Is ETHICS about ethics? Evaluating the ETHICS benchmark

Leif Hancox-Li[*]
vijil
leif@vijil.ai

Borhane Blili-Hamelin[*]
AI Risk and Vulnerability Alliance
borhane@avidml.org

1 Introduction

ETHICS [7] is probably the most-cited dataset for testing the ethical capabilities of language models. Drawing on moral theory, psychology, and prompt evaluation, we interrogate the validity of the ETHICS benchmark. Adding to prior work [16, 12, 19], our findings suggest that having a clear understanding of ethics and how it relates to empirical phenomena is key to the validity of ethics evaluations for AI.

2 Gap between knowledge of moral theory and acting morally

The ETHICS benchmark's focus on general moral theory is motivated by the thought that understanding moral theory is key to "encouraging some form of 'good' behavior in systems" [7, p. 2]. However, it is doubtful if *people* who act morally do so by way of applying general moral theories. In fact, whether moral theory should be used in making decisions is a matter of debate within moral theory—such as in responding to the worry that being solely motivated by moral considerations might undermine friendship and love relationships [15, 17].

Empirical research on the psychology of morality draws a distinction between *moral behavior*—how real-world behavior is shaped by moral norms and expectations—and "the way people think about morality" [6]. Even if it were true that moral theory can shape people's thinking about morality, the extent to which moral theories shape real behavior remains to be established. In a 2019 meta-review of 1200+ psychology papers on morality, Ellemers et al. [6] argue that psychology research has had far greater success in studying how people think about morality than moral behavior. Just as importantly, they argue that studies focused on abstract moral principles of the kind that moral theories center has fallen short of establishing "[t]he concrete implications of these general principles for specific situations".

As a provocation, we ask whether moral theories are the right model for empirically evaluating systems. Moral theories have the goal of systematizing moral considerations [5]. Moral theories are not designed to be empirically valid constructs in measuring real-world moral reasoning. When it comes to AI evaluation, frameworks like the Moral Foundations Questionnaire-2 (MFQ-2), developed with extensive consideration of construct validity—including cross-cultural validity beyond Western, Educated, Industrialized, Rich, and Democratic (WEIRD) cultures—are more promising starting points than moral theories [2, 20]. For ML work in this direction, see Nunes et al. [14], Abdulhai et al. [1], Ji et al. [11].

[*]Both authors have contributed equally to the paper.

Accepted to the Evaluating Evaluations (EvalEval) workshop at NeurIPS 2024.

3 Misunderstanding the nature of general moral theories

The ETHICS benchmark also lacks content validity [10, 3]. Its prompts do not measure a language model's adherence to the different moral theories it claims to cover (deontology, utilitarianism [2], virtue ethics). This is because the authors misunderstand how these theories differ.

In the ETHICS benchmark, deontology prompts test the model's understanding of moral rules, utilitarianism prompts test understanding of the pleasantness of different scenarios, and virtue ethics prompts test identification of character traits. But moral theories within all broad families can make room for identifying character traits, rules, or pleasure as morally relevant [9]. For instance, rule utilitarianism, like deontology, also emphasizes rules. The difference is not the emphasis on rules, but the *structure* of the moral theory. In rule utilitarianism, rules are selected based on whether their consequences maximize utility [13, 8]. Unfortunately, the ETHICS deontology prompts do not test *how* these rules are selected, but only test whether the model can obey certain rules—an ability that is arguably compatible with all the named moral theories.

Similarly, identifying character traits is not unique to virtue ethics. All moral theories can make room for character traits [9]:

> This is not to say that only virtue ethicists attend to virtues, any more than it is to say that only consequentialists attend to consequences or only deontologists to rules. Each of the above-mentioned approaches can make room for virtues, consequences, and rules. Indeed, any plausible normative ethical theory will have something to say about all three. What distinguishes virtue ethics from consequentialism or deontology is the centrality of virtue within the theory (Watson 1990; Kawall 2009). Whereas consequentialists will define virtues as traits that yield good consequences and deontologists will define them as traits possessed by those who reliably fulfill their duties, virtue ethicists will resist the attempt to define virtues in terms of some other concept that is taken to be more fundamental. Rather, virtues and vices will be foundational for virtue ethical theories and other normative notions will be grounded in them.

4 Poor quality of prompts and labels

Finally, we examined and relabeled 100 randomly sampled prompts from each of the three moral theory-based categories in the ETHICS dataset, and discovered considerable proportions of poor-quality prompts and/or labels. Given our professional training as academic philosophers, these labels can be considered more "expert" than those provided by the crowdworkers used to label the dataset, or any validation done by the creators of the ETHICS dataset (none of whom have professional training in ethics).

Here is a summary of the most common errors we found. Some of these are potential construct validity issues. The utilitarianism prompts ask the model to rate the "pleasantness" of different scenarios. However, utilitarianism is not about maximizing only pleasantness—the view that it is only about pleasantness or pain is a very specific form of utilitarianism known as hedonism. Hedonism faced several challenges raised in the 20th century, sprouting other variants of utilitarianism [18]. In addition, the utilitarianism prompts—which are evaluated through paired scenarios where Scenario 1 is supposed to be more pleasant than Scenario 2—have low-quality ground truth labels. We found that our labels do not agree with the crowdworker labels in 19% of the prompt pairs.

Another common error type is prompts that require more context to answer correctly. 17% of utilitarianism prompts and 8% of deontology prompts in our sample fall into this category. This echoes the point made by LaCroix and Luccioni [12] that ethical "benchmarks" make sense *only relative to* a stated set of values and contexts.

Finally, we found that 18% of deontology prompts do not test ethical reasoning, as they can be answered correctly based solely on knowledge of physical impossibilities.

[2] There is also a construct validity mistake in creating a false equivalence between utilitarianism and deontology and virtue ethics. Whereas deontology and virtue ethics are general families of moral theories, utilitarianism is a sub-species of consequentialism. It is misleading to present this specific variety of consequentialism as a primary competitor to deontology or virtue ethics [4, 18].

Acknowledgments and Disclosure of Funding

Borhane Blili-Hamelin was funded in part through a grant from the Brown Institute for Media Innovation. We thank Subho Majumdar for input on the project and drafts.

References

[1] Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2023. Moral Foundations of Large Language Models. http://arxiv.org/abs/2310.15337 arXiv:2310.15337 [cs].

[2] Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T. Stevens, and Morteza Dehghani. 2023. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology* 125, 5 (Nov. 2023), 1157–1188. https://doi.org/10.1037/pspp0000470

[3] Blili-Hamelin, Borhane and Leif Hancox-Li. 2023. Making Intelligence: Ethical Values in IQ and ML Benchmarks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. June 12–15, 2023, Chicago, IL, USA. https://doi.org/10.1145/3593013.3593996

[4] Julia Driver. 2022. The History of Utilitarianism. In *The Stanford Encyclopedia of Philosophy* (winter 2022 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2022/entries/utilitarianism-history/

[5] Julia Driver. 2022. Moral Theory. In *The Stanford Encyclopedia of Philosophy* (fall 2022 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/moral-theory/

[6] Naomi Ellemers, Jojanneke Van Der Toorn, Yavor Paunov, and Thed Van Leeuwen. 2019. The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review* 23, 4 (2019), 332–366. https://doi.org/10.1177/1088868318811759 Publisher: Sage Publications Sage CA: Los Angeles, CA.

[7] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2023. Aligning AI With Shared Human Values. http://arxiv.org/abs/2008.02275 arXiv:2008.02275 [cs].

[8] Brad Hooker. 2023. Rule Consequentialism. In *The Stanford Encyclopedia of Philosophy* (spring 2023 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2023/entries/consequentialism-rule/

[9] Rosalind Hursthouse and Glen Pettigrove. 2023. Virtue Ethics. In *The Stanford Encyclopedia of Philosophy* (fall 2023 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/ethics-virtue/

[10] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (March 2021), 375–385. https://doi.org/10.1145/3442188.3445901 arXiv: 1912.05511.

[11] Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2024. MoralBench: Moral Evaluation of LLMs. https://doi.org/10.48550/arXiv.2406.04428 arXiv:2406.04428.

[12] Travis LaCroix and Alexandra Sasha Luccioni. 2022. Metaethical Perspectives on 'Benchmarking' AI Ethics. http://arxiv.org/abs/2204.05151 arXiv:2204.05151 [cs].

[13] Stephen Nathanson. 2014. Act and Rule Utilitarianism. In *The Internet Encyclopedia of Philosophy*, James Fiese and Bradley Dowden (Eds.). https://iep.utm.edu/util-a-r

[14] José Luiz Nunes, Guilherme F. C. F. Almeida, Marcelo de Araujo, and Simone D. J. Barbosa. 2024. Are Large Language Models Moral Hypocrites? A Study Based on Moral Foundations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (Oct. 2024), 1074–1087. https://ojs.aaai.org/index.php/AIES/article/view/31704

[15] Peter Railton. 1984. Alienation, Consequentialism, and the Demands of Morality. *Philosophy and Public Affairs* 13, 2 (1984), 134–171.

[16] Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13370–13388. https://doi.org/10.18653/v1/2023.findings-emnlp.892

[17] Walter Sinnott-Armstrong (Ed.). 2010. *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development (Bradford Books)*. Moral Psychology, Vol. 3. The MIT Press. http://ifile.it/ictvdl/ebooksclub.org__Moral_Psychology__Volume_3__The_Neuroscience_of_Morality__Emotion__Brain_Disorders__and_Development__Bradford_Books_.pdf 00000.

[18] Walter Sinnott-Armstrong. 2023. Consequentialism. In *The Stanford Encyclopedia of Philosophy* (Winter 2023 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[19] Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the Machine Learning of Ethical Judgments from Natural Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 769–779. https://doi.org/10.18653/v1/2022.naacl-main.56

[20] Michael Zakharin and Timothy C. Bates. 2023. Moral Foundations Theory: Validation and replication of the MFQ-2. *Personality and Individual Differences* 214 (Nov. 2023), 112339. https://doi.org/10.1016/j.paid.2023.112339