A developmentally-situated approach to teaching normative behavior to AI

Response to the EthicsNet Guardians’ Challenge

*By G Gordon Worley III*

# Introduction

Based on the usefulness of ImageNet, MovieLens, and other comprehensive datasets for machine learning, it seems reasonable that we might create an EthicsNet of ethical data we could use to train AI systems to behave ethically (Watson, 2018). Such a dataset would aid in addressing issues of AI safety, especially as they relate to AGI, since it appears learning human values will be a key component of aligning AI with human interests (Bostrom, 2014). Unfortunately, building a dataset for ethics is a bit more complicated than it is for images or movies because ethics is primarily learned by situated, embodied agents acting in the world and receiving feedback on those actions rather than by non-situated agents who learn about the world without understanding themselves to be part of it (Varela, 1999). Therefore we consider a way to fulfill the purpose of EthicsNet based on the idea that ethical knowledge is developmentally situated and so requires a generative procedure rather than a traditional dataset to train AI to adopt values and behave ethically.

# Ethics is developmentally situated

In philosophy the study of ethics quickly turns to metaethics because those are the sorts of questions that are of interest to philosophy, so it’s tempting to think that, based on the philosophical literature of ethics, learning to behave ethically (i.e. learning behavioral norms) is primarily about resolving ethical dilemmas and developing ethical theories that allow us to make consistent choices based on values. However, this would be to ignore the psychology of how people learn what behaviors are normative and apply those norms to engage in ethical reasoning (Peters, 1974). Rather than developing a coherent ethical framework from which to respond, humans learn ethics by first learning how to resolve particular ethical questions in particular ways, often without realizing they are engaged in ethical reasoning, and then generalizing until they come to ask question about what is universally ethical (Kohlberg, Levine, & Hewer, 1983).

This is to say that ethics is both situated in general—ethics is always about some agent deciding what to do within some context it is itself a part of—and situated developmentally—the context includes the present psychological development and behavioral capabilities of the agent. Thus to talk about providing data for AI systems to learn ethics we must consider what data makes sense given their developmental situation. For this reason we will now briefly consider the work of Kohlberg and Commons.

Kohlberg proposed a developmental model of ethical and moral reasoning correlated with general psychological development (Kohlberg & Hersh, 1977). We will summarize it here as saying that the way an agent reasons about ethics changes as it develops a more complex ontology so that young children, for example, reason about ethics in ways appropriate to their ability to understand the world and this results in categorically different reasoning and often different actions than that of older children, adolescents, adults, and older adults. Although Kohlberg focused on humans, Commons has argued that we can generalize developmental theories to other beings, and there is no special reason to think AI will be exceptional with regards to development of ontological and behavioral complexity, thus we should expect AI to experience psychological development (or something functionally analogous to it) and thus will develop in their moral reasoning as they learn and grow in complexity (Commons, 2006).

It’s within the context of developmentally situated ethics that we begin to reason about how AI systems can be taught to behave ethically. We might expect to train ethical behavior in AI systems the same way we teach them to recognize objects in images or extract features from text—viz. by providing a large data set with some predetermined solutions that we can train AI systems against—but this would be to believe that AI is exceptional and allows learning ethics in a way very different from the way both humans and non-human animals learn behavioral norms. Assuming AI systems are not exceptional in this regard, we consider a formulation of EthicsNet compatible with developmentally situated learning of ethical behavior.

# Outline for a generative EthicsNet

From a young age, human children actively seek to learn behavioral norms, often to the point of overfitting, by aggressively deriving “ought” from “is” (Schmidt et al., 2016). They do this based on a general social learning motivation to behave like conspecifics seen in both primates and other non-human animals (van de Waal, Borgeaud, Whiten, 2013), (Dingemanse et al., 2010). This strongly suggests that, within their developmental context, agents learn norms based on a strong, self-generated motivation to do so, thus foundational to to our proposal for teaching AI systems ethical behavior is a self-sustaining motivation to discover behavioral norms from examples. Other approaches may be possible, but the assumption of a situated, self-motivated agent agrees with how all agents known to learn normative behavior do so now, so we take up this assumption out of a desire to conserve uncertainty. Thus we will assume for the remainder of this work that such a motivation exists in AI systems to be trained against the EthicsNet dataset, although note that implementation of this motivation to learn norms strictly lies outside the scope of the EthicsNet proposal and will not be considered in detail here.

So given that we have a situated agentic AI that is self-motivated to learn normative behaviors, what data should be provided to it? Since the agent is situated it cannot, strictly speaking, be provided data of the sort that we normally think of when we think of datasets for AI systems. Instead, since the agent is to be engaged in actions that offer it the opportunity to observe, practice, and infer behavioral norms, it needs to be a dataset in the form of situations it can participate in. For humans and non-human animals this “dataset” is presented naturally through the course of living, but for AI systems the natural environment does not necessarily present such opportunities. Thus we propose that the goal of EthicsNet is to give a framework in which to generate such opportunities.

Specifically we suggest creating an environment where AI agents can interact with humans with the opportunity to observe and query humans about behavioral norms based on the agents’ behavior in the environment. We do not envision this as an environment like ReCAPTCHA where providing ethical information to AI systems via EthicsNet is the primary task in service of some secondary task (von Ahn, 2008). Instead, we expect EthicsNet to be secondary to some primary human-AI interaction that is inherently meaningful to the human since this is the same way normative behavior is learned in humans and non-human animals, viz. as a secondary activity to some primary activity.

By way of example, consider an AI system that serves as a personal assistant to humans that interacts with humans via a multi-modal interface (e.g. Siri, Google Assistant, Cortana, and Alexa). The primary purpose of the AI-human interaction is for the AI assistant to help the human with completing tasks and finding information they might otherwise have neglected. As the AI assistant and human interact, the human will demonstrate behaviors that will give the AI assistant an opportunity to observe and infer behavioral norms based on the way the human interacts with it. Further, the AI assistant will take actions, and about those actions the human may like what the AI assistant did or may prefer it did something else. We see the goal of EthicsNet as providing a way for the human to provide the AI assistant in this scenario feedback about those likes and preferences so the AI assistant can use the information to further its learning of behavioral norms.

# Caveats of a generative EthicsNet

As mentioned, ethical learning is developmentally situated, so this means that feedback from guardian humans to learning AI systems should differ depending on how complexly an AI system models the world. Explaining by way of example, consider that young children are often presented corrections on their behavior to get them to conform to norms in ways that focus on categorizing actions as right and wrong. A simple example might be telling a child to *always* hold hands while cross the street and to *never* hit another child. Such an approach, of course, leaves out many nuances of normative behavior adults would consider, as in some cases a serious threat may mean a child should risk crossing the street unattended or hitting another child in defense. The analogous cases for AI systems will of course be different, but the general point of presenting developmentally appropriate information holds, such as eliding nuances of norms for children that adults would normally consider.

In order to ensure developmentally appropriate feedback is given, it’s important to give contextual clues to humans about the AI system’s degree of development. For example, we might want to give clues that the human should treat the AI the same way it would treat a child if that were developmentally appropriate, or treat them as an adult if that were developmentally appropriate. Experimentation will be necessary to find the cues that encourage humans to give developmentally appropriate feedback, so EthicsNet will need to be able to a provide rapidly iterable interface to allow developers to find the best user experience for eliciting maximally useful responses from humans for helping AI systems learn normative behaviors.

Since EthicsNet, as proposed here, is to be a secondary function to an AI system serving some other primary function, an implementation difficulty is that it must be integrated with a system providing the primary functionality. This will likely involve forming partnerships with leading AI companies to integrate EthicsNet into their products and services. This is more complicated than if EthicsNet could be developed in isolation, but we believe for reasons laid out above that it cannot, so this added complexity is necessary. For similar reasons this will make development of EthicsNet more complicated since it will require integration with one or more existing systems owned by other organizations in order to allow EthicsNet to get feedback from humans serving the guardian role to AI systems, but we believe the additional cost and complexity is worthwhile since something short of this seems unlikely to succeed at the task of teaching AI systems to behave ethically based on what we know about how normative behaviors are learned in humans and non-human animals.

Given this context in which EthicsNet will be deployed, it will also be important to make sure to choose partners that enable AI systems being trained through EthicsNet to learn from humans from multiple cultures since different cultures have differing behavioral norms. Note, though, that this will also make it harder for the AI systems being trained to infer what behavior is normative because they will receive conflicting opinions from different guardians. How to resolve such normative uncertainty is an open question in ethics, so EthicsNet may prove vital in research to discover how, in an applied setting, to address conflicts over behavioral norms (MacAskill, 2014).

# Conclusion

The view of EthicsNet we have presented here is not one of a typical dataset for machine learning like ImageNet but rather as a framework in which AI systems can interact with humans who serve as guardians and provide feedback on behavioral norms. Based on the situated—particularly the developmentally situated—nature of ethical learning, we believe this to be the best approach possible and that a more traditional dataset approach will come up short towards fulfilling the goal of enabling AI systems to learn to act ethically. Although this approach offers less opportunity for rapid training since it requires interaction with humans on human timescales and requires integration with other systems since ethical learning is a secondary activity to some other primary activity, the outcome of producing AI systems that can conform to human interests via ethical behavior makes it worth the additional effort.

# References

L. von Ahn, B. Maurer, C. McMillen, D. Abraham, M. Blum. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*. 321 (5895): 1465–1468. DOI:[10.1126/science.1160379](https://doi.org/10.1126/science.1160379)

N. Bostrom. (2013). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

M. L. Commons. (2006). [Measuring an Approximate g in Animals and People](https://www.dareassociation.org/documents/Commons,%20Measuring%20an%20Approximate%20g%202006.pdf). *Integral Review*, *3*, 82-99.

N. J. Dingemanse, A. J.N. Kazem, D. Réale, J. Wright. (2010). Behavioural reaction norms: animal personality meets individual plasticity. *Trends in Ecology & Evolution,* Volume 25, Issue 2, Pages 81-89, DOI: [10.1016/j.tree.2009.07.013](https://doi.org/10.1016/j.tree.2009.07.013).

L. Kohlberg & R. H. Hersh. (1977). Moral development: A review of the theory. *Theory Into Practice*, 16:2,53-59. DOI: [10.1080/00405847709542675](https://doi.org/10.1080/00405847709542675).

L. Kohlberg, C. Levine, & A. Hewer. (1983). Moral stages: A current formulation and a response to critics. *Contributions to Human Development, 10,* 174.

W. MacAskill. (2014). *Normative Uncertainty*. Dissertation, University of Oxford.

R. S. Peters. (1974). *Psychology and ethical development. A collection of articles on psychological theories, ethical development and human understanding*. George Allen & Unwin, London.

M. F. H. Schmidt, L. P. Butler, J. Heinz, M. Tomasello. (2016). Young Children See a Single Action and Infer a Social Norm: Promiscuous Normativity in 3-Year-Olds. *Psychological Science*. DOI: [10.1177/0956797616661182](http://dx.doi.org/10.1177/0956797616661182).

F. J. Varela (1999). *Ethical know-how: Action, wisdom, and cognition*. Stanford University Press.

E. van de Waal, C. Borgeaud, A. Whiten. (2013). Potent Social Learning and Conformity Shape a Wild Primate’s Foraging Decisions. *Science*, 340 (6131): 483-485. DOI: [10.1126/science.1232769](https://doi.org/10.1126/science.1232769).

N. Watson. (2018). EthicsNet Overivew. URL: <https://www.herox.com/EthicsNet>