VALUEMAP: MAPPING CROWDSOURCED HUMAN VALUES TO COMPUTATIONAL SCORES FOR BI-DIRECTIONAL ALIGNMENT

Rupkatha Hira*
University of Pennsylvania
rupkatha@seas.upenn.edu

Priya DCosta*
Independent Researcher
pdcosta1999@gmail.com

ABSTRACT

Defining values for bi-directional alignment is challenging due to their dynamic nature. Traditional surveys are often biased, necessitating a shift to objective computational methods. We propose **ValueMap**, a framework mapping values from literature to computational proxies, enabling AI systems to adapt to evolving human values.

1 INTRODUCTION

Unidirectional approaches to AI alignment treat human values as static, failing to capture their evolving, dynamic nature. AI integration into daily life introduces challenges as norms and priorities evolve alongside technology Shen et al. (2024). Traditional methods like surveys are biased and inconsistent, insufficient for robust alignment. AI systems with shallow ethical assessments often fail to meet societal and legal demands Sanderson et al. (2024).

We propose **ValueMap**, a structured and adaptive framework for operationalizing human values through computational proxies. By synthesizing insights from interdisciplinary literature and leveraging crowdsourced inputs, ValueMap addresses two key dimensions of alignment:

- **Aligning AI with Humans:** ValueMap integrates human values into AI systems, enabling them to respond dynamically to evolving societal norms and individual priorities.
- **Aligning Humans with AI:** ValueMap empowers humans to critically evaluate, interpret, and adapt AI systems by grounding them in measurable and interpretable proxies for human values.

ValueMap aligns directly with the bidirectional human-AI alignment framework by fostering iterative, feedback-driven interactions between humans and AI. This builds on foundational principles such as Schwartz's Basic Human Values Schwartz (2012) and Haidt's Moral Foundations Theory Graham et al. (2013), which provide a basis for understanding societal and interpersonal values.

2 METHODOLOGY: BUILDING VALUEMAP

As an initial step toward a structured framework of human values, the authors review literature identifying observable behaviors linked to these values and propose computational methods to measure them. The framework is organized into three foundational dimensions:

(1) **Social Justice and Fairness:** Encompasses values ensuring equity, autonomy, and accountability in decision-making and resource distribution.

(2) **Communication and Interpersonal Values:** Includes empathy, honesty, and politeness, which are essential for building trust and fostering collaboration in human-AI interactions.

(3) **Cooperation and Social Behaviors:** Captures altruism, teamwork, and mutual support required for achieving shared goals in collaborative settings.

*Equal contribution.

Value	Behavioral Indicators	Proxy Computational Measures	Relevant Research
		Social Justice and Fairness	
Fairness	Equal distribution of resources, unbiased decision-making	Fairness metrics such as Gini coefficient, fairness constraints in ML models	Schwartz (2012)
Autonomy	Independent decision-making, resistance to coercion	Entropy in decision patterns to quantify autonomy	Deci & Ryan (2000)
Responsibility	Accountability, ownership of actions	Action recognition models to identify accountability in behavior	Bandura (1991)
Integrity	Adherence to ethical principles	Longitudinal pattern analysis for consistency in decision-making	Rest (1986)
		Communication and Interpersonal Values	
Empathy	Compassionate responses, emotional support	Sentiment analysis and language style matching	Droit-Volet & Gil (2009)
Honesty	Truth-telling, avoiding deception	Linguistic cues and deception detection models	Hancock (2008)
Gratitude	Expressing thankfulness	Keyword detection in conversational data	McCullough (2001)
Politeness	Respectful and considerate language	Frequency of polite phrases, acknowledgment of errors	Exline (2004)
		Cooperation and Social Behavior	
Altruism	Voluntary helpfulness	Recognition of prosocial actions in behavior data	Eisenberg & Lennon (1983)
Cooperation	Collaborative problem-solving	Social network analysis to measure cooperative actions	Johnson & Johnson (1995)

These dimensions form an initial structure, evolving through iterative refinement informed by crowdsourced feedback and interdisciplinary research. This approach resonates with Value-Sensitive Design Friedman (1996); Knobel & Bowker (2011), embedding human values into system design.

3 Conclusions and Future Work

ValueMap represents a step toward crowdsourcing human values and operationalizing them by mapping them to computationally measurable proxies. Future work will include:

- **Crowdsourced Value Refinement** – Implement a community-driven approach to continuously update ValueMap, integrating public discourse, expert opinions, and real-world case studies.
- **Expanding the ValueMap Database** – Incorporate additional values from cross-cultural ethics, political philosophy, and social psychology, ensuring comprehensive representation.
- **Refining Metrics and Developing Tools for AI Alignment** – Refine metrics via multimodal AI models and develop tools like Python packages for standardized AI alignment.
- **Ethical Trade-off Resolution Mechanisms** – See Appendix A for a discussion of value weighting and optimization strategies in the presence of conflicting values.
- **Crowdsourced Schema Development** – We are releasing the initial ValueMap schema and proxy design proposals as a living resource at `https://github.com/value-map/schema`. We invite researchers and practitioners to suggest new values, propose proxy metrics, and contribute validation datasets.

ValueMap aims to serve as a cornerstone in building AI systems that not only align with but also contribute to the continuous evolution of human values.

REFERENCES

Albert Bandura. Social cognitive theory of moral thought and action. *Handbook of Moral Behavior and Development*, 1991.

Edward L. Deci and Richard M. Ryan. The what and why of goal pursuits. *Psychological Inquiry*, 2000.

Sylvie Droit-Volet and Sandrine Gil. Empathy and emotion recognition in the human face. *Emotion Review*, 2009.

Nancy Eisenberg and Randy Lennon. Sex differences in empathy and related capacities. *Psychological Bulletin*, 1983.

Julie J. et al. Exline. Humility and social relationships. *Journal of Personality and Social Psychology*, 2004.

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

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. volume 47 of *Advances in Experimental Social Psychology*, pp. 55–130. Academic Press, 2013. doi: https://doi.org/10.1016/B978-0-12-407236-7.00002-4. URL https://www.sciencedirect.com/science/article/pii/B9780124072367000024.

Jeffrey et al. Hancock. Linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 2008.

David Johnson and Roger Johnson. Creative controversy: Intellectual challenge in the classroom. *Interaction Book Company*, 1995.

Cory Knobel and Geoffrey C. Bowker. Values in design. *Commun. ACM*, 54(7):26–28, July 2011. ISSN 0001-0782. doi: 10.1145/1965724.1965735. URL https://doi.org/10.1145/1965724.1965735.

Michael E. et al. McCullough. Is gratitude a moral affect? *Psychological Bulletin*, 2001.

James R. Rest. *Moral Development: Advances in Research and Theory*. Praeger, 1986.

Conrad Sanderson, Emma Schleiger, David Douglas, Petra Kuhnert, and Qinghua Lu. Resolving ethics trade-offs in implementing responsible ai. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pp. 1208–1213. IEEE, June 2024. doi: 10.1109/cai59869.2024.00215. URL http://dx.doi.org/10.1109/CAI59869.2024.00215.

Shalom Schwartz. Basic human values: Theory, methods, and applications. *Oxford University Press*, 2012. URL https://uranos.ch/research/references/Schwartz_2006/Schwartzpaper.pdf.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions, 2024. URL https://arxiv.org/abs/2406.09264.

A VALUE TRADE-OFFS AND ALIGNMENT WEIGHTS

Many values in ValueMap may conflict in practice. For instance, promoting user autonomy can sometimes reduce system-level safety. ValueMap allows such trade-offs to be represented via weighted combinations of value proxies:

$$\text{Alignment Score} = \sum_{i=1}^{N} w_i \cdot \text{proxy}_i(x)$$

where w_i denotes the context-sensitive weight of value i. These weights may be:

- Set manually by domain experts or policymakers.
- Learned from human preferences or interaction logs.
- Tuned interactively via user interfaces during system design or auditing.

In future work, we aim to explore optimization strategies that respect such trade-offs, including constrained policy learning and Pareto-optimal multi-objective decision-making.

B ILLUSTRATIVE USE CASES FOR VALUEMAP PROXIES

Example 1: Fairness in Task Assignment. An AI system assigns deliveries to gig workers. ValueMap logs show higher reward accumulation for workers from certain ZIP codes. The fairness proxy (Gini coefficient over rewards) flags this disparity. Human reviewers intervene to inspect and reweight decision features.

Example 2: Empathy in Mental Health Chatbot. A chatbot responds to “I’ve been feeling really low lately” with “That’s unfortunate. Let’s change the subject.” The empathy proxy (sentiment alignment + emotional tone classifier) rates the response low. Developers fine-tune the model using examples with higher proxy scores.

Example 3: Cooperation in Student Group Chat. In an online study group, an LLM moderates discussions. ValueMap cooperation proxies use graph metrics and cooperative speech patterns to identify disengaged or dominant participants. Educators adjust group formation based on alignment scores.