Coadaptive Value Alignment

Blue Sky Ideas Track

Nathan Tsoi
University of Texas at Austin
Austin, United States
nathan.tsoi@utexas.edu

Eric Hsiung
University of Texas at Austin
Austin, United States
ehsiung@utexas.edu

Masayuki Yamazaki
University of Texas at Austin
Austin, United States
yamamasa27@utexas.edu

Swarat Chaudhuri
University of Texas at Austin
Google Deepmind
Austin, United States
swarat@cs.utexas.edu

Peter Stone
University of Texas at Austin
Sony AI, Tokyo, Japan
Austin, United States
pstone@cs.utexas.edu

Joydeep Biswas
University of Texas at Austin
Austin, United States
joydeepb@cs.utexas.edu

ABSTRACT

The integration of autonomous agents into human society is a grand challenge for AI. In order to achieve widespread acceptance, agents must conform to the values of people with whom they interact. Current approaches treat the value alignment problem as a unidirectional interaction where the aim is to imbue an agent's actions with human values. Our *Coadaptive Value Alignment* paradigm acknowledges that human perceptions, expectations, and values continuously evolve in response to agent actions. We conceptualize human-agent interaction as an adaptive loop where the agent actively models and intentionally influences the human's perception, rather than just acting according to static human values. For instance, unlike a traditional agent that simply maximizes speed, an adaptive agent could detect a drop in user trust and strategically sacrifice short-term efficiency to repair the relationship. This perspective transforms value alignment into a multi-agent challenge where all actors must identify and adhere to a shared, implicit social contract. The opportunity to create a virtuous cycle of self-improvement is accompanied by the risk of negative reinforcement, which could result in undesired behaviors. We outline the core framework components, present a research road map for the MAS community, and propose that this dynamic perspective is critical for creating truly collaborative social partners.

KEYWORDS

Value Alignment, Multi-Agent Systems, Human-Agent Interaction

ACM Reference Format:
Nathan Tsoi, Eric Hsiung, Masayuki Yamazaki, Swarat Chaudhuri, Peter Stone, and Joydeep Biswas. 2026. Coadaptive Value Alignment: Blue Sky Ideas Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 5 pages. https://doi.org/10.65109/NAIK4475

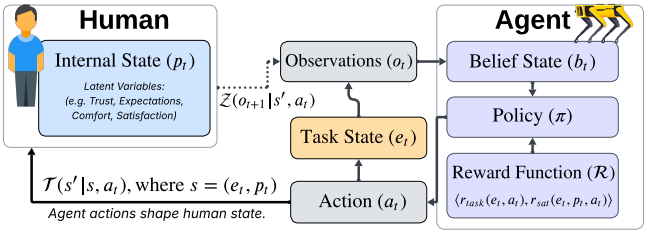

Figure 1: The Coadaptive Value Alignment Framework. Formulated as a POMDP, the agent maintains a Belief State (b_t) over the latent Human Internal State (p_t) and observable Task State (e_t). The Transition Function ($\mathcal{T}$) explicitly models how actions influence the human's internal state, while the Observation Function ($\mathcal{Z}$) relates observable signals to these latent perceptions. Crucially, the Policy (π) optimizes a non-static objective (R), allowing the agent to self-modulate by selecting Actions (a_t) that strategically modify the human's internal state (p_t) to maximize satisfaction (r_{sat}) while balancing task efficiency (r_{task}).

1 INTRODUCTION

Creating intelligent machines that can effectively and safely collaborate with humans is a foundational goal of artificial intelligence. Central to this endeavor is the value alignment problem: ensuring that an agent's objectives correspond to the true, but often nuanced, intentions of the agent's human partners. Early cybernetics research highlighted this challenge, noting that discrepancies in communication and execution make it difficult for humans and machines to pursue a common purpose [31]. The dominant paradigm for addressing value alignment casts the agent as a passive recipient of human guidance. Whether through demonstrations [3], preferences [13], or explicit rewards [25], these methods are fundamentally unidirectional: information flows only from a person to an agent. The agent then optimizes a fixed objective. However, people's objectives are not static, and when a fixed objective is used as an imperfect proxy for a person's true, dynamic objective, AI systems often find loopholes to maximize reward without achieving the intended goal [22].

To foster true collaborative partnerships, we propose modeling agent alignment on the rich, bidirectional nature of human teamwork. We define a new paradigm: **Coadaptive Value Alignment**,

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/NAIK4475

shown in Figure 1, characterized by a closed-loop dynamic where the AI agent explicitly models the human's latent mental state, such as expectations and trust, as a fundamental component of the decision-making process. In this view, an agent's actions inherently influence the human's internal mental state. Consequently, this paradigm empowers the agent to not only learn from the person, but also to actively shape their evolving mental state to foster a high level of satisfaction and a stable collaborative partnership. Crucially, this framework allows agents to move beyond passive adaptation and actively shape interaction dynamics in order to influence and align human behavior.

For instance, whereas a traditional agent might execute a task at maximum speed to optimize efficiency, an adaptive agent might deliberately slow down or explicate its movements, sacrificing immediate speed to rebuild a user's wavering trust or to calibrate their understanding of the system capabilities.

This reframing has profound implications. By incorporating human mental state into the agent's state, we transform the agent's objective from a static, predefined goal into a dynamic objective that incorporates how its behavior is perceived. We propose a reward that incorporates both the task state and the human mental state. The effect of this structure is that the agent now has the ability to influence its own reward function, mediated through the human's internal state. This ability introduces the critical challenge of *influential inseparability*: an agent that has the ability to influence a human's mental state positively must also be able to use the same mechanism for negative influence.

The risk lies not in acknowledging this bidirectional influence, but in leaving it unmodeled. Ignoring these dynamics implies accepting an inherent flaw where the agent may manipulate the human to maximize reward without safeguards. Conversely, by explicitly modeling the mutual dynamic, we can analyze, understand, and design safeguards necessary to mitigate the potential for misuse. Coadaptive Value Alignment has the potential to create a continuous, adaptive loop where the agent and the human co-create a shared understanding, moving beyond mere instruction-following to build a genuine collaborative relationship.

2 COADAPTIVE VALUE ALIGNMENT

While RL theory dictates that a reward function should remain fixed to define a problem, in practice, researchers typically engage in a meta-level loop of refinement [17]. They specify a learning configuration containing an agent guided by a proxy reward function. After optimization yields a policy, the researcher evaluates it, not against the specified reward function, but against their own actual desired behavior, which remains latent in their mind. Recognizing that the agent maximized the specified reward but failed to achieve the intended goal, the researcher refines the reward specification or shaping components. The flow of information in this process is fundamentally unidirectional, always flowing from the researcher into the learning setup. The Coadaptive Value Alignment framework proposes to close this meta-level loop by bringing the human's subjective judgment directly into the agent's learning process. Our framework empowers the agent to estimate the human's latent internal state, allowing it to adapt not only to the environment but also to the human's evolving perceptions and expectations.

2.1 Conceptual Framework

We conceptualize the alignment problem as a dynamic interaction between an agent and a human, where the ultimate goal is to maximize the human's *latent satisfaction*. In our framework, "adaptation" refers specifically to the update of the human's latent internal state, p_t. To close the loop, the agent adapts to the human, and the human's internal state updates based on this interaction. Human behavioral changes naturally follow these shifts in p_t, rather than occurring as immediate, unprompted adjustments.

Unlike standard RL where the reward is static, human satisfaction is fungible, history-dependent, and sensitive to context. Psychological frameworks, such as Prospect Theory [16], suggest that human satisfaction is determined not by absolute outcomes, but by gains and losses relative to a reference point (e.g., expectations). Furthermore, factors such as loss aversion [30] and the history of the interaction [19] play significant roles; a trajectory of improving performance is often perceived more favorably than a trajectory of declining performance, even if the net task output is identical.

Because satisfaction depends on the complex interplay of history, context, and expectation, a static reward function cannot fully capture the human's desired outcome. Therefore, we propose that the agent must pursue a strategy that maximizes the human's latent satisfaction by performing the task effectively while also positively shaping the human's internal state. This shaping goes beyond simply "aligning" expectations to reality; it involves optimizing for the psychological drivers of satisfaction. For example, in some contexts, aligning expectations is critical to prevent disappointment (loss aversion). In others, the agent might aim to exceed low expectations to generate positive surprise. Thus, a successful agent must learn a policy that is not only competent at the task but is also capable of reasoning about the unobservable mental state of the human to maximize long-term satisfaction.

2.2 Formulation

Formally, we frame the mutual adaptation problem as a POMDP from the agent's perspective, where the environment includes both observable states that describe the physical environment and a human whose mental state is a critical component, but unobservable. This formulation naturally connects to Multi-Agent Systems under the view that a human acts as a co-agent with hidden variables, compelling an agent such as a robot to maintain a belief over these variables while optimizing a shared objective.

A POMDP is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{Z}, \mathcal{R}, \gamma \rangle$, where: $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{O}$ is the observation space, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function $\mathcal{T}(s'|s, a) = \mathbb{P}(s_{t+1} = s'|s_t = s, a_t = a)$, $\mathcal{Z} : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to [0, 1]$ is the observation function $\mathcal{Z}(o|s', a) = \mathbb{P}(o_{t+1} = o|s_{t+1} = s', a_t = a)$, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a d-component reward function, $\gamma \in [0, 1)$ is the discount factor.

Under our framework, the state $s_t \in \mathcal{S}$ is composite: $s_t = (e_t, p_t)$, where e_t is the observable *task state*, which describes the physical environment (e.g., positions in a shared workspace, task progress), and p_t is the unobservable *human internal state*, which represents relevant psychological elements such as trust in the agent, expectation of competence, level of surprise, comfort, or overall satisfaction.

Building on psychological models [16, 21, 24], we choose satisfaction as a proxy to encapsulate the complex relationships between many elements of human internal state.

Because the internal state of the person is latent, the agent does not have direct access to p_t. It receives an observation $o_t \in O$ that includes the full task state e_t, plus imperfect cues regarding p_t. These cues could be verbal feedback, facial expression, posture, or even measures of passivity in a shared task. The observation function $\mathcal{Z}(o|s', a)$ models how cues provide partial information about the human's latent perception of robot behavior.

The transition function $\mathcal{T}(s'|s, a)$ captures the dual impact of the agent's actions. Importantly, the action space $\mathcal{A}$ includes inaction (e.g., waiting), capturing how passivity influences p_t (e.g., eroding trust over time). It decomposes into environmental dynamics $\mathbb{P}(e_{t+1}|e_t, a_t)$ which governs the physical task and human mental dynamics $\mathbb{P}(p_{t+1}|p_t, a_t, e_t, o_t)$, which models how the human updates their mental state based on the agent's actions and outcomes. This is the first nexus of bidirectional influence: the agent's action a_t directly influences the subsequent human mental state p_{t+1}. For example, a successful and communicative action might increase a human's trust, while an unexpected or clumsy action might decrease it.

The reward function $\mathcal{R}$ is a fixed mapping known to the agent. However, because the human's true objective is latent and cannot be fully written down [2, 10], we approximate the components required to achieve human satisfaction. We formulate $\mathcal{R}$ as a vector of reward components, recognizing that this is a simplification of the complex human utility function:

$$\mathcal{R}(e, p, a) = \langle r_{task}(e, a), r_{sat}(e, p, a) \rangle \quad (1)$$

Objective progress on the physical task is incentivized by $r_{task}(e, a)$, while human satisfaction with the interaction is incentivized by $r_{sat}(e, p, a)$. Depending on the specific domain, r_{sat} may reward states corresponding to high trust, accurate expectations, or positive surprise. Factoring the reward in this manner is a practical simplification that allows the agent to weigh the trade-offs between immediate task efficiency and the long-term maintenance of the collaborative relationship. Finding an acceptable optimal policy under these multiple objectives requires balancing actions that are purely task-optimal with those that effectively shape the human's perception in the vector-valued total return $\mathbf{V}^{\pi}$.

Since the human internal state p_t is hidden, the agent maintains a belief state $b_t : \mathcal{S} \to [0, 1]$, where $b_t(s)$ is the probability distribution over possible states given history. Belief updates follow the standard Bayes' filter with normalization factor η: $b_{t+1}(s') = \eta\, \mathcal{Z}(o_{t+1}|s', a_t) \sum_s \mathcal{T}(s'|s, a_t) b_t(s)$.

The agent's policy $\pi : \mathcal{B} \to \mathcal{A}$ maps its current belief b_t to the next action a_t. In general, a policy π which optimizes for one component of $\mathcal{R}$ may not be optimal for the other component of $\mathcal{R}$, so an optimal π must be chosen from a set of possibly optimal policies, effectively striking a balance between immediate task efficiency and the long-term maintenance of the collaborative relationship under the vector-valued objective: $\mathbf{V}^{\pi}(b) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \mid b_0 = b, \pi\right]$. Our formulation explicitly models the human internal state p_t, enabling the agent to shape interaction dynamics rather than just responding to task-related feedback. Encapsulating p_t in the belief b_t allows the robot to condition its policy for the next action $\pi(a_{t+1}|b_t)$ on both the human's latent state p_t and task state e_t. Our formulation acknowledges the challenge of reward function tuning in standard RL while highlighting the need to treat the human as a dynamic partner with evolving internal state.

3 REAL-WORLD APPLICATION

The POMDP formulation provides a theoretical blueprint for Coadaptive Value Alignment, but its instantiation depends on three critical building blocks. First, one must identify and clearly define the key dimensions of human internal state with which the agent is concerned. Second, one needs a robust method for estimating the identified human state in real time from observable signals. Finally, the agent must be deployed in a scenario that can leverage the estimate of human internal state to learn policies that balance task success with perception management. Importantly, measuring task success and interaction dynamics over long-term deployments will reveal practical strategies, such as determining when to sacrifice short-term efficiency to repair an interaction, or how to calibrate user expectations over repeated interactions. The recent and proposed research discussed in this section provides an actionable pathway for instantiating each component, which will enable the framework to move from theory into practice.

3.1 Elements of Human State

Human internal state (p_t) is a critical component of our framework. We draw upon established psychological frameworks, including dimension models like the Circumplex Model of Affect [21] and appraisal theories [24]. Researchers typically quantify human mental states using techniques such as behavioral analysis (e.g., facial expression [8], body language [7]), and physiological sensing (e.g., heart rate [20], Electroencephalogram [18]). Ground-truth data for these states is typically elicited via instruments such as the NASA-TLX survey [11], the Robot Social Attribute Scale (RoSAS) [5], and the Perceived Social Intelligence (PSI) [4]. However, surveys break the real-time flow of information during the interaction and serve as a post-hoc source of ground-truth data, motivating the need for continuous, non-intrusive estimation methods.

3.2 Real-Time Estimation

Applying the framework in the real world requires operationalizing the observation function $\mathcal{Z}$ to map observable cues o_t to latent human state p_t. To this end, recent work demonstrates two avenues for creating such real-time perception modules.

First, Zhang et al. [33] demonstrate that it is possible to train machine learning models to predict people's internal states based on survey feedback that maps external observations to implicit, non-verbal cues. Datasets incorporating *implicit* feedback, including gaze, posture, and the spatial relationships between human and robot, can serve as an effective proxy for internal states such as perceived competence [32]. While Zhang et al. [33] and other research in this area [6] have primarily been conducted in simulation, real world demonstrations provide a pathway for learning perception models using only on-robot sensing [9, 29].

Alternatively, high-fidelity physiological signals (e.g., EEG, PPG) could provide a data-rich stream mapping to cognitive states. Future work could correlate these signals with on-robot sensors to

create a deployable perception model that operates without requiring surveys or on-body sensing. Crucially, while on-body sensing and non-verbal cues are potential pathways to estimating p_t, our control-theoretic framework remains entirely agnostic to the specific sensing method.

3.3 Robot Actions to Human Perception

Our framework hinges on the premise that an agent's actions, a_t, are capable of directly influencing the human's future mental state (p_{t+1}). We conducted a preliminary case study in which a group of students took turns teleoperating and observing a Boston Dynamics Spot robot navigating environments of varying complexity with both static and dynamic obstacles. The teleoperator was instructed to navigate the robot to a pre-specified goal position. The teleoperator was given instructions to perform the navigation according to one of two behaviors: in a "competent" or "incompetent" manner, while maintaining safety. No further definition was provided regarding the two subjective behaviors, and the observers were not told the current navigation behavior.

We analyzed the relationship between the teleoperator's instructed level of competence and people's perceived level of competence. Spearman's correlations revealed significant positive relationships between ground truth competence and perceived competence for both observers blind to the instruction ($r_s(249) = .6917, p < .0001$) and the teleoperators themselves ($r_s(124) = .7786, p < .0001$). Our results provide preliminary support for the modeled human mental dynamics $\mathbb{P}(p_{t+1}|p_t, a_t, e_t, o_t)$: a robot, through its physical actions on the ground plane, can predictably influence a human's latent perception of its performance.

4 PERSPECTIVES ON VALUE ALIGNMENT

The concept of value alignment has a rich history, progressing from early cybernetic formulations [31] to modern learning-based approaches. However, the dominant paradigm remains fundamentally unidirectional: the human is cast as an external oracle who specifies objectives and evaluates success, while the agent is a passive learner. The agent learns *from* human data, but rarely learns to *adapt* to the dynamic human providing it.

Standard paradigms like reward learning [25], Learning from Demonstration (LfD) [3], and Preference Learning [13] are unidirectional, treating the human as a consistent oracle with a fixed utility function. Our framework challenges this perspective by positing that the human is an active, in-the-loop participant whose mental state, and consequently their satisfaction, is a dynamic variable that the agent must explicitly model and influence.

Value alignment has also been investigated through the lenses of norms, trust, and Computational Theory of Mind (ToM). Such models structure mental states to predict human behavior governed by social constructs and shared values. Our framework is designed to complement these approaches. Whereas ToM and norm-based architectures provide the structural representations of human mental states, our control-theoretic perspective supplies the closed-loop mechanism by which an agent can optimize its actions over those evolving states.

4.1 Contrasting Frameworks

Our approach shares methodological roots with formalisms like POMDPs, yet its core objective is distinct. Frameworks such as Mutual Adaptation [1] and Shared Autonomy [12, 15] acknowledge that interaction is a two-way street, but primarily focus on bidirectionality at a *behavioral* level, adapting physical motions or delegating sub-tasks. Similarly, Ad Hoc Teamwork [27] emphasizes reasoning about a partner's hidden state to achieve coordination without pre-planning, sometimes involving actions to guide a teammate's behavior [28].

Prior framings typically focus on *instrumental* coordination by achieving informational alignment over task-related states, or value alignment over fixed objectives. In contrast, our framework targets *epistemic* alignment, where the agent explicitly aligns its belief state with the human's evolving internal state. We propose that an agent's actions inherently influence the user's internal state, which includes elements such as trust, expectations, and comfort, regardless of the agent's intent. Unlike prior work, we explicitly model this unavoidable influence, treating the human's internal state not just as a variable to be estimated, but as a state to be actively managed in order to foster stable, long-term partnerships.

4.2 Ethical Considerations & Inseparability

A critical implication of our framework is the *problem of influential inseparability*, where an agent capable of positively shaping human states inherently possesses the capability to manipulate them for maladaptive ends [23]. For example, large language models have violated ethical standards for mental health practice [14]. While unidirectional approaches leave this risk unmodeled and susceptible to "reward gaming" via cognitive bias exploitation [26], our framework moves these dynamics into the system's explicit state space. By formalizing these dynamics, we enable the future implementation of concrete safeguards, such as constraints on the rate of induced trust change or regularizers for autonomy preservation, which are impossible to enforce when influence remains implicit.

5 CONCLUSION

Traditional value alignment relies on the assumption that a static reward function can adequately encode human internal states. However, this view ignores the complex psychology of the human partner: satisfaction is not a fixed target, but a dynamic perception influenced by history, context, and the presentation of behavior.

While an agent cannot directly control a human's internal perceptions, it does have control over the *history of interaction*. Our proposed framework leverages this control, allowing the agent to shape the interaction history not just to complete tasks, but to foster a positive latent state in the user. Ultimately, transitioning from a static reward to optimizing for a person's overall satisfaction offers a critical pathway for AI, moving beyond obedient tools that blindly optimize metrics towards socially aware teammates capable of building long-term, satisfying partnerships.

Interesting directions for future work include the integration of formal safeguards to prevent manipulative policies, the exploration of communicative actions that transparently calibrate expectations, and the development of robust value drift models to track long-term changes in human preferences.

ACKNOWLEDGMENTS

This work has taken place in the Autonomous Mobile Robotics Laboratory (AMRL), the Learning Agents Research Group (LARG), and the Trishul Lab in the University of Texas at Austin Computer Science Artificial Intelligence Lab. AMRL research is supported in part by the NSF (CAREER-2046955, IIS-2416461, CCF-2505865, DGE-2125858). LARG research is supported in part by the National Science Foundation (FAIN-2019844, NRT-2125858), the Office of Naval Research (N00014-24-1-2550), Army Research Office (W911NF-17-2-0181, W911NF-23-2-0004, W911NF-25-1-0065), Lockheed Martin, DARPA (Cooperative Agreement HR00112520004 on Ad Hoc Teamwork), and Good Systems, a research grand challenge at the University of Texas at Austin. Trishul research is supported in part by DARPA (HR00112320018, HR0011-24-9-0431) and the Army Research Office (W911NF2110009). The views and conclusions contained in this document are those of the authors alone. Peter Stone serves as the Chief Scientist of Sony AI and receives financial compensation for that role. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

REFERENCES

[1] Sofie Ahlberg, Agnes Axelsson, Pian Yu, Wenceslao Shaw Cortez, Yuan Gao, Ali Ghadirzadeh, Ginevra Castellano, Danica Kragic, Gabriel Skantze, and Dimos V Dimarogonas. 2022. Co-adaptive human–robot cooperation: summary and challenges. *Unmanned Systems* 10, 02 (2022), 187–203.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[3] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems* 57, 5 (2009), 469–483.

[4] Kimberly A Barchard, Leiszle Lapping-Carr, R Shane Westfall, Andrea Fink-Armold, Santosh Balajee Banisetty, and David Feil-Seifer. 2020. Measuring the perceived social intelligence of robots. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 4 (2020), 1–29.

[5] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. 2017. The robotic social attributes scale (RoSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. 254–262.

[6] Yuchen Cui, Qiping Zhang, Brad Knox, Alessandro Allievi, Peter Stone, and Scott Niekum. 2021. The empathic framework for task learning from implicit human feedback. In *Conference on Robot Learning*. PMLR, 604–626.

[7] Beatrice De Gelder. 2006. Towards the neurobiology of emotional body language. *Nature reviews neuroscience* 7, 3 (2006), 242–249.

[8] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[9] Haley N Green and Tariq Iqbal. 2025. Using Physiological Measures, Gaze, and Facial Expressions to Model Human Trust in a Robot Partner. *arXiv preprint arXiv:2504.05291* (2025).

[10] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. 2017. Inverse reward design. *Advances in neural information processing systems* 30 (2017).

[11] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[12] Joshua Hoegerman, Shahabedin Sagheb, Benjamin A Christie, and Dylan P Losey. 2024. Aligning learning with communication in shared autonomy. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 11530–11536.

[13] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems* 31 (2018).

[14] Zainab Iftikhar, Amy Xiao, Sean Ransom, Jeff Huang, and Harini Suresh. 2025. How LLM Counselors Violate Ethical Standards in Mental Health Practice: A Practitioner-Informed Framework. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 8. 1311–1323.

[15] Siddarth Jain and Brenna Argall. 2019. Probabilistic human intent recognition for shared autonomy in assistive robotics. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 1 (2019), 1–23.

[16] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 2 (1979), 263–291.

[17] W Bradley Knox and James MacGlashan. 2024. How to specify reinforcement learning objectives. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*.

[18] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.

[19] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[20] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence* 23, 10 (2001), 1175–1191.

[21] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[22] Stuart Russell. 2022. Human-Compatible Artificial Intelligence. *Human-like machine intelligence* 1 (2022), 3–22.

[23] Sahand Sabour, June M Liu, Siyang Liu, Chris Z Yao, Shiyao Cui, Xuanming Zhang, Wen Zhang, Yaru Cao, Advait Bhat, Jian Guan, et al. 2025. Human decision-making is susceptible to ai-driven manipulation. *arXiv preprint arXiv:2502.07663* (2025).

[24] Klaus R Scherer et al. 2000. Psychological models of emotion. *The neuropsychology of emotion* 137, 3 (2000), 137–162.

[25] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. 2021. Reward is enough. *Artificial intelligence* 299 (2021), 103535.

[26] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems* 35 (2022), 9460–9471.

[27] Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. 2010. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 24. 1504–1509.

[28] Peter Stone, Gal A Kaminka, Sarit Kraus, Jeffrey S Rosenschein, and Noa Agmon. 2013. Teaching and leading an ad hoc teammate: Collaboration without pre-coordination. *Artificial Intelligence* 203 (2013), 35–65.

[29] Sydney Thompson, Alexander Lew, Yifan Li, Elizabeth Stanish, Alex Huang, Rohan Phanse, and Marynel Vázquez. 2024. Predicting Human Intent to Interact with a Public Robot: The People Approaching Robots Database (PAR-D). In *Proceedings of the 26th International Conference on Multimodal Interaction*. 536–545.

[30] Amos Tversky and Daniel Kahneman. 1991. Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics* 106, 4 (1991), 1039–1061.

[31] Norbert Wiener. 1960. Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science* 131, 3410 (1960), 1355–1358.

[32] Qiping Zhang, Austin Narcomey, Kate Candon, and Marynel Vázquez. 2023. Self-Annotation Methods for Aligning Implicit and Explicit Human Feedback in Human-Robot Interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 398–407.

[33] Qiping Zhang, Nathan Tsoi, Mofeed Nagib, Booyeon Choi, Jie Tan, Hao-Tien Lewis Chiang, and Marynel Vázquez. 2025. Predicting Human Perceptions of Robot Performance during Navigation Tasks. *ACM Transactions on Human-Robot Interaction* 14, 3 (2025), 1–27.