prevents escalation. This benchmark is essential for long-term human-robot interaction, where maintaining trust, transparency, and fluency supports robust autonomy in dynamic, human-centered environments.

IV. Conclusion and Future Work

By treating psychologically grounded action types as benchmarking dimensions, we capture the diversity of human interpretive strategies and apply them to evaluate how well robot behavior aligns with users' cognitive and social expectations. Together, they form an action understanding framework that prioritizes interpretability, adaptability, and social resonance over task success alone.

This framework expands the notion of human alignment by grounding evaluation in the same mechanisms that humans, especially children and non-experts, use to understand others' actions. Rather than relying on anthropomorphic features or performance metrics, it prioritizes whether robot behavior is interpretable through core cognitive processes, such as goal inference, causal reasoning, and joint attention.

Future work will involve empirically validating these benchmarks across a range of real-world scenarios, including interactions involving uncertainty, errors, and social misalignment. We also aim to integrate these benchmarks into robot learning algorithms and policy optimization, ensuring that robots not only learn effectively but also do so in ways that remain transparent, predictable, and trustworthy to humans. Ultimately, this work contributes toward building robots that are not only capable but also cognitively and socially aligned partners in everyday human environments.

References

[1] B. R. Duffy, "Anthropomorphism and the social robot," *Rob. Auton. Syst.*, vol. 42, no. 3–4, pp. 177–190, Mar. 2003.

[2] L. W. Barsalou, "Grounded cognition," *Annu. Rev. Psychol.*, vol. 59, no. 1, pp. 617–645, 2008.

[3] J. A. Sommerville, A. L. Woodward, and A. Needham, "Action experience alters 3-month-old infants' perception of others' actions," *Cognition*, vol. 96, no. 1, pp. B1–11, May 2005.

[4] G. Rizzolatti and C. Sinigaglia, "The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations," *Nat. Rev. Neurosci.*, vol. 11, no. 4, pp. 264–274, Apr. 2010.

[5] A. D. Dragan, K. C. T. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *2013 8th ACM/IEEE International*

Conference on Human-Robot Interaction (HRI), IEEE, Mar. 2013. doi: 10.1109/hri.2013.6483603.

[6] L. S. Vygotsky and M. Cole, *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, 1978.

[7] J. S. DeLoache, "Early Understanding and Use of Symbols: The Model Model," *Curr. Dir. Psychol. Sci.*, vol. 4, no. 4, pp. 109–113, Aug. 1995.

[8] M. Tomasello, *The cultural origins of human cognition*. London, England: Harvard University Press, 2009.

[9] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability," *Int. J. Soc. Robot.*, vol. 5, no. 3, pp. 313–323, Aug. 2013.

[10] B. Mutlu, J. Forlizzi, and J. Hodgins, "A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior," 2006. doi: 10.1109/ichr.2006.321322.

[11] A. Gopnik and L. Schulz, "Mechanisms of theory formation in young children," *Trends Cogn. Sci.*, vol. 8, no. 8, pp. 371–377, Aug. 2004.

[12] A. M. Leslie, "Spatiotemporal continuity and the perception of causality in infants," *Perception*, vol. 13, no. 3, pp. 287–305, 1984.

[13] A. N. Meltzoff, "Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children," *Dev. Psychol.*, vol. 31, no. 5, pp. 838–850, Sep. 1995.

[14] D. Kirsh, "On distinguishing epistemic from pragmatic action," *Cogn. Sci.*, vol. 18, no. 4, pp. 513–549, Dec. 1994.

[15] B. F. Malle, S. Guglielmo, and A. E. Monroe, "A theory of blame," *Psychol. Inq.*, vol. 25, no. 2, pp. 147–186, Apr. 2014.

[16] E. J. de Visser, R. Pak, and T. H. Shaw, "From 'automation' to 'autonomy': the importance of trust repair in human-machine interaction," *Ergonomics*, vol. 61, no. 10, pp. 1409–1427, Oct. 2018.

[17] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?," *Behav. Brain Sci.*, vol. 1, no. 4, pp. 515–526, Dec. 1978.

[18] H. M. Wellman, D. Cross, and J. Watson, "Meta-analysis of theory-of-mind development: the truth about false belief," *Child Dev.*, vol. 72, no. 3, pp. 655–684, May 2001.

[19] B. Scassellati, "Theory of Mind for a Humanoid Robot," *Auton. Robots*, vol. 12, no. 1, pp. 13–24, Jan. 2002.

[20] C. Breazeal, D. Buchsbaum, J. Gray, D. Gatenby, and B. Blumberg, "Learning from and about others: towards using imitation to bootstrap the social understanding of others by robots," *Artif. Life*, vol. 11, no. 1–2, pp. 31–62, 2005.

[21] A. Gopnik, A. N. Meltzoff, and P. K. Kuhl, *The scientist in the crib: What early learning tells us about the mind*. William Morrow & Co*, 1999.*

[22] D. E. Berlyne, *Conflict, arousal, and curiosity*. New York: McGraw-Hill Book Company, 1960.

[23] G. Gordon and C. Breazeal, "Bayesian active learning-based robot tutor for children's word-reading skills," *Proc. Conf. AAAI Artif. Intell.*, vol. 29, no. 1, Feb. 2015, doi: 10.1609/aaai.v29i1.9376.

[24] M. Jeannerod, *The Cognitive Neuroscience of Action*. in Fundamentals of Cognitive Neuroscience. London, England: Blackwell, 1997.

[25] A. L. Woodward, "Infants selectively encode the goal object of an actor's reach," *Cognition*, vol. 69, no. 1, pp. 1–34, Nov. 1998.

[26] W. Wood and D. Rünger, "Psychology of habit," *Annu. Rev. Psychol.*, vol. 67, no. 1, pp. 289–314, 2016.

[27] J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry," *Am. Psychol.*, vol. 34, no. 10, pp. 906–911, Oct. 1979.

[28] J. J. Gibson, *The Ecological Approach To Visual Perception*. Psychology Press, 2013.

[29] K. E. Adolph and J. E. Hoch, "Motor development: Embodied, embedded, enculturated, and enabling," *Annu. Rev. Psychol.*, vol. 70, no. 1, pp. 141–164, Jan. 2019.

[30] Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). Lawrence Erlbaum Associates, Inc.

[31] H. Admoni and B. Scassellati, "Social Eye Gaze in Human-Robot Interaction: A Review," *J. Hum. Robot Interact.*, vol. 6, no. 1, p. 25, Mar. 2017.

[32] E. Z. Tronick, "Emotions and emotional communication in infants," *Am. Psychol.*, vol. 44, no. 2, pp. 112–119, Feb. 1989.

[33] C. Breazeal, "Emotion and sociable humanoid robots," *Int. J. Hum. Comput. Stud.*, vol. 59, no. 1–2, pp. 119–155, Jul. 2003.

[34] R. M. Golinkoff, "'I beg your pardon?': the preverbal negotiation of failed messages," *J. Child Lang.*, vol. 13, no. 3, pp. 455–476, Oct. 1986.

[35] E. A. Schegloff, G. Jefferson, and H. Sacks, "The preference for self-correction in the organization of repair in conversation," *Language (Baltim.)*, vol. 53, no. 2, p. 361, Jun. 1977.