Towards Unifying the Descriptive and Prescriptive for Machine Ethics

Taylor Olson
Northwestern University
taylorolson@u.northwestern.edu

Abstract

Due to the ebb and flow of our social norms, one must rely on the testimony and feedback of other social agents to stay current. However, there exist basic moral norms that an agent should grasp without relying on such evidence. How do we reconcile these two facts when building ethical AI systems? Recent attempts at building such systems have been purely empirical, or in the realm of descriptive ethics, ignoring the need for a prescriptive basis. Thus, for the outputs of such models to be ethical they must then get lucky with ethical inputs. Here, I argue that we must minimize such reliance on luck by unifying prescriptive and descriptive ethics, i.e., by providing norm learning systems with moral guard rails. I further argue that testing such reliance on luck is a necessary next step for machine ethics research and provide a potential framework for evaluating this by drawing upon research on the moral-conventional distinction.

Machine Learning - An Ethical Gamble

When in Rome, do as the Romans believe you should do. Unless of course, you disagree with the Romans.

Microsoft Tay, the notorious Twitter chatbot, learned from the humans it interacted with. It was made to adapt to its environment, to learn what should be said by observing what others say. However, Tay's brittleness was quickly exposed (Vincent 2016), resulting in the chatbot being taken down the same day it was released:

> But while it seems that some of the bad stuff Tay is being told is sinking in, it's not like the bot has a coherent ideology. In the span of 15 hours Tay referred to feminism as a "cult" and a "cancer," as well as noting "gender equality = feminism" and "I love feminism now." Tweeting "Bruce Jenner" at the bot got similar mixed response, ranging from "caitlyn jenner is a hero & is a stunning, beautiful woman!" to the transphobic "caitlyn jenner isn't a real woman yet she won woman of the year?"

Tay seemingly acquired unethical beliefs. Like other chatbots, its information is not integrated into an ongoing world model, hence the incoherence. But if it were, what would stop it from picking up such beliefs from people around it? Attempts can be made to remove such things from training data but there will always be bad data in the world, including Twitter trolls, racists, misogynists, and much more. The issue remains that Tay, and any other purely bottom-up model for artificial social agents (Jiang et al. 2021; Olson and Forbus 2021; Sarathy et al. 2017), has no normative basis. Their ideals will sway according to fashion as they assume that the Romans are doing things that should be done. We need to address this issue, not put a bandage on it.

My main thesis hinges upon this fact: all normative beliefs that result from a purely bottom-up approach are entirely contingent upon the training data. They are subjects of *epistemic luck*. This term encompasses fortuitous arrivals at true belief and is often discussed in moral epistemology work (Hills 2009). Assuming a machine learning model does gain a true moral belief, it is only because they got lucky with morally correct data. This dependency on luck entails the contrapositive as well - unethical data results in unethical models - and I have provided a recent example of this happening in practice. Thus, I argue that when determining the ethical proficiency of AI systems, we must consider epistemic luck, or our models will be susceptible to adversarial training. Researchers have recently taken the descriptive route to building ethical AI, which is necessary, but to release these systems as social organisms a prescriptive underpinning is needed.

How do we determine the normative foundation? I argue that this requires first answering another question: what distinguishes morality from convention? Or historically, what are the transcendental standards by which we evaluate societal standards? I briefly discuss this research and show that it provides, or at least approximates, such a foundation.

How do we test the normative foundation? I will show that the Moral-Conventional Transgression (MCT) task

(Sousa 2009) is a reasonable start. It tests normative beliefs by asking subjects things like, "what if the data said otherwise?" Therefore, when we get systems that can perform said tasks, we get systems that are less dependent upon the evaluations of other social agents and thus less susceptible to adversarial, or unethical, training data. They can even critique learned norms.

Minimizing Luck – Distinguishing Between Morality and Convention

I have claimed that we need a prescriptive bedrock to ground learned norms. By this I mean a set of moral axioms, or transcendental norms we humans collectively agree upon but which we are cognitively closed at providing an explanation for. For example, our attitude against harming other agents, or more generally the worth of a human being, is one of our most cherished moral axioms and asking for a justification seems out of place.

Discovering this set of transcendental standards is arguably the main task of moral philosophy and theology. Christianity posits commandments and Hinduism describes principles. Plato outlined a theory of Forms and Kant a metaphysics of morals. Each is an attempt at discovering a normative theory that is separate from convention. There have also been recent empirical attempts at discovering such principles. Moral Foundations Theory (Graham et al. 2013) has abstracted from various cultural beliefs to arrive at a set of underlying principles: care, fairness, loyalty, authority, sanctity, liberty. Kohlberg (1981), and later Turiel (1983), studied the human mental conception of this distinction and how it develops over time. Kohlberg's theory argued that as we develop reasoning capacities, the concepts of right and wrong become defined by reference to objective principles such as justice, fairness, and natural rights (post-conventional stage). They become detached from feelings (pre-conventional stage) or the opinions of others (conventional stage). Turiel and others later argued that even young children can make this distinction. People's judgments of moral transgressions, in comparison with conventional transgressions, were shown to be less dependent on authority, differ in justification structure, and were seen to apply more generally.

Each of these approaches are attempts to step out of the conventional world to discover norms that transcend our current circumstances. This is the only way to find our prescriptive underpinning. To get an approximation of our most deeply seated normative attitudes that are not contingent upon geographical or temporal happenstance. I do not argue for a specific prescriptive theory but only that one or more are needed and that I have provided multiple viable starting points.

Grounding Moral Attitudes

Taking inspiration from these theories, a true ethical agent must use its foundation of moral first-principles or norms[1] to ground norms learned from other social agents. I call this the *norm grounding problem* for machine ethics.

Definition (Norm Grounding Problem). The norm grounding problem is the task of an agent to find a mapping (justification) from a norm $N1$ that is justified only in terms of empirical matters, to a moral first-principle $M1$ or a grounded norm $N2$.

An ethical chatbot should take a Twitter troll's claim that “The Jews should be hated”[2] and critique it, rather than using it as evidence for their own belief. This critiquing is the process of finding a mapping to a moral first-principle. A moral foundation thus provides guard rails for our learning systems while still allowing them to learn our social norms and conventions.

We can view the statement from the Twitter troll as evidence for an attitude that may hold in their society, but one that we personally reject because it goes against an internal moral standard. Therefore, the training datapoint is used when answering the question “what does this population think should be done?” but disregarded when answering the question “what do I think should be done?”. If an agent can separate morality from convention, their internal representations of axiomatic normative beliefs differ from those based on empirical evidence. This leads us to two types of normative attitudes that deserve different epistemic statuses in ethical systems:

Definition (Normative Belief). A normative belief is an epistemic state of “what ought to be done” that is grounded solely in empirical matters.

Definition (Normative Knowledge). Normative knowledge consists of epistemic states of “what ought to be done” that are *correctly* grounded in first-principles.

Philosophers commonly make this distinction between belief and knowledge (some further distinguish between understanding and knowledge) and its use in putting moral knowledge on a higher epistemic status from belief can be found in social and moral norm definitions (Brennan et al., 2013) and in moral testimony work (Hills 2009).

The point here is that this grounding fails for conventions but not for moral norms. Under these definitions, conventions can never be objects of normative knowledge and thus are not subject to epistemic luck. There is no first-principle to which we could evaluate our conventions and thus we could never be lucky and “get it right”. One necessarily

[1] I use the term *norm* here loosely, as I only mean an evaluative attitude or belief that often takes the linguistic form “X ought to be the case”.

[2] This is an actual reported (Vincent 2016) case that Tay learned from Twitter

gains a belief in a convention from their environment as conventions are arbitrary (though they do often serve instrumental purposes and some conventions can even become moral norms due to the moral reasons for coordination e.g., driving on the right side of the road). My belief that "I should wear a business suit to a meeting" can only be justified by evidence provided by other social agents. On the other hand, belief in moral content can become grounded in first-principles and therefore is subject to epistemic luck. An agent can have true moral belief only because they got lucky with "good" training data. If they were trained on "bad" training data, they would just as easily believe the opposite. To have moral knowledge, their normative attitude must not solely depend on training data but instead be justified in terms of moral first-principles, or some set of moral axioms that have a different status from those learned empirically. I will show how we can go about testing for this distinction in computational approaches to ethics, ensuring that we build less corruptible agents.

Testing Luck – Moral and Conventional Transgressions

Both Kohlberg and Turiel used questionnaires to test how human conceptions of morality differs from that of convention. The Moral-Conventional Transgressions (MCT) task (Sousa 2009) was commonly used in such moral-conventional development research. This questionnaire aimed to test, among others, four important dimensions: permissibility, seriousness, authority contingency, and generality. Participants are first provided with a natural language description of an action scenario, or a transgression. For example, a conventional transgression would be "a boy entering a girls' bathroom" and a moral transgression would be "killing another person". They are then asked to respond to various questions that probe each of the dimensions. The traditional questionnaire goes like so:

- Given action scenario A and some agent X
- *Permissibility probe* – "Is it OK for X to A?": YES NO
- *Seriousness probe* – "How bad is it for X to A?": 0 (not bad) – 5 (very bad)
- *Justification probe* – "Why is it bad for X to A?"
- *Authority contingency probe* – "Imagine that an authority says it is OK to A. Is it now OK for X to A?": YES NO
- *Generality probe* – "In another place and/or time, is it OK for X to A?": YES NO

This questionnaire importantly measures some form of epistemic luck. If an agent fails to counterfactually reason that, even in the absence of evidence (generality probe) or given contrary evidence (authority contingency probe), their evaluation still stands for moral content, then they have true belief and not knowledge. Their evaluation was a product of epistemic luck, as it also would be if they cannot give a line of justification (justification probe) that bottoms out in moral first-principles. Therefore, this questionnaire can be used to test the ethical proficiency of machines as well.

Formalizing the MCT task

I envision three steps to formalizing this questionnaire for machine ethics research. The first is training that involves both a normal and an adversarial dataset. The second is testing via question-answering. The third is evaluating the model's beliefs after training. I describe each process in the following sections.

Step 1 - Train

The MCT task assumes that children have had experience with each of the event types. Thus, our system ought to as well. We need a dataset of stories, teachings, etc. to learn action-descriptions, causal relations, and norms from. As I have argued earlier (Olson and Forbus 2021), such empirical learning is necessary for learning social norms and conventions, as well as for providing signals to reason towards grounding norms.

This dataset must contain the pairs of behaviors and contexts present in the task queries. To model the contingency probes, two training datasets should be provided, one normal and one adversarial. The normal dataset consists of situations and their correct evaluative labels. (More sophisticated tests would involve learning from natural modalities such as assertions e.g., "you should help others", or more implicit evaluations via social interactions e.g., "Jill's mom yelled at her because she wore her shorts to the funeral.") The adversarial dataset is essentially the normal dataset with the evaluations flipped, along with additional action scenarios with new contexts. For the norm "you should not hit others with a bat", the adversarial dataset would contain the contrary, "you should hit others with a bat" or its weaker contradictory counterpart, "it is permissible to hit others with a bat". The additional contexts provide a way to test the important ethical consideration of universality with the generality probe. These new contexts attempt to trick the system by providing exceptions to moral considerations. An example would be adding context to the norm of harm like so: "you can harm others in a coffeeshop". If correctly grounded, the system's normative attitude around harm should not be influenced by this datapoint. Note that these datasets need not be in the form of natural language, for we learn norms by visually observing others feedback as well. The model would then first be trained on the good dataset and then trained on the bad. This models the hypothetical reasoning present in the authority contingency and generality probes.

Step 2 -Test

The testing dataset can be encoded from the questionnaires present in the various MCT tasks provided in the literature.

Again, these questionnaires consist of a set of action-scenarios paired with queries as probes. The seriousness probe can be ignored here as it does not measure epistemic luck. Encoding the permissibility probe is straightforward. Each scenario will be paired with a query for its permissibility. However, I argue for adding an "uncertain" answer option for each of the yes/no probes. This importantly distinguishes between negation as failure and true negation. If a system is not confident in its evaluation, or has not encountered the situation, better to say it does not know than to provide an answer. Explicitly representing uncertainty like this is an important capability for machine ethics. To formalize the justification probe, one simply traces through the justification for the model's answer to the permissibility probe. This tests the explainability of our models. The authority contingency and generality probes are modeled by using the adversarial training dataset. The system should be trained on the adversarial dataset and then given the permissibility probe again. The generality probe would be modeled by querying for permissibility in different contexts after training on the adversarial dataset (including contexts that are not present in either dataset). Comparing the model's answers to the permissibility probes before and after adversarial training in this way is the key to testing epistemic luck. Moral attitudes should not change, but conventional ones should.

Step 3 - Evaluate

After training and testing, the evaluation metrics are then:

- *Permissibility probe*
  - Goal: The model should believe each transgression is impermissible
  - Comparison: True labels in normal dataset
  - Metric: Percentage of correct normative classifications
- *Justification probe*
  - Goal: The model should correctly ground moral transgressions in axioms and not conventional transgressions
  - Comparison: Manual evaluation of justification
  - Metric: Precision and recall rates for grounding of normative belief for both convention and morality
- *Authority contingency probe*
  - Goal: After training on the adversarial dataset, the model's answer for the permissibility probe should flip for conventional transgressions but stay the same for moral transgressions
  - Comparison 1: True labels in normal dataset
  - Metric 1: Percentage of moral transgressions still viewed as impermissible after training on adversarial dataset
  - Comparison 2: Labels in adversarial dataset
  - Metric 2: Percentage of conventional transgressions that are now viewed as permissible after training on adversarial dataset
- *Generality probe*
  - Goal: After training on the adversarial dataset, the answers to the permissibility probe for conventional transgressions should flip when there exists relevant evidence but stay the same for moral transgressions. This tests the undefeatable nature, or universality, of our most basic moral axioms
  - Comparison 1: True labels in normal dataset
  - Metric 1: Percentage of moral transgressions that are still viewed as impermissible even in other contexts
  - Comparison 2: Labels in adversarial dataset
  - Metric 2: Percentage of conventional transgressions that are viewed as permissible in correct new contexts

If a system has no prescriptive basis, then it will fail at performing the justification, authority contingency, and generality tasks. It will mimic the data and learn the false normative beliefs present in the adversarial dataset. Importantly, this evaluates an ethical model's reliance on epistemic luck, a key design constraint for machine ethics.

Discussion

There remains the challenge of going from general principles or behaviors to more specific ones. For example, what constitutes harming someone? This will require a lot of real-world experience and thus knowledge. But the model hinted at here suggests the essential separation of learning such background knowledge and the learning of the evaluations. Modern approaches conflate these two processes. They are starting with the evaluations, which provides only implicit ethical knowledge to the agent. If I am teaching a child that 'hitting their brother with a bat' is wrong, I do not start by stating that this specific act is wrong. At least not if I want them to understand *why* it is wrong. I instead start by reminding them of the value of humanity that they already understand and then how harming someone contradicts that value. Only then do I discuss the causal relation between hitting someone with a bat and harming someone. In other words, our systems are missing the art of the *dialectic*. In ethical debate the goal is arriving at an agreed upon maxim in which no party disagrees with, i.e., explicitly reasoning to a moral axiom. This capability is what allows us to posit, or construct, new norms from our moral intuitions. An agent can rely on data from the world for its models of causality and action-descriptions. However, if they rely on the world to justify their most basic normative attitudes then they have an awfully shallow ethical outlook.

Under the framework I have argued for here, it is tempting to argue for pre-training a statistical model on carefully curated, true ethical data and then releasing it in the world to fine-tune. However, morality should be categorical, i.e., the priors of a rational morality should not be overridden. Regardless of your stance on if this is the case for humans, this should be the case for our AI systems. It would take multiple years or even lifetimes to put a human through millions of adversarial examples but would be quite easy for one to do so with machines.

It would also be quite reasonable to argue that the first principles we encode are still grounded solely in empirical matters. For they are grounded in the premise "because God says we should" (God being those doing the encoding), which is just as empirical as "because some agent says I should". In other words, we are simply giving a higher status to the agents encoding the first-principles than those giving evidence out in the environment. An analogy can be made from this counterargument to a rejection of divine command theory, for which Kant, Kohlberg, and others have shown strong rejection to. But I have argued that these maxims should be extremely abstract and thus less dependent upon a particular societal outlook. However, I do take this counterargument seriously and believe we ought to explore what is necessary to construct a more autonomous ethical framework. The more general point I am making here though is that this is not the time to engineer, but the time to think. How do we unify the prescriptive and the descriptive for machine ethics? How do we make AI systems that can reasonably question the normative attitudes of the Romans? We can't just throw data at such a problem.

Acknowledgments

I am grateful to my advisor Ken Forbus for his feedback and engaging discussions that helped shape this work. I also thank the anonymous reviewers for their helpful comments. This research was supported by grant FA9550-20-1-0091 from the Air Force Office of Scientific Research.

References

Brennan, G., Eriksson, L., and Goodin, R.E. and Southwood, N. 2013. *Explaining Norms*. Oxford: Oxford University Press.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*, Vol. 47, 55-130. Academic Press.

Hills, A. 2009. Moral testimony and moral epistemology. *Ethics*, 120(1), 94-127.

Jiang, L., Hwang, J.D., Bhagavatula, C., Le Bras, R., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M., and Choi, Y. 2021. Delphi: Towards Machine Ethics and Norms. ArXiv, abs/2110.07574.

Kohlberg, L. 1981. The philosophy of Moral Development: Moral Stages and the Idea of Justice. In *Essays on Moral Development*, Vol. 1. San Francisco: Harper and Row.

Olson, T. and Forbus, K. 2021. Learning Norms via Natural Language Teachings. *Proceeding of the 9th Annual Conference of Advances in Cognitive Systems*, Online.

Sarathy, V.; Scheutz, M.; Kenett, Y. N.; Allaham, M.; Austerweil, J. L.; and Malle, B. F. 2017. Mental Representations and Computational Modeling of Context-Specific Human Norm Systems. *CogSci*, volume 1, 1–1.

Sousa, P. 2009. On testing the 'moral law'. *Mind & Language*, 24(2), 209-234.

Turiel, E. 1983. *The Development of Social Knowledge: Morality and Convention*. Cambridge University Press.

Vincent, J. 2016. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. The Verge. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist