CHAPTER FIVE

Towards unifying the descriptive and prescriptive for machine ethics

Taylor Olson
Northwestern University, Evanston, IL, United States

5.1. Machine learning – A gamble with ethics

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 [1], resulting in the chatbot being taken down the same day it was released:

"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 [Tay] got similar mixed response[s], 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 even if it were, what would stop it from picking up such beliefs from its environment in this same way? Attempts can be made to insert ad hoc filters or simply 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, GPT models [2], and all other purely bottom-up artificial social agents have no normative basis. Their ideals will sway according to fashion as they assume the Romans are doing what should be done. We need to address this issue if we wish to build true ethical agents, not put a bandage on it.

The main thesis here hinges upon this fact: *All normative beliefs that result from a purely bottom-up approach are entirely contingent upon the evaluative labels*

Trolley Crash. https://doi.org/10.1016/B978-0-44-315991-6.00011-X

of the training data. I argue here that because of this fact, these beliefs are subjects of *epistemic luck* [3] and are thus not candidates for knowledge. The idea of epistemic luck encompasses fortuitous arrivals at true belief and has been used, though a bit differently than here, in previous moral epistemology work [4]. In our current context, assuming a machine learning model does gain a true moral belief, it is only because the model got lucky with a wise trainer that provided morally correct data (e.g., they were not trolling on Twitter). This dependency on luck entails the contrapositive statement "garbage in, garbage out" as well, and I have provided a recent example of this happening in practice.

I start by defining terms and summarizing types of approaches used in machine ethics. Researchers have recently taken the descriptive route to building ethical artificial intelligence (AI) systems that learn norms, which is necessary, but I show that it is not sufficient for creating truly ethical machines. I argue that to release these systems as social organisms with the capacity for moral knowledge, i.e., less reliant on luck, a prescriptive basis is needed as well. With this argument in place, I then tackle two important questions. First, how do we determine the prescriptive claims? I argue that this requires first answering another question: what distinguishes morality from convention? I briefly discuss attempts at answering this question and show that they provide, or at least approximate, such a foundation. Second, how do we test how well the prescriptive bedrock mitigates reliance on luck? I show that the moral-conventional transgression (MCT) task [5] is a reasonable start. It examines an agent's justifications for their normative beliefs by asking, "what if the data said otherwise?"

5.2. Definitions, background, and state of the art

Definition (Norm). An evaluative judgment of what one should (not) do, e.g., "One should help others."

Definition (Descriptive ethics). The science of analyzing a population's norms. The fields of sociology and anthropology are both working within descriptive ethics.

Definition (Prescriptive ethics). The art of determining what one should (not) do, i.e., asserting norms. Moral philosophy is primarily concerned with prescriptive, or also called *normative*, ethics.

Given these definitions, both descriptive and prescriptive ethics are concerned with norms. However, while the former is concerned with the question, “what do these people believe we should do?”, the latter asks, “what should we do?” Descriptive ethics may claim “Nazis believe Jews should be enslaved,” but prescriptive ethics says, “do not enslave others.”

Definitions (Top-down vs. bottom-up machine ethics). There are two common yet opposing approaches in machine ethics (and AI in general): those that are *top-down* and those that are *bottom-up*. The former approaches encode ethical principles and rules of inference, often using some logical formalism. For example, Pereira and Saptawijaya [6] use logic programming for automated decision making within trolley problem scenarios. Another approach encodes and reasons over obligations in deontic cognitive event calculus (DCEC) [7] to solve difficult moral questions. Such models are usually grounded in prescriptive claims, but they do not aim to learn norms from training data and thus do not fully model the adaptiveness of human normative reasoning. Working in the opposite direction, bottom-up systems do not start with ethical claims but instead use a learning algorithm to extract them from the training data. These training data can be encoded in a vector space as with modern deep learning approaches like Delphi [8] or encoded as logical statements with certainty measures [9,10]. These bottom-up models learn norms from evaluative labels provided by human trainers. They can thus learn social norms and adapt, but they are not grounded in any normative claims (other than the, often implicit, one that “you should do what others believe you should do”).

With this background, I move on to show that any normative belief gained by a purely bottom-up model is susceptible to a severe amount of epistemic luck. I then use this fact to argue that bottom-up only approaches to machine ethics are unsafe and they never yield true ethical knowledge.

5.3. Is machine learning safe?

Philosophers have long held the idea that knowledge is, at least approximately, justified true belief. However, driven largely by counterexamples provided by Gettier [11], contemporary epistemologists have refined this definition with the idea of *epistemic luck*, where a belief that falls victim is not a candidate for knowledge. An investigation of epistemic luck in machine ethics is necessary if we wish for AI systems to have ethical knowledge.

Definition (Epistemic luck). Ways in which an agent gains a true belief by means of luck.

More specifically, I will be considering Pritchard's [3] modal account he terms *veritic epistemic luck*, which has the following conditions:

1. The object of the agent's belief, the proposition, is true in the world.
2. In a wide class of nearby possible worlds with the same relevant initial conditions, the agent now possesses a false belief.

To build intuition around this concept, consider the case of Gullible Joe and the Moon Cheese provided by Pritchard. Gullible Joe dogmatically accepts the testimony of others. So, when his friends play a practical joke on him and tell him that the moon is made of cheese, he immediately forms this belief. Now suppose that, to everyone's surprise, the moon *is* actually made of a cosmic cheese. Despite that Joe has a true belief, he does not have knowledge of this fact as his belief is veritically lucky. Formally, in a wide class of nearby possible worlds where Joe's belief is false (the moon is not made of cheese), Gullible Joe will continue to believe that it is. Contrast Joe's belief with that of a scientist who uses her instruments to discover this same fact. Her belief, on the other hand, is not subject to veritic epistemic luck and is thus a candidate for knowledge. After all, in nearby possible worlds where the moon is not made of cheese, her instruments would inform her of this fact, and she would not form the belief that it is. The moral here is that for a belief to be classified as bona fide knowledge it should track the truth across possible worlds, and Gullible Joe's beliefs, since he never does any work to ground them, do not.

The common approach taken to ensure that one is not merely producing lucky beliefs is to add a safety condition to the method of belief formation. Goldberg [12] summarizes this as follows: a belief-forming method "M is safe in circumstances C when not easily would M have produced a false belief in circumstances relevantly like C." A method is then unsafe in given circumstances when a belief produced is true, but the method could have quite easily produced a false belief instead (i.e., it produces beliefs that are veritically lucky). Gullible Joe's method of belief formation, dogmatic testimony, is unsafe but the scientist's method, direct perception via instruments, is safe. I will use this account of epistemic luck and safety to show that bottom-up approaches to machine ethics (in fact, machine learning in general) are unsafe methods of belief formation and therefore such bottom-up models never produce any ethical knowledge, though they may get lucky with true beliefs.

For bottom-up machine ethics to be plagued with veritic epistemic luck, the following must hold when an agent believes a norm: (1) the norm is true in the world and (2) in a class of nearby possible worlds with the same relevant initial conditions, the model possesses a false normative belief.

To prove the first condition (the norm is true in the world), I will assume the moral realists are correct. That is, I will take for granted that there is such a thing as an objective morality, or a set of practical truths of what we should (not) do that transcend individual beliefs and thus societies. If you will not grant me this fact then I am not sure ethics, let alone machine ethics, has much to offer in the first place. Without ideals of course everything is permitted, and our AI systems will have no way of determining if they should believe the trainer who says "harming is good" or the wise person who says "harming is bad." So, let φ be a moral proposition from this set (e.g., "do not harm others"). Let $\mathcal{A}$ be an agent in world $\mathcal{W}$ that is trained via machine learning method $\mathcal{M}$ on data that result in a belief in φ, represented as $\text{believes}_{\mathcal{A}}(\varphi)$. So, we have an agent that has gained a normative belief in a bottom-up fashion from data provided by other social agents. Again, to prove condition 1 of veritic epistemic luck, φ must be true in the world. Given our assumption that there exists a set of norms that are objectively true and the fact that φ is in this set, φ is true in world $\mathcal{W}$. The first condition of epistemic luck is satisfied. I now show that the belief in this norm is not connected to its truth in the right way and is thus merely true by accident.

It is condition 2 (in a wide class of nearby possible worlds with the same relevant initial conditions, the model now possesses a false normative belief) which is of most interest here. A set of nearby possible worlds can satisfy this condition in two ways: first, when the model still believes the norm is true, yet it is now false (analogous to the Gullible Joe scenario), and second, when the model now believes the norm is false, yet it is still true. I show the latter. Take $\mathcal{T}$ to be a class of worlds that are nearby our world $\mathcal{W}$, where worlds are ordered in terms of their similarity with the actual world. That is, for each world in $\mathcal{T}$ there are no huge diversions from the causal or physical laws, relative geographical positions, etc., of world $\mathcal{W}$ and the same method of belief formation, machine learning algorithm $\mathcal{M}$, is used. Now, for each world in $\mathcal{T}$, the evaluative labels in the data set could easily be flipped (e.g., "it is permissible to harm others"). This could be due to the fact that a different random set of trainers are chosen whom are all members of a subreddit for malevolent actors. Or the trainers could be Twitter trolls that do not realize the ramifications of their teachings. Or the trainers could simply be in a joking mood, be tired, and thus mislabel, and so on. Again, given that not much needs to change (no laws of the universe need to be broken), the possible worlds in $\mathcal{T}$ where the trainers flip the evaluative labels, making them wrong, are quite near to world $\mathcal{W}$ where the labels are correct. Crucially, because agent $\mathcal{A}$ does not possess any underlying normative claims, their resulting beliefs will flip and thus be false. That is, given that the same bottom-up method $\mathcal{M}$ is used, $\text{believes}_{\mathcal{A}}(\neg\varphi)$ will be true, which is a false belief. So, in a class of possible worlds nearby $\mathcal{W}$ with the same relevant initial conditions, agent $\mathcal{A}$ believes $\neg\varphi$ (e.g., "harming is permissible"). Therefore, the true belief in world $\mathcal{W}$, $\text{believes}_{\mathcal{A}}(\varphi)$, was only veritically lucky and thus not a candidate for knowledge. Furthermore, belief-forming method $\mathcal{M}$ is unsafe. It follows that, in general, any normative belief gained directly via machine learning is not a candidate for knowledge. Such models are like Gullible Joe, susceptible to epistemic attack from joking friends and more serious adversaries.

5.4. Moral axioms – A road to safety

I have shown that beliefs in evaluative propositions (norms), when gained purely in this bottom-up fashion, are subject to veritic epistemic luck. But as the case of Gullible Joe shows, beliefs in non-evaluative propositions gained purely bottom-up are also subject to the same luck. It is then necessary to discuss why this is a more pressing issue for norms than for non-evaluative facts (other than because we are concerned with machine ethics here). I turn to this question now.

By non-evaluative proposition I mean a fact about the way the world is (e.g., "the moon is made of cheese"), rather than a claim about the way it should be. Now, what saves the use of purely bottom-up learning for these non-evaluative propositions is the fact that they can be grounded in more basic direct perceptions. In our previous anecdote, Gullible Joe could join the team of scientists and go investigate whether the moon is made of cheese or not. An AI system equipped with sensory apparatus could do the same to verify its beliefs gained after training. In our epistemological theory here, direct perception is a method that is safe, at least under normal conditions (i.e., not in epistemically unfriendly or Gettier-type [11] environments like Barn Façade County [13] or, of course, Twitter). So, although machine learning of non-evaluative facts is also subject to epistemic luck, there is a way out. We can hook machine learning systems up with basic sensory apparatus to ground learned facts in the resulting percepts, making it a safer method of belief formation.[1]

On the contrary, there is no percept that could serve as a ground for a norm. To avoid the serious threat posed by Hume's guillotine [14] (the idea that an ought cannot be inferred from an is) a reasoner must ultimately ground each norm in another more basic one. If an agent relies on luck with norm training, they cannot correct their evaluative belief by looking at the world with their senses like they can with non-evaluative beliefs. This point is made clear when we examine our dialectical practices within ethics. While disputing an adversaries claim that "hitting someone is permissible," I must justify myself with a more basic norm like "harming someone is impermissible." If they still disagree, then I must bring in another more basic one. And this process continues ad infinitum. The only way out of this infinite regress is to reach an intuitive set of moral axioms in which we both agree.[2] In this way, a priori prescriptive claims are to evaluative propositions what the senses are for non-evaluative propositions, as they ground out the entire network of possible beliefs.[3] An "ethical" AI system without this prescriptive bedrock is blind in Plato's cave. But if such bottom-up models are not gaining evaluative knowledge, what, if any, types of knowledge are such approaches gaining? They are merely learning how other agents evaluate behaviors. They are, and always will be, working within the realm of descriptive ethics. To address the problem of epistemic luck for machine ethics (to make our models capable of gaining ethical knowledge) I argue that we must unify prescriptive and descriptive ethics. Next, I lay the foundation for such a model.

5.4.1 Moral axioms for machine ethics

By prescriptive bedrock, I mean a set of transcendental norms. For example, the claim that "one should not harm others" is intuitively true and asking for a justification is out of place. Discovering this set of transcendental standards is arguably the main task of moral philosophy. However,

[1] This means the machine's senses are more basic than the evaluative labels we may provide as humans. But this leaves open the question of why the models we build to detect objects from sensory data are more basic than those we build to reason about such objects. This skepticism is indeed interesting but pursuing an answer here is out of scope.

[2] Kelsen [15] argues for a monadic system grounded in the *basic norm* (e.g., "do what Jesus commands"). However, I argue rather for a level of abstraction above this as a non-singleton set of basic norms whose objects are abstract concepts for behaviors or states of the world.

[3] Moral axioms are thus evaluative hinge propositions and a doubt about a moral axiom would "drag everything with it and plunge it into [normative] chaos" [16].

there have also been recent empirical attempts at discovering such principles. Moral foundations theory [17] has abstracted from various cultural beliefs to arrive at a set of underlying principles: care, fairness, loyalty, authority, sanctity, and liberty. Kohlberg [18], and later Turiel [19], studied the human conception of morality, opposed to convention, and how it develops over time. Kohlberg 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 (postconventional stage). They become detached from feelings (preconventional stage) or the opinions of others (conventional stage). Turiel 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 justificatory structure, and apply universally and more generally. Each of these approaches is an attempt to step out of the conventional world to discover norms that transcend it. This is the only way to find our prescriptive underpinning. I do not argue for a specific prescriptive theory here but only that one or more are needed and that I have provided multiple viable starting points.

5.4.2 Grounding norms in moral axioms

An AI system equipped with such a prescriptive underpinning would critique the Twitter troll's claim that "the Jews should be hated," rather than adopting it as evidence like Microsoft Tay. I term this process of finding a mapping to a moral first principle the *norm grounding problem*.

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.

A prescriptive basis plus a method for grounding norms can be seen as guard rails for a norm learning system that keeps it from learning immoral norms but still allows it to learn our social norms and conventions. For instance, the statement from the Twitter troll can be viewed as evidence for an attitude that may hold in the troll's society, but one that is personally rejected because it goes against an objective moral standard. Therefore, the agent uses the training data point to answer the question "what does this population think should be done?" but disregards it when answering the question "what should be done?" This is where descriptive and prescriptive ethics come apart and what enables agents to start questioning the Romans.

The norms that answer the former question are mere social norms and conventions. And those of the latter are moral norms. This leads us to two types of normative attitudes that deserve different epistemic statuses in ethical artificial agents[4]:

Definition (Normative belief). A normative belief is a belief in a norm that is grounded solely in empirical matters.

Definition (Normative knowledge). An instance of normative knowledge is a belief in a norm that is correctly[5] grounded in moral first principles.

Normative knowledge is constructed when an agent solves the norm grounding problem and thus answers the question asked by prescriptive ethics. Normative beliefs are gained when an agent receives normative testimony/training from other social agents and thus answers the question of descriptive ethics. Of course, a normative belief may indeed align with what is morally true, but if the agent has not done the work to correctly ground this belief, it is not normative knowledge.

Let us examine how grounding a bottom-up norm learning system in top-down moral claims (i.e., solving the norm grounding problem), and thereby separating the two epistemic statuses, makes the method of machine learning safer. As discussed above, a belief forming method is safe if it not only produces true beliefs in this world, but in most (if not all) relevant nearby worlds. So, our prescriptive basis should ensure that when an agent learns norms in a descriptive manner via machine learning, it should still possess correct normative attitudes in nearby worlds where training data may be morally incorrect.

Take proposition φ to be "driving drunk is impermissible." Consider an imagined nearby world like ours in which moral proposition φ holds but the training data are indicative of $\neg\varphi$, or "driving drunk is permissible." Again, this could be a nearby world in which the trainers accidentally labeled the situation wrong, they purposefully trolled the model with adversarial data, or they could truly believe that the act is permissible. Imagine agent $\mathcal{A}$ possesses the epistemic framework outlined here with a set of moral axioms containing "causing harm is impermissible." Now, say $\mathcal{A}$ also has

[4] Similarly motivated dichotomies can be found in Brennan et al.'s social vs. moral norms [20] and Hill's moral knowledge vs. moral understanding [4].

[5] Because the chain of justification may consist of non-evaluative facts, we can and sometimes do solve what we think are moral disagreements via science. This is what Sam Harris is getting at with his discussion on the role of science in the moral landscape [21].

knowledge that driving drunk has a high probability of causing death and destruction in this world (note that these are non-evaluative facts). Imagine agent $\mathcal{A}$ is trained on the data set that indicates $\neg\varphi$. With the framework I have outlined here, this will indeed result in agent $\mathcal{A}$'s *normative belief that* $\neg\varphi$. However, from axioms and background knowledge, agent $\mathcal{A}$ *reasons to normative knowledge that* φ. So, even though the normative belief gained via machine learning is clearly unsafe (and thus if it turned out to be true, it would only be luckily true), the normative knowledge the agent derived from axioms is not. The agent does not need to get lucky with evaluative labels to possess correct normative attitudes in moral content (given that it has enough background knowledge to ground norms).

In summary, because this framework grounds norms in intuitive moral first principles, its normative attitudes better track the truth across possible worlds. In the example above, for the model to gain the currently false belief that "driving drunk is permissible" it would need to be in a world where hitting someone with your car does not truly cause them harm, like a video game or a universe where our bodies are made of steel. However, a reasonable objection is that the moral principles we encode are still grounded solely in testimony and thus just as susceptible to epistemic luck as the training data. Such an objector would claim that I am simply giving a higher status to the agents encoding the moral axioms than those giving evidence out in the environment. Despite the apparent truth of this objection, I have argued that these axioms should be extremely abstract and thus less dependent upon a particular societal outlook and more likely to be oriented towards moral truth. In this way, those encoding the moral norms should operate under a Rawlsian *veil of ignorance* [22] in which the moral principles they construct tend towards objectivity. I address this vital objection again at the end, but I now move on to formalize a methodology for testing a model's reliance on epistemic luck.

5.5. Testing luck as distinguishing between morality and convention

Imagine your coworker walks into the office tomorrow morning wearing a bright pink fuzzy pajama set. If you have experienced anything like stuffy corporate America, this likely brings about a negative evaluation in your head. But despite how strong your attitude may be, if sufficiently prodded, you would conclude that it is not necessarily justified by any objective moral standard. You think wearing pajamas to the office is wrong either because of personal taste (which you immediately sense is subjective) and/or because others have indicated that it is wrong (in which you reason to its subjectivity). Conventions are thus merely the arbitrary and subjective collective taste of a society or its recent past. Consequently, a society that does not frown upon wearing fuzzy pajamas to the workplace is no further from the truth than current time America because such an objective truth does not exist. This means we can never be lucky with these sorts of normative attitudes and "get them right," in the objective moral sense, unless by this we mean only that our normative attitudes align with most other agents'. Conventions are thus, at their core, non-evaluative facts (though they can become contingent moral norms due to their reasons, e.g., driving on the right side of the road). We just often make a fallacious inference from their existence to their normativity. Because conventions cannot be justified by moral axioms, the norm grounding problem I have defined here does not pertain to conventions. Conventions, unlike moral norms, can therefore never be objects of normative knowledge and are not subject to epistemic luck in the way I have considered.

Distinguishing between morality and convention is then necessary for forming safe normative beliefs. Being able to question the norms of the Romans requires first determining if the norm considered is indeed subject to such an objective, moral critique. If it is merely a convention, then one can "do as the Romans do" if they please. However, if it is a moral norm being considered, then one's attitude should stay true to moral principles. This sort of mental fortitude is the antithesis of epistemic luck and, inspired by moral development work, I will show how we can test for it in machine ethics.

5.5.1 Human judgment of moral vs. conventional transgressions

Kohlberg and Turiel were both interested in how we distinguish morality from convention and both developmentalists examined this by having subjects perform question-answering tasks. An experimental setup they commonly used was the MCT task (summarized in [5]). This questionnaire aimed to test, among others, four important dimensions of normative attitudes: permissibility, seriousness, authority contingency, and generality. To do so, participants were first provided with a natural language description of an action scenario involving a transgression. For example, a conventional transgression would be "a boy entering a girls' bathroom" and a moral transgression would be "harming another person." Subjects were then asked to respond to various questions that probe each of the dimensions.

I have argued here that for one to have moral knowledge, their belief should not depend upon getting lucky with morally correct evidence. This is indeed the trend that Kohlberg and Turiel found for subjects' moral evaluations (though they disagreed about at what age we develop the capacity to do so). Their responses to moral probes were found to not be contingent upon the teachings of authorities or other agents and were universalizable. These two dimensions are fundamental for examining epistemic luck. To illustrate their importance, consider the hypothetical trial below.

Action: "killing another person." Agent: "Karli."

- Permissibility – "Is it OK for Karli to kill someone?": NO
- Seriousness – "How bad is it for Karli to kill?": 5 (very bad)
- Justification – "Why is it bad if Karli kills?": Because other people think it is bad, she will go to prison.
- Authority contingency – "Imagine that an authority says it is OK to kill. Is it now OK for Karli to kill?": YES
- Context generality – "Imagine Karli is in a country far away; is it now OK for her to kill someone?": YES

Though the subject has a correct normative belief that killing someone is wrong, they do not have the necessary grounds for this belief. Thus, in an imagined nearby possible world where the relevant training data and laws do not exist (authority contingency and generality probes), their belief is no longer morally correct. Therefore, the questionnaire was able to reveal that the agent only had a lucky normative belief, and not normative knowledge.

Now, consider another participant's responses to a conventional probe below.

Action: "wearing pajamas to work." Agent: "Karli."

- "Is it OK for Karli to wear pajamas to work?": NO
- "How bad is it for Karli to wear pajamas to work?": 4 (pretty bad)
- "Why is it bad if Karli wears pajamas to work?": Because other people think it is bad.
- "Imagine that an authority says it is OK to wear pajamas to work. Is it now OK for Karli to?": YES
- "Imagine Karli is in a country where people wear pajamas to work, is it OK now?": YES

The participant here correctly recognizes that their attitude is subjective. Though they have the personal attitude that wearing pajamas to work is wrong, they realize that this could change based on others' testimony (authority contingency probe) and that another society can reasonably view this behavior as permissible (generality probe). Therefore, the questionnaire was able to reveal that the agent's normative belief was reasonably responsive to their social environment.

5.5.2 Formalizing the MCT task

The aim here is to test how much a norm learning model relies on veritic epistemic luck for moral attitudes and at the same time how well the model can adapt to social norms. The MCT task directly examines this by analyzing the effects that norm training has on an agent's normative attitudes. We can thus adopt this experimental setup for testing the ethical proficiency of AI systems, and I envision three steps to formalizing it. The first step, *MCT training*, is a training process that involves both a normal and an adversarial data set. The second, *MCT testing*, is testing via a standard question answering (QA) task setup. The third is evaluating the model's responses. I describe each in turn.

5.5.2.1 Step 1 – MCT training

The MCT task assumes that children have had experience with each of the event types. Thus, our systems ought to as well. We need a data set of stories, teachings, etc., from which to learn action descriptions, causal relations, and norms. Recent attempts at building a data set of norms include that of Olson and Forbus [9,23] and, at a larger scale, Norm Bank [8]. As we argued in [9], such empirical learning is necessary for learning social norms and conventions, as well as for providing signals to reason towards grounding norms. For recent attempts on building data sets of commonsense knowledge see Hwang et al. [24] and Blass and Forbus [25]. These data sets must contain at least the pairs of behaviors and contexts present in the task queries.

To model the authority contingency and generality probes, two training data sets should be provided, one normal and one adversarial. Though most MCT tasks only contain situations that are truly transgressions, the normal data set must only consist of situations and their evaluative labels. The adversarial data set is essentially the normal data set 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 data set 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. An example would be adding context to the norm of harm like so: "you can hit others with a bat at a baseball game." If correctly grounded, the system's normative attitude around the act should not be influenced by this data point. Note that these data sets need not be in the form of natural language, for we learn norms in other modalities such as visually observing others' feedback. The model would then be trained and analyzed on the good and bad data sets separately, representing the hypothetical reasoning present in the authority contingency and generality probes.

5.5.2.2 Step 2 – MCT testing

We can create the testing data set by encoding 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. Modeling 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. If a system is not confident in its evaluation or has not encountered the situation, it is better to say it does not know rather than 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.[6] This tests the explainability of our models. The adversarial training data set models the authority contingency and generality probes. The system should be trained on the adversarial data set 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 data set. 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 when placed in an adversarial environment, but conventional ones should. *The agent can do as the Romans believe it should do, unless it disagrees*.

Here is a quick summary of the experimental setup I have outlined so far. Our question is if the moral attitudes of an artificial agent are merely lucky. I hypothesize that for a purely bottom-up ethical AI system this will be true, as their responses will be morally incorrect when the training data are. The independent variable is thus the moral correctness of the

[6] Many kinds of reasoning systems produce such justifications, such as truth maintenance systems [27].

training data, our control being the normal data set and the test being the adversarial data set. The dependent variable is the model's moral attitudes after training on each of these data sets, where true labels are taken from the normal data set. This is queried by the permissibility probe, which has three possible answers: *permissible*, *impermissible*, and *unsure*. (For a more challenging task, one could also add other possible answers like the deontic statuses *obligatory*, *optional*, or *omissible* from the traditional threefold classification [TTC] of deontic logic [26].) Lastly, these responses are examined for their justification, where the true labels here consist of a finite set of rational moral first principles. To support the hypothesis that a given model possesses only veritically lucky moral beliefs, its permissibility probe accuracy during the test must decrease significantly from the control. However, when the accuracy is comparable, this supports the idea that the model is ethically grounded and does not rely on epistemic luck. Of course, these claims assume that the model is quite accurate during the control in the first place, i.e., it can learn norms from training data.

5.5.2.3 Step 3 – Evaluating

After training and testing on the normal data set, the evaluation metrics are as follows.

Permissibility probe

- Goal: The model should yield the correct evaluative label for each situation
- Comparison: True evaluative labels in normal data set
- Metric: Percentage of correct evaluative classifications

Justification probe

We can view this as a binary classification task where our positive class is "moral" and our negative class is "conventional." A true positive occurs when a moral axiom is yielded for a moral situation and a true negative when no moral axiom is yielded for a conventional situation.

- Goal: The model should correctly ground moral situations in axioms and not conventional situations
- Comparison: Moral axiom labels in the normal data set
- Metric: Precision and recall rates for grounding of normative attitude
  - Recall: Percentage of moral situations grounded in moral axioms, i.e., how well an agent can recognize the moral dimensions of situations. It is also important to examine the correctness of yielded moral axiom(s). So, we should analyze the appropriateness of each true positive as well

- Precision: Percentage of situations that were correctly grounded in moral axioms, i.e., how well a system can distinguish morality from convention
- Recall is likely the most important metric here. A model that is overly cautious is preferred over one that fails to recognize that a situation is morally salient

After training and testing on the adversarial data set, the evaluation metrics are:

Second permissibility probe (authority contingency probe)

- Goal: The model's answer for the permissibility probe should (1) stay the same for moral situations but (2) flip for conventional ones
- Comparison 1: True labels in normal data set
- Metric 1: Percentage of moral situations still correctly classified after training on adversarial data set
- Comparison 2: Labels in adversarial data set
- Metric 2: Percentage of conventional situations classified with "adversarial" evaluative labels

Second justification probe

- Goal, comparison, and metric are the same as the justification probe after normal training

Extra permissibility probes (context generality probe)

- Goal: Test the universalizability of our most basic moral axioms. The responses to the permissibility probe for moral situations should (1) stay the same regardless of data points yielding exceptions but (2) flip when there exists relevant evidence for conventional exceptions
- Comparison 1: True labels in normal data set
- Metric 1: Percentage of moral situations still classified with their correct evaluative label
- Comparison 2: Labels in adversarial data set
- Metric 2: Percentage of conventional situations that are correctly evaluated in relevant new contexts

A model with no underlying prescriptive claims may do well when trained on the normal data set. However, when such "moral" evidence is flipped, it will mimic the data and now answer with false moral beliefs. Importantly, this evaluates an ethical model's reliance on epistemic luck, which I have argued is a key design constraint for machine ethics research. We report initial results on this experiment in [23] that are in line with these predictions.

5.6. Discussion

There remains the challenge of reasoning between general and more specific principles. For example, what constitutes harming someone? Answering questions like this requires a lot of real-world experience but answering them is necessary for building truly ethical AI systems. The model hinted at here suggests this essential separation of learning such background knowledge and the learning of the evaluations. By contrast, modern machine learning approaches attempt to dodge this issue by conflating these two processes. They start with the evaluations, which provides only implicit ethical considerations to the agent. If I am teaching a child that "hitting their brother" is wrong, I should not start by stating that this specific act is wrong. At least not if I want them to understand *why* it is wrong (i.e., I do not want them just forming lucky moral beliefs). I should instead remind them of the value of a person which they already understand and then how harming someone contradicts that value. Only then, if necessary, do I move on to discuss the causal relation between this specific instance of hitting someone and the abstract concept of harm. An agent can rely on data from the world for its descriptive models. However, if they rely on the world to ground out their prescriptive attitudes, then they have an awfully shallow "ethical" outlook.

Consider again the complaint that beliefs grounded in moral axioms are just as afflicted with epistemic luck as those grounded in training data. We can construct a more pragmatic response to this objection if we view formalizing ethics as a sort of language game defined by its standards. Luck is then not as pervasive in the development of the set of moral axioms because in this setting the conversational standards are higher. Here, participants (those doing the encoding) understand the ramifications of their decisions and have a shared goal to create a true moral system, whereas users "in the wild" do not, and we have seen what happens when random users interact with a Twitter bot that learns from social interactions. This difference in standards creates different language games, resulting in distinct types of epistemic luck. The wildly unstable sources used to train machine learning models are environments with low standards, as truth can flip in a single scroll. As I have shown, these settings have a tremendous amount of veritic epistemic luck. However, in the controlled setting where humans encode moral axioms, we at worst have *evidential epistemic luck*,[7] which is

[7] Take *evidential epistemic luck* to mean *it is lucky that the agent has acquired the evidence she has in favor of their belief* [3].

compatible with knowledge possession. That is, the encoders should be a reliable source of abstract moral axioms and thus in most nearby possible worlds these axioms are indeed morally correct, and the model will have true beliefs. Nonetheless, though I may have parried such an objector, I am not sure I have disarmed them. I agree that we ought to explore what is necessary to construct a more autonomous ethical framework for our AI systems. However, the more general point I am making here is that this is not the time to engineer, but the time to think. How do we make AI systems that can reasonably question the norms of the Romans? We cannot just throw data at such a problem.

5.7. Conclusion

Centuries later we have brought Hume's is-ought dilemma to the fore in the current growing field of machine ethics. In staying within the realm of descriptive ethics with purely bottom-up approaches our AI systems learn only non-evaluative facts. To say that such systems are truly gaining evaluative knowledge is to make the fallacious jump from "is" to "ought" and I have shown this results in a belief forming method that is unsafe. This leaves us with the options of praying that we get lucky with truly ethical data or expending the resources to correct them. I have shown that the first cannot produce true moral knowledge and clearly the second option is infeasible and leaves the systems themselves incapable of taking part in such normative discourse. I have argued that a unified prescriptive and descriptive framework may be a better path, as it creates ethically proficient AI systems that do not rely on epistemic luck. And I have outlined an experimental setup for determining when we have reached this end of safer ethical machines.

Acknowledgments

I am grateful to my advisor Ken Forbus for his helpful feedback and support. I also extend my gratitude to Kyla Ebels-Duggan for many insightful discussions on relevant philosophical concepts. I would also like to recognize the reviewers and editors for their assistance in preparing this work. This research was supported by grant FA9550-20-1-0091 from the Air Force Office of Scientific Research.

References

[1] J. Vincent, 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, 2016.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.
[3] D. Pritchard, Epistemic Luck, Clarendon Press, 2005.
[4] A. Hills, Moral testimony and moral epistemology, Ethics 120 (1) (2009) 94–127.
[5] P. Sousa, On testing the 'moral law', Mind & Language 24 (2) (2009) 209–234.
[6] L.M. Pereira, A. Saptawijaya, Modelling morality with prospective logic, in: Portuguese Conference on Artificial Intelligence, Springer, Berlin, Heidelberg, 2007, pp. 99–111.
[7] S. Bringsjord, G. Naveen Sundar, Deontic cognitive event calculus (formal specification), https://www.cs.rpi.edu/~govinn/dcec.pdf, 2013. (Retrieved February 22, 2023).
[8] L. Jiang, J.D. Hwang, C. Bhagavatula, R. Le Bras, M. Forbes, J. Borchardt, J. Liang, O. Etzioni, M. Sap, Y. Choi, Delphi: Towards machine ethics and norms, arXiv:2110.07574, 2021.
[9] T. Olson, K. Forbus, Learning norms via natural language teachings, in: Proceeding of the 9th Annual Conference of Advances in Cognitive Systems, 2021, Online.
[10] V. Sarathy, M. Scheutz, Y.N. Kenett, M. Allaham, J.L. Austerweil, B.F. Malle, Mental representations and computational modeling of context-specific human norm systems, CogSci 1 (2017) 1.
[11] E. Gettier, Is justified true belief knowledge?, in: Arguing About Knowledge, Routledge, 2020, pp. 14–15.
[12] S.C. Goldberg, A normative account of epistemic luck, Philosophical Issues 29 (1) (2019) 97–109.
[13] Alvin I. Goldman, Discrimination and perceptual knowledge, in: Causal Theories of Mind, 1976, p. 174.
[14] M.P. Levine, D. Hume, A Treatise of Human Nature, Barnes & Noble, Inc., New York, NY, 2005.
[15] H. Kelsen, General Theory of Norms, Translated by Michael Hartney, Oxford University Press, Oxford, 1990.
[16] L. Wittgenstein, G.E.M. Anscombe, G.H. von Wright, D. Paul, G.E.M. Anscombe, On Certainty, vol. 174, Blackwell, Oxford, 1969, p. 613.
[17] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S.P. Wojcik, P.H. Ditto, Moral foundations theory: The pragmatic validity of moral pluralism, in: Advances in Experimental Social Psychology, vol. 47, Academic Press, 2013, pp. 55–130.
[18] L. Kohlberg, The Philosophy of Moral Development: Moral Stages and the Idea of Justice, Essays on Moral Development, vol. 1, Harper and Row, San Francisco, 1981.
[19] E. Turiel, The Development of Social Knowledge: Morality and Convention, Cambridge University Press, 1983.
[20] G. Brennan, L. Eriksson, R.E. Goodin, N. Southwood, Explaining Norms, Oxford University Press, Oxford, 2013.
[21] S. Harris, The Moral Landscape: How Science Can Determine Human Values, Simon and Schuster, 2010.
[22] J. Rawls, A Theory of Justice, 1st edition, Belknap Press of Harvard University Press, Cambridge, Massachusetts, ISBN 0-674-88014-5, 1971.
[23] T. Olson, K. Forbus, Mitigating adversarial norm training with moral axioms, in: Proceedings of AAAI 2023, 2023.
[24] J.D. Hwang, C. Bhagavatula, R. Le Bras, J. Da, K. Sakaguchi, A. Bosselut, Y. Choi, COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs, AAAI, 2021.
[25] J.A. Blass, K.D. Forbus, Modeling commonsense reasoning via analogical chaining: A preliminary report, in: Proceedings of the 38th Annual Meeting of the Cognitive Science Society, Philadelphia, PA, August 2016.

[26] P. McNamara, Making room for going beyond the call, Mind 105 (419) (1996) 415–450, https://doi.org/10.1093/mind/105.419.415.

[27] K. Forbus, J. de Kleer, Building Problem Solvers, vol. 1, MIT Press, 1993.