Title: Value learning

URL: https://en.wikipedia.org/wiki/Value_learning

PageID: 80279928

Categories: Category:Artificial intelligence, Category:Cognitive science, Category:Machine learning, Category:Moral psychology

---

Value learning is a research area within artificial intelligence (AI) and AI alignment that focuses on building systems capable of inferring, acquiring, or learning human values, goals, and preferences from data, behavior, and feedback. The aim is to ensure that advanced AI systems act in ways that are beneficial and aligned with human well-being, even in the absence of explicitly programmed instructions. [1] [2]

Unlike traditional AI that focuses purely on task performance, value learning aims to ensure that AI decisions are ethically and socially acceptable. It is analogous to teaching a child right from wrong—guiding an AI to recognize which actions align with human moral standards and which do not. The process typically involves identifying relevant values (such as safety or fairness), collecting data that reflects those values, training models to learn appropriate responses, and iteratively refining their behavior through feedback and evaluation. Applications include minimizing harm in autonomous vehicles, promoting fairness in financial systems, prioritizing patient well-being in healthcare, and respecting user preferences in digital assistants. Compared to earlier techniques, value learning shifts the focus from mere functionality to understanding the underlying reasons behind choices, aligning machine behavior with human ethical expectations. [3]

Motivation

The motivation for value learning stems from the observation that humans are often inconsistent, unaware, or imprecise about their own values. Hand-coding a complete ethical framework into an AI is considered infeasible due to the complexity of human norms and the unpredictability of future scenarios. Value learning offers a dynamic alternative, allowing AI to infer and continually refine its understanding of human values from indirect sources such as behavior, approval signals, and comparisons. [4] [5]

A foundational critique of traditional reinforcement learning (RL) highlights its limitations in aligning artificial general intelligence (AGI) with human values. It is argued that RL systems optimize fixed reward signals, which can incentivize harmful or deceptive behavior if such actions increase rewards. As an alternative, he proposes value-learning agents that maintain uncertainty over utility functions and update beliefs based on interactions. These agents aim not to maximize static rewards but to infer what humans truly value. This probabilistic framework enables adaptive alignment with complex, initially unspecified goals and is viewed as a foundational step toward safer AGI. [6]

The growing importance of value learning is reflected in how AI products are increasingly evaluated and marketed. A notable shift occurred with the release of GPT-4 in March 2023, when OpenAI emphasized not just technical improvements but also enhanced alignment with human values. This marked one of the first instances where a commercial AI product was promoted based on ethical considerations. The trend signals a broader transformation in AI development—prioritizing principles like fairness, accountability, safety, and privacy alongside performance. As AI systems become more integrated into society, aligning them with human values is critical for public trust and responsible deployment. [7]

Key approaches

One central technique is inverse reinforcement learning (IRL), which aims to recover a reward function that explains observed behavior. IRL assumes that the observed agent acts (approximately) optimally and infers the underlying preferences from its choices. [8] [9]

Cooperative inverse reinforcement learning (CIRL) extends IRL to model the AI and human as cooperative agents with asymmetric information. In CIRL, the AI observes the human to learn their hidden reward function and chooses actions that support mutual success. [10] [11]

Another approach is preference learning , where humans compare pairs of AI-generated behaviors or outputs, and the AI learns which outcomes are preferred. This method underpins successful applications in training language models and robotics. [12] [13]

Recent research introduces a novel framework for learning human values directly from behavioral data, without relying on predefined models or external annotations. The method distinguishes between value specifications (contextual definitions) and value systems (agents' prioritizations among values). A demonstration in route choice modeling—using tailored inverse reinforcement learning (IRL) techniques—infers how agents weigh options such as speed, safety, or scenic routes. The results confirm that value learning from demonstrations can effectively capture complex decision-making preferences, supporting the feasibility of value-aligned AI in applied settings. [14]

Concept alignment

A major challenge in value learning is ensuring that AI systems interpret human behavior using similar conceptual models. Recent research distinguishes between "value alignment" and "concept alignment," the latter referring to the internal representations that humans and machines use to describe the world. Misalignment in conceptual models can lead to serious errors even if value inference mechanisms are accurate. [15]

Challenges

Value learning faces several difficulties:

Ambiguity of human behavior – Human actions are noisy, inconsistent, and context-dependent. [16]

Reward misspecification – The inferred reward may not fully capture human intent, particularly under imperfect assumptions. [17]

Scalability – Methods that work in narrow domains often struggle with generalization to more complex or ethical environments. [18]

Research from Purdue University reveals that AI training datasets disproportionately emphasize certain human values—such as utility and information-seeking—while underrepresenting others like empathy, civic responsibility, and human rights. By applying a value taxonomy grounded in moral philosophy, researchers found that AI systems trained on these datasets may struggle in morally complex or socially sensitive contexts. To address these gaps, the study employed reinforcement learning from human feedback (RLHF) and value annotation to audit and guide dataset improvements. This work underscores the importance of comprehensive value representation in data and contributes tools for more equitable, value-aligned AI development. [19]

Hybrid and cultural approaches

Recent work highlights the importance of integrating diverse moral perspectives into value learning. One framework, HAVA (Hybrid Approach to Value Alignment), incorporates explicit (e.g., legal) and implicit (e.g., social norm) values into a unified reward model. [20] Another line of research explores how inverse reinforcement learning can adapt to culturally specific behaviors, such as in the case of "culturally-attuned moral machines" trained on different societal norms. [21]

An important global policy initiative supporting the goals of value learning is UNESCO's Recommendation on the Ethics of Artificial Intelligence, unanimously adopted by 194 member states in 2021. Although the term "value learning" is not explicitly used, the document emphasizes the need for AI to operationalize values such as human dignity, justice, inclusiveness, sustainability, and human rights. It establishes a global ethical framework grounded in four core values and ten guiding principles, including fairness, transparency, and human oversight. Tools like the Readiness Assessment Methodology (RAM) and Ethical Impact Assessment (EIA) help translate these principles into practice. [22]

Applications

Value learning is being applied in:

Robotics – Teaching robots to cooperate with humans in household or industrial tasks. [23]

Large language models – Aligning chatbot behavior with user intent using preference feedback and reinforcement learning. [18]

Policy decision-making – Informing AI-assisted decisions in governance, healthcare, and safety-critical environments. [20]

See also

AI alignment

Inverse reinforcement learning

Preference learning

Ethics of artificial intelligence

Reward hacking

References