Transparent Value Alignment

Lindsay Sanneman
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
lindsays@csail.mit.edu

Julie Shah
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
julie_a_shah@csail.mit.edu

ABSTRACT

As robots become increasingly prevalent in our communities, aligning the values motivating their behavior with human values is critical. However, it is often difficult or impossible for humans, both expert and non-expert, to enumerate values comprehensively, accurately, and in forms that are readily usable for robot planning. Misspecification can lead to undesired, inefficient, or even dangerous behavior. In the value alignment problem, humans and robots work together to optimize human objectives, which are often represented as reward functions and which the robot can infer by observing human actions. In existing alignment approaches, no explicit feedback about this inference process is provided to the human. In this paper, we introduce an exploratory framework to address this problem, which we call Transparent Value Alignment (TVA). TVA suggests that techniques from explainable AI (XAI) be explicitly applied to provide humans with information about the robot's beliefs throughout learning, enabling efficient and effective human feedback.

CCS CONCEPTS

• **Computing methodologies → Theory of mind**; *Reinforcement learning*; *Cognitive robotics*; • **Human-centered computing →** *HCI theory, concepts and models*; *Collaborative interaction.*

KEYWORDS

Value Alignment, Transparency, Explainable AI

ACM Reference Format:
Lindsay Sanneman and Julie Shah. 2023. Transparent Value Alignment. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3568294.3580147

1 INTRODUCTION

Alignment of autonomous agent objectives with those of humans could greatly enhance agents' ability to act flexibly to safely and reliably meet humans' goals across a variety of contexts. However, a key barrier to alignment is that it is often difficult for humans to specify their objectives comprehensively in ways that produce desired agent behavior across all contexts and in forms that are readily usable for agent planning [32]. Value alignment is an open challenge in artificial intelligence that aims to address this problem by enabling agents to infer human goals and values through interaction [10, 16, 20]. Though not explicitly accounted for in most existing approaches to alignment, direct and appropriate feedback about this value learning process could enable humans to more readily verify alignment or identify gaps in agent models and subsequently amend these gaps efficiently and effectively. In this paper, we introduce an exploratory framework which captures this two-way communication and inference process which we call *Transparent Value Alignment (TVA)*. An overview of the closed-loop TVA process is depicted in Figure 1.

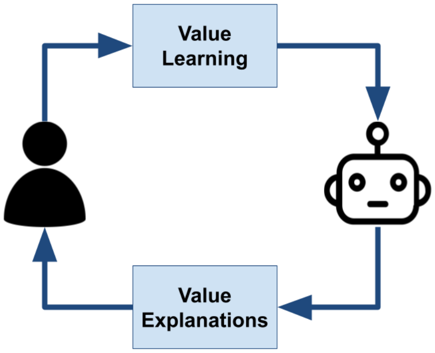

Figure 1: Transparent Value Alignment (TVA) Framework

To see the value of applying the TVA lens to approaches for alignment, we can consider a manufacturing-based scenario in which a human and a robot work together to efficiently move boxes from a conveyor belt onto a shipping pallet. Suppose the robot initially learns to perform the task while working with a cautious human partner who never enters the robot's reachable workspace. However, suppose that this robot also needs to be able to accommodate future human partners who are less cautious and prefer to work in closer proximity to the robot when moving boxes. In this scenario, the robot might learn to prioritize efficiency over maintaining a safe distance from human partners or might not learn to consider proximity to humans at all, which could lead to dangerous interactions. Providing explicit feedback about the trade-offs the robot has learned to the humans in this scenario could enable them to amend the robot's priorities, thereby enhancing the safety of future interactions and the overall performance of the human-robot team. The TVA problem framing applies not only to manufacturing applications such as this example, but also has broad applicability across many other human-robot interaction domains such as search and rescue robotics, space robotics, autonomous driving, and others.

While human objectives and values can be represented in a variety of ways, reward functions are a common representation in the value alignment setting [10, 16, 20]. Reward functions encode

HRI '23 Companion, March 13–16,2023, Stockholm, Sweden
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9970-8/23/03.
https://doi.org/10.1145/3568294.3580147

the benefit of taking different actions from different states and can be used to develop autonomous agent plans in fields such as reinforcement learning [38]. They are widely applied to agent planning problems because they provide an efficient way to model desired and undesired behaviors, can be leveraged within a number of convenient modeling frameworks such as Markov Decision Processes (MDPs), and are generalizable across a variety of environments. Additionally, there exist numerous approaches for both learning reward functions from humans and optimizing them. However, misspecified reward functions can lead to problems such as "reward hacking" where an agent optimizes reward in a way that is not intended by the designer [2, 32], which motivates the need for alignment in reward-driven planning scenarios in particular. Given these considerations, we focus on learning and explaining reward functions in our discussion of TVA.

Some existing approaches to reward alignment make the assumption that humans can infer an agent's beliefs about the reward function over the course of an interaction by simply observing the agent's optimal actions at each step [16, 20]. However, this might be difficult or intractable given human cognitive limitations, especially in complex scenarios. Therefore, enabling agents to provide direct feedback to humans about their current beliefs about reward and doing so in a way that leverages the most effective techniques for agent transparency could enhance the alignment process. State-of-the-art techniques in explainable AI (XAI) can be explicitly designed and applied to enable agents to achieve this end. Gunning and Aha [19] define XAI as "AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future." We adopt this definition of XAI and define explanations in the TVA context as the information necessary to support human inference of the agent's reward function.

The key insight of the TVA approach to alignment is that studying each component of the TVA loop (value learning and value explanations) while considering the entire closed-loop process has the potential to enhance alignment overall. For example, certain approaches for value learning may be better suited to supporting interpretability than others, and certain approaches to AI explainability may enable humans to more naturally guide a robot's value learning process. Therefore, studying each component in light of the other is an important future direction of this work.

In the remainder of this paper, we discuss a representative but non-comprehensive set of existing approaches to the two primary components of the TVA process depicted in Figure 1: value learning (Section 2) and value explanations (Section 3). Since we consider reward functions as the representation of value in this work, we outline approaches for reward learning and reward explanations in particular. In these sections, we also identify future research directions for reward learning and reward explanations as they relate to the closed-loop, bi-directional TVA process. In Section 4, we highlight additional future research directions motivated by the exploratory TVA framework. Section 5 further discusses how TVA can be applied as a tool in human-centered design processes. Finally, Section 6 concludes the paper.

2 APPROACHES FOR REWARD LEARNING

Various approaches for reward learning have been proposed in the AI literature. Many approaches apply inverse reinforcement learning-based techniques, which leverage demonstrations of optimal or near-optimal behavior to infer objectives and rewards [31]. For example, learning from demonstration is one common approach in which a human provides demonstrations of desired behavior and the agent derives the reward and associated optimal policy from these demonstrations [4]. While most approaches assume that human demonstrations are either optimal or approximately optimal (according to a Boltzmann rational assumption) when performing inference over human rewards, recent work has considered a more structured approach to modeling human sub-optimality which accounts for the possibility of missing features [6].

Beyond learning from demonstration-based approaches, active learning involves a "learner" agent which learns about a human teacher's reward function through a sequence of queries about agent behavior [36]. These queries often ask humans to provide their preference for agent behavior from two or more demonstrations. A popular strategy for query selection is uncertainty sampling in which the agent generates queries that maximally reduce uncertainty about the reward parameters [36]. Another recent strategy selects queries based on both uncertainty over reward parameters and the human's ability to answer the queries confidently [5]. Bıyık et al. [5] show that the agent's estimate of the human's reward function is better aligned with the human's true reward function and that human teachers find queries easier to answer with their proposed query selection strategy than with a baseline approach. Other approaches involve incorporating corrections or critiques of agent behavior into active learning in order to reduce the total number of queries required to learn human reward functions [12]. While these strategies ask about the ideal agent behavior that results from the human's underlying reward function, other approaches ask humans about the features in their reward function directly [21, 29]. For example, Hadfield-Menell et al. [21] leveraged human-crafted reward functions as observations about the true reward and performed inference over the true reward function given these observations.

Existing approaches for learning from demonstration and active learning do not provide explicit feedback to the human about what the agent has learned throughout the interaction. Recent work in value alignment has accounted for the implicit feedback humans receive during reward learning through their observations of agent actions. This work models human-agent interaction as a two-player game in which only the human has full information about the reward function that the team is aiming to optimize [16, 20]. In this setup, the human infers the agent's current beliefs through observing its actions at each point during the interaction and then uses this inferred information to choose their own maximally informative actions for the agent to observe in turn. As discussed in the Introduction, while feedback is implicitly accounted for in this problem setup, these techniques have not explored how to effectively provide feedback through the explicit use of explanation techniques within the value alignment context, which the TVA framework addresses.

3 APPROACHES FOR REWARD EXPLANATION

As with reward learning, numerous approaches for explaining reward have also been proposed. These fall broadly into two categories: feature-based explanations, which explain reward functions in terms of their features and relative weights, and policy-based explanations, which explain reward functions in terms of the policies that result from the reward function [34]. Feature-based techniques introduced to date include directly providing the reward function in terms of its features and weights [34], providing subsets of the most important, prototypical, or unknown features [25, 33, 39], and providing weighted abstractions of reward features [26, 37]. Policy-based techniques include providing the most optimal, least optimal, or most legible trajectories based on the reward function [14, 34], summarizing the policy by either providing sets of informative demonstrations [1, 9, 22, 27, 28] or rationales that describe how state-action pairs relate to the overall policy [13, 15], and providing factored policy information which indicates the contribution of each state-action pair to each reward feature [3, 23]. While not yet studied, combinations of feature- and policy-based explanation techniques may be most ideal in certain contexts, and further research is needed to determine whether these combinations would be of value in the alignment context.

Applying these explanation techniques successfully in the alignment setting requires not only an understanding of how effectively each technique communicates reward information but also consideration of additional factors which affect human information processing and reliance on automation, such as trust and cognitive workload [35]. Sanneman and Shah [34] performed a recent study comparing a variety of reward explanation techniques and found that while directly providing reward information most effectively supported human understanding of reward functions, providing abstractions of reward features was most effective in terms of balancing reward understanding with cognitive workload. Another consideration for presenting reward explanations is how to appropriately scaffold and communicate information in ways that account for human learning strategies, which both Lee et al. [28] and Booth et al. [8] address. Further study of these techniques and others within the closed-loop TVA context in particular is needed to determine which individual or combinations of explanation techniques are most effective for enhancing human feedback in the value learning process.

4 CLOSING THE TVA LOOP: FUTURE RESEARCH DIRECTIONS

Future TVA research will require consideration and study of the full closed-loop process of learning and explaining reward as a whole. Beyond studying techniques such as the ones outlined in Sections 2 and 3 in light of this closed-loop TVA process, there are a variety of additional avenues for future research which would also be of value. For example, developing a common language [24] or a set of common representations [7] for reward features between humans and robots could make both the processes of reward learning and explanation more efficient. While initial steps have been taken in this direction [7, 24], more research is necessary to extend these initial approaches and characterize their efficacy in the closed-loop alignment context more broadly.

In addition, learning and explaining reward in the TVA context may be most effectively achieved by leveraging different types of information (e.g. feature- or policy-based information or combinations of them) at different points throughout the learning process, and further study of this would be of value. Individual differences between humans, such as expertise, skills, and communication needs, should also be studied and accounted for in future research on TVA [11]. Finally, while human studies have been performed to determine the efficacy of algorithms for reward learning and explanation in isolation, future human studies will be necessary to characterize any differences in and nuances of these results within the context of the closed-loop TVA process.

While this paper has primarily discussed scenarios where a human holds the ground truth reward function, an additional future research direction relates to cases where the human must instead learn about the reward function from an agent which holds the ground truth, flipping the proposed loop. In this case, algorithms that explicitly support human reward learning and evaluation of the human's knowledge throughout the learning process must be developed and studied. Previous research on AI tutoring could serve as a launching point for such approaches [18]. Finally, complex human-robot teaming scenarios of the future will most likely involve cases where neither the human nor the robot has full information about the team's shared reward function or cases where multiple humans have different rewards or preferences that must be negotiated in order to converge on an agreed-upon shared reward. Algorithms for learning, explanation, and negotiation in these circumstances must also be developed.

5 TVA AS A TOOL FOR HUMAN-CENTERED DESIGN

In this paper, we have discussed reward functions as a possible representation for human values, since reward lends itself well to the context of robot decision-making. It is important to note that we refer to values in a limited sense here: the values encapsulated by reward functions model only the subset of human values which translate cleanly to trade-offs between observable features of a robot's environment that influence its decision-making. However, robots are also part of a broader societal context which includes the full set of human values. These values influence much more about robot adoption than robot decision-making, including both whose values the robot aligns with and how robots are adopted in any given context, if at all.

Processes such as Participatory Design [17] and Value Sensitive Design (VSD) [30] can be applied to ensure that the perspectives of all relevant stakeholders are accounted for in the application of robotics across new domains. TVA can be applied as a tool within these processes to ensure that the values they capture are appropriately translated into robot reward functions. Since TVA increases transparency about the value learning process, it will also ideally empower non-robot programmers and those who might not otherwise have power over how robots are applied in their respective domains to more directly control the robot adoption process.

6 CONCLUSION

In this paper, we proposed the exploratory Transparent Value Alignment framework, which contributes to the design and study of effective algorithms for bi-directional communication between humans and robots in the value alignment setting. This framework incorporates the explicit consideration of suitable strategies for explaining the robot's reward function throughout the learning process. This can enable a person to provide enhanced feedback which leads to more efficient and higher-quality alignment. There are an abundance of future research directions which build on the proposed exploratory TVA framework, spanning both algorithm development and human study design. This future research will facilitate more effective human-robot teaming by lowering the barrier to safe and high-performance human-robot interaction.

REFERENCES

[1] Ofra Amir, Finale Doshi-Velez, and David Sarne. 2018. Agent strategy summarization. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems.* 1203–1207.

[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] Andrew Anderson et al. 2019. Explaining reinforcement learning to mere mortals: An empirical study. *arXiv preprint arXiv:1903.09708* (2019).

[4] 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.

[5] Erdem Bıyık, Malayandi Palan, Nicholas C Landolfi, Dylan P Losey, and Dorsa Sadigh. 2019. Asking easy questions: A user-friendly approach to active reward learning. *arXiv preprint arXiv:1910.04365* (2019).

[6] Andreea Bobu, Andrea Bajcsy, Jaime F Fisac, and Anca D Dragan. 2018. Learning under misspecified objective spaces. In *Conference on Robot Learning.* PMLR, 796–805.

[7] Andreea Bobu, Marius Wiggert, Claire Tomlin, and Anca D Dragan. 2021. Feature Expansive Reward Learning: Rethinking Human Input. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction.* 216–224.

[8] Serena Booth, Sanjana Sharma, Sarah Chung, Julie Shah, and Elena L Glassman. 2022. Revisiting human-robot teaching and learning through the lens of human concept learning. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 147–156.

[9] Daniel S Brown and Scott Niekum. 2019. Machine teaching for inverse reinforcement learning: Algorithms and applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7749–7758.

[10] Daniel S Brown, Jordan Schneider, Anca Dragan, and Scott Niekum. 2021. Value alignment verification. In *International Conference on Machine Learning.* PMLR, 1105–1115.

[11] Jessie YC Chen, Stephanie A Quinn, Julia L Wright, and Michael J Barnes. 2013. Effects of individual differences on human-agent teaming for multi-robot control. In *International Conference on Engineering Psychology and Cognitive Ergonomics.* Springer, 273–280.

[12] Yuchen Cui and Scott Niekum. 2018. Active reward learning from critiques. In *2018 IEEE international conference on robotics and automation (ICRA).* IEEE, 6907–6914.

[13] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. Explainable AI for robot failures: Generating explanations that improve user assistance in fault recovery. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction.* 351–360.

[14] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 301–308.

[15] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O Riedl. 2018. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* 81–87.

[16] Jaime F Fisac, Monica A Gates, Jessica B Hamrick, Chang Liu, Dylan Hadfield-Menell, Malayandi Palaniappan, Dhruv Malik, S Shankar Sastry, Thomas L Griffiths, and Anca D Dragan. 2020. Pragmatic-pedagogic value alignment. In *Robotics Research.* Springer, 49–57.

[17] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[18] Kevin A Gluck and John E Laird. 2018. Interactive task learning: Humans, robots, and agents acquiring new tasks through natural interactions. 169–191.

[19] David Gunning and David W Aha. 2019. DARPA's Explainable Artificial Intelligence Program. *AI Magazine* 40, 2 (2019), 44–58.

[20] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. Cooperative inverse reinforcement learning. *arXiv preprint arXiv:1606.03137* (2016).

[21] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. 2017. Inverse reward design. *arXiv preprint arXiv:1711.02827* (2017).

[22] Sandy H Huang, David Held, Pieter Abbeel, and Anca D Dragan. 2019. Enabling robots to communicate their objectives. *Autonomous Robots* 43, 2 (2019), 309–326.

[23] Zoe Juozapaitis et al. 2019. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on Explainable Artificial Intelligence.*

[24] Subbarao Kambhampati, Sarath Sreedharan, Mudit Verma, Yantian Zha, and Lin Guan. 2022. Symbols as a lingua franca for bridging human-ai chasm for explainable and advisable ai systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12262–12267.

[25] Been Kim, Cynthia Rudin, and Julie A Shah. 2014. The bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in neural information processing systems* 27 (2014).

[26] Isaac Lage and Finale Doshi-Velez. 2020. Human-in-the-Loop Learning of Interpretable and Intuitive Representations. In *ICML Workshop on Human Interpretability in Machine Learning, Vienna, Austria.*

[27] Isaac Lage, Daphna Lifschitz, Finale Doshi-Velez, and Ofra Amir. 2019. Exploring computational user models for agent policy summarization. *arXiv preprint arXiv:1905.13271* (2019).

[28] Michael S Lee, Henny Admoni, and Reid Simmons. 2022. Reasoning about Counterfactuals to Improve Human Inverse Reinforcement Learning. *arXiv preprint arXiv:2203.01855* (2022).

[29] Sören Mindermann, Rohin Shah, Adam Gleave, and Dylan Hadfield-Menell. 2018. Active inverse reward design. *arXiv preprint arXiv:1809.03060* (2018).

[30] Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Commun. ACM* 36, 6 (1993), 24–28.

[31] Andrew Y Ng, Stuart J Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *Icml*, Vol. 1. 2.

[32] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544* (2022).

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* 1135–1144.

[34] Lindsay Sanneman and Julie A Shah. 2022. An empirical study of reward explanations with human-robot interaction applications. *IEEE Robotics and Automation Letters* 7, 4 (2022), 8956–8963.

[35] Lindsay Sanneman and Julie A Shah. 2022. The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems. *International Journal of Human–Computer Interaction* 38, 18-20 (2022), 1772–1788.

[36] Burr Settles. 2009. Active learning literature survey. (2009).

[37] Sarath Sreedharan et al. 2020. Bridging the Gap: Providing Post-Hoc Symbolic Explanations for Sequential Decision-Making Problems with Black Box Simulators. *arXiv preprint arXiv:2002.01080* (2020).

[38] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.

[39] Aaquib Tabrez, Shivendra Agrawal, and Bradley Hayes. 2019. Explanation-based reward coaching to improve human performance via reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 249–257.