From Bench to Bedside: Implementing AI Ethics as Policies for AI Trustworthiness

Jeffrey M. Bradshaw[1], Larry Bunch[1], Michael Prietula[1,2], Edward Queen[2], Andrzej Uszok[1], Kristen Brent Venable[1,3]

[1]Institute for Human and Machine Cognition
[2]Emory University
[3]University of West Florida
jbradshaw@ihmc.org, lbunch@ihmc.org, mj.prietula@emory.edu, equeen@emory.edu, auszok@ihmc.org, bvenable@ihmc.org

Abstract

It is well known that successful human-AI collaboration depends on the perceived trustworthiness of the AI. We argue that a key to securing trust in such collaborations is ensuring that the AI competently addresses ethics' foundational role in engagements. Specifically, developers need to identify, address, and implement mechanisms for accommodating ethical components of AI choices. We propose an approach that instantiates ethics semantically as ontology-based moral policies. To accommodate the wide variation and interpretation of ethics, we capture such variations into ethics sets, which are situationally specific aggregations of relevant moral policies. We are extending our ontology-based policy management systems with new representations and capabilities to allow trustworthy AI-human ethical collaborative behavior. Moreover, we believe that such AI-human ethical encounters demand that trustworthiness is bi-directional – humans need to be able to assess and calibrate their actions to be consistent with the trustworthiness of AI in a given context, and AIs need to be able to do the same with respect to humans.

Overview and Objectives

Today, widespread calls and directives exist that AI should be ethical. However, there is a pressing need for more clarity and guidance on how AI could be ethical. Concerns about AI trust grow as AI assumes increasingly important roles woven into the fabric of our society. Public, business, and governmental organizations, both domestic and international, struggle to articulate guidelines to address this concern as highly visible reports emerge on the potential or demonstrated "unethicality" of these technologies, eroding the potential for trust (Bonnell, 2023).

Specific gaps in these efforts highlight the complexity of the ethical challenges in AI. First, defining general moral statements that lead to appropriate ethical choices in any context is difficult in and of itself. Second, it is challenging to accommodate cultural, institutional, or situational differences that impact ethical choices derivable from those moral statements. Finally, it is hard to implement them computationally, weighing derived options against one another in a fashion that affords trust in the process. Of course, we are all aware of the tidal wave of asserted concerns by institutions, governments, and organizations worldwide, often taking the form of desired statements of ethical guidelines – what AI should or should not be or do.

The progress of AI-human collaboration necessitates ethical reasoning components that attend to those concerns. Interesting work is emerging on assessing and managing expectations of ethics and trust in AI (Kinney et al. 2024), but there also needs to be guidance on designing AI that specifically addresses implementing ethical reasoning components at the core of AI choice. Our long-term objective is to improve the quality of autonomous and collaborative human-machine ethical decision-making by leveraging the strengths of each to mitigate the weaknesses, biases, and experiences of the other. Our approach is to develop and demonstrate that trusted automated and collaborative ethical decision support – in the form of a policy-based, semantically-formed, ethical architecture capable of explanation and learning – can be successfully applied to complex real-world domains.

The Global Call for AI Ethics

In 2016, the White House, the European Parliament, and the United Kingdom's House of Commons released a report on AI's social, ethical, and economic impact. Cath et al. (2018) analyzed each report but noted that the U.S. recommendations were unique in stating that innovation experimentation should be encouraged, with research (including basic long-

term research) informing regulatory policy. Since then, calls for AI ethics have escalated. The 193 member states support UNESCO's 2021 Recommendation on the Ethics of Artificial Intelligence (UNESCO 2021), which was also adopted by all G20 members. In addition to the European Union, all 38 members and 8 non-member countries have adopted OECD Recommendations of the Council on Artificial Intelligence (OECD 2024). The European Union's AI Act (EU-AIACT 2024), implemented in August, is the world's first all-encompassing framework specifically on AI. This global consensus of intent underscores the universal concern of AI ethics writ large. Yet, there is uncertainty about moral matters (Macaskill, Bykvist, and Ord 2020), and it is unlikely that any invariant and universal set of ethics-morals will be agreed upon (Moore 2006). Consequently, we emphasize the need to define an architecture within which specific contextual sets of ethics (defined as moral policies) can be manually defined or autonomously and situationally learned, examined, and implemented, taking the preferences of specific stakeholders into account.

We want to stress an often-neglected step in this process: implementation. Fortunately, implementation issues for AI ethics have not gone unnoticed. Efforts range from specific algorithmic implementations of proposed moral stances (Leben 2019) to broader, encompassing discussions of defining code for AI ethics (Boddington 2017) to the ethics of AI developers (Griffin, Green, and Welie, J. 2024). Arkin (1998) proposed a "moral faculty" as a component of a general architecture that functioned as an ethical governor in autonomous weapons. In that design, constructs such as "guilt" and "emotion" were included. This raises the recurring issue of the risks of modeling poorly understood (and variously defined) components of human decision-making. It should be noted that there is a difference between modeling ethical deliberation in a way that takes human considerations into account vs. as a process that slavishly follows how humans deliberate ethics — explicitly or implicitly (Howard and Borenstein 2018). Put bluntly, should we design aircraft based on how birds have adapted to the physics of flight, or is it better to construct artifacts that directly adapt to the physics of flight?

Whatever approach we take should ultimately result in a design that assures internal coherence to these ethical sets. As technology evolves, it remains uncertain whether these sets can be considered "truly" human-ethical or something different (Arkin 2018). The complexities of the environment—social, legal, and cultural—indicate that no single moral theory will be sufficient (Pagallo 2017).

It is one thing to determine whether an autonomous AI agent "behaves" in a way consistent with a specific moral stance, but quite another to ascribe moral agency to an AI agent taking that stance (Gunkel 2012). The latter tack necessitates non-trivial complications regarding AI and legal theory (Abhivardhan 2019). Thus, an architecture that is agnostic to and capable of implementing and applying the broadest possible range of context and role-dependent moral stances is needed.

Addressing these issues, we rely on AI to ensure conformance to sets of conflicting and complex ethical imperatives. Humans play a crucial role in ethical decision-making by determining when to make exceptions for nuanced and/or consequential circumstances. Our experience to date leads us to believe that in such situations, humans and machines can perform better working together than either could do alone and that the general approach proposed can be applied, in whole or in part, to a wide range of AI ethical contexts. We will now discuss three key innovation challenges in this project.

Three Challenges

The first challenge is *to extend semantically rich mediating representations for diverse ethical theories and their eventual instantiated applications*. The formal representation of ethics needs to be semantically rich to sufficiently represent the universe of competing ethical theories and their moral interpretation of laws, rules, and codes as policies across widely different application domains. The semantic-free perspective of large language models can be insufficient in theory and practice, and the consequences of those deficiencies are now widely evidenced (Burtsev, Reeves, and Job 2023; Mitchell and Krakauer 2023). Today's formal and informal reports on "AI ethics" reveal a diversity of definitions. Therefore, the core concept of ethics must remain flexible and tailored to the moral policies within our specific ethical framework, noting that the interpretation of ethics as instantiated computationally as moral policies was discussed initially thirty years ago by Moore (1995).

In addition, to enable human-AI collaboration, formalisms must function in a mediating role—representing the moral policies in a way that is both understandable and explainable to humans and computationally efficient and expressive for automated learning and reasoning. As an example of the backbone for such ontological representations, consider the W3C Web Ontology Language (OWL) standard (OWL2 2012). This approach enables incremental, principled extension of actions, actors, and situations (including spatial ontologies) to be expressed within and governed by policies. For example, we have demonstrated the expressive power of ontologies in representing policies in multiple domains (Bradshaw, Montanari, and Uszok 2014), and related work has similarly extended OWL to represent legislation (Ceci et al. 2016; Palmirani et al., 2021). We assert that the elemental components of ethics can be captured and represented as moral policies. Such ontology-based policies

would be both computable and human-understandable to allow, forbid, obligate, or waive the performance of actions in dynamically assessed contexts. We are developing and testing new extensions to these previous representational efforts to capture the varying nuances of ethics implemented as moral policies across several critical decision contexts. OWL contains the necessary constructors for formally describing the rich and diverse contexts needed to represent moral policies.

The second challenge is *to refine and resolve moral policy reasoning and computational mechanisms using those implemented representations*. An ontology-based representation is not sufficient on its own. One also needs a rule engine to decide these policies (Ceci and Aldo 2016). Representational ontologies exploit the power of description logic to perform complex reasoning tasks on large-scale problems with high efficiency. Additional mechanisms, such as AI-based preference reasoning, must be integrated to relate policies and establish priority among conflicting moral policies. These mechanisms should also include handling uncertainties, under-specifications, and other relevant forms of conflict. This use of ontologies permits extending the instantiated moral reasoning framework by adding actions, actors, and related concepts from new domains. A description logic reasoner operating over OWL is a pivotal contributor to deciding the applicable policies for a given action in context. Furthermore, the integration of contextually specific reasoners affords the unique and necessary components to simulate the policy consequences of hypothetical changes within and between ethics sets to enable the environment to report deliberations, histories, and explanations that human users can understand.

The third and most significant challenge is *to understand and design the human-AI engagement needed to collaboratively decide which moral policies apply in each context, anticipate policy violations, and guide humans and AI to ethical actions.* Central to our approach is developing an interactive policy explanation capability for AI that is aware of human input at multiple levels. This capability enables AI to learn from and exchange knowledge about ethical decisions with humans and other AI agents. The AI policy framework must be able to govern the human-AI system, determining when AI can act autonomously, when human oversight is necessary, when humans can act independently, and when both humans and AI need to collaborate for adequate assessment of the context and policies in making ethical decisions. To do so, humans need to be able to assess and calibrate their actions to be consistent with the trustworthiness of AI in a given context, and AIs need to be able to do the same with respect to humans.

Building on KAoS

Leveraging our experience in research, development, and use of a mature general-purpose policy management system built upon OWL (Bradshaw et al. 2014), we are adding new representations and reasoning mechanisms suited for enumerating and enforcing the multiple interrelated layers of rules, laws, regulations, and social norms that frame ethical decision-making. KAoS supports authorization policies that either allow or forbid an action, as well as obligation policies that either require or waive the requirement of an action in a given context.

The breadth of KAoS semantics has allowed its use across multiple application domains and operating environments. It has well-defined interfaces for programmatic access to functionality and the ability to import and export OWL ontologies. Several government agencies and private organizations have sponsored research and development efforts to mature KAoS for scalability, more powerful and flexible reasoning, and for use within distributed enterprises (Bradshaw et al. 2003; Uszok et al. 2011).

KAoS uses a logical precedence mechanism as an alternative to numeric priorities (Bradshaw et al. 2014). This allows administrators to specify an almost infinite variety of precedence relationships among moral policies, mirroring the kinds of rationale that people use when deciding which policy will dominate another. KAoS has provided rich experience in policy representation (via our policy representation language). It also provided insight into building the basics of reasoning about policies. But KAoS does not yet have the ability to learn new policies based on experience nor to link policies with values or the capability of representing and reasoning values or risks. What is needed is a specific rule engine to decide policies, and OWL-2 logical reasoners alone are insufficient (e.g., Ceci and Aldo 2016). Consequently, we are electing to design components based on open emerging standards, such as LegalRuleML (Palmirani et al. 2021), with related reasons like SPINdle (Lam and Governatori 2009) and Houdini (Cristani et al. 2023).).

Building on Ethics

Accordingly, and in conclusion, we have defined five core requirements for this project regarding ethics-related capabilities supporting AI trustworthiness. Specifically, an implemented AI agent must be able to

...know the laws, rules, norms, and other types of policies that forbid, allow, oblige, and waive certain actions in specified contexts.

...determine whether an action would violate one or more policies before taking the action.

...decide when to take an action that violates a policy for ethical reasons.

...explain the decision to take or forestall actions based on policies, values, outcomes, and goods.

...share own knowledge and learn from others' contextual decisions and outcomes (human and AI).

References

Abhivardhan. 2019. *Artificial intelligence ethics and international law*. New Delhi, India: BPB Publications.

Arkin, R. 1998. *Behavior-based robots.* Cambridge, MA: MIT Press.

Arkin, R. 2018. Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE* 100(3): 571-589. doi: 10.1109/JPROC.2011.2173265

Boddington, P. 2017. *Towards a code of ethics for artificial intelligence.* Switzerland: Springer Nature.

Bonnell, A. 2023. Navigating the Paradox: Restoring Trust in an Era of AI and Distrust. https://napawash.org/standing-panel-blog/navigating-the-paradox-restoring-trust-in-an-era-of-ai-and-distrust. Accessed: 2024-7-24

Bradshaw, J.; et al. 2003. Representation and reasoning for DAML-based policy and domain services in KAoS and nomads. *Proceedings of the second international joint conference on Autonomous agents and multiagent systems* (AAMAS '03). New York, NY. doi:10.1145/860575.86070

Bradshaw, J.; Montanari, R.; and Uszok, A. 2014. Policy-based governance of complex distributed systems: What past trends can teach us about future requirements. *Adaptive, Dynamic, and Resilient Systems*. Auerbach Publications. 259-284.

Burtsev, M.; Reeves, M.; and Job, A. 2023. The Working Limitations of Large Language Models. *MIT Sloan Management Review* Winter.

Cath, C. et al. 2018. Artificial Intelligence and the 'Good Society':The US, EU, and UK approach. *Science and Engineering Ethics* 24:505-528.doi:10.1007/s11948-017-9901-7

Cristani, M. et al. 2023. The architecture of a reasoning system for Defeasible Deontic Logic. *Procedia Computer Science* 225: 4214-4224. doi:10.1016/j.procs.2023.10.418

EUAIACT 2024. *The EU Artificial Intelligence Act*. http://data.europa.eu/eli/reg/2024/1689/oj Accessed: 2024-8-19

Griffin, T.; Green, P.; and Welie, J. 2024. The ethical agency of AI developers. *AI and Ethics* 4:179-188. doi: 0.1007/s43681-022-00256-3

Gunkel, D. 2012. *The machine question: Critical perspectives on AI, robots, and ethics.* Cambridge, MA: MIT Press.

Howard, A.; and Borenstein, J. 2018. The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics* 24: 1521-1536. doi: 10.1007/s11948-017-9975-2

Kinney, M. et al. 2024. Expectation management in AI: A framework for understanding stakeholder trust and acceptance of artificial intelligence systems. *Heliyon* 10(7). doi: 10.1016%2Fj.heliyon.2024.e28562

Lam, HP; and Governatori, G. 2009. The making of SPINdle, In *Rule Representation, Interchange and Reasoning on the Web*, edited by G. Governatori, J. Hall, and A. Paschke, 315-322. Berlin: Springer-Verlag.

Leben, D. 2019. *Ethics for robots: How to design a moral algorithm*. New York, NY: Routledge.

Macaskill, M.; Bykvist, K.; and Ord, T. 2020. *Moral Uncertainty*. Oxford: Oxford University Press.

Mitchell, M.; and Krakauer, D. 2023. The debate over understanding AI's large language models. *PNAS* 120(13): e2215907120. doi:10.1073/pnas.2215907120

Moore, J. 1995. Is ethics computable? *Metaphilosophy* 26: 1-31. doi: 10.1111/j.1467-9973.1995.tb00553.x

Moore, J. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems Jul*/Aug: 18-21. doi: 10.1109/MIS.2006.80

OECD. 2024. Principles for Trustworthy AI. https://oecd.ai/en/ai-principles Accessed: 2024-7-28.

OWL2. 2012. Web Ontology Language: Document Overview (Second Edition). World Wide Web Consortium (W3C). https://www.w3.org/TR/owl2-overview Accessed: 2024-7-22**.**

Pagallo, U. 2017. When morals ain't enough: Robots, ethics, and the rules of the law. *Minds & Machines* 27: 625-638. doi: 10.1007/s11023-017-9418-5

Palminani, M. et al. 2021. *LegalRuleML Core Specification Version 1.0*. https://docs.oasis-open.org/legalruleml/legalruleml-core-spec/v1.0/os/legalruleml-core-spec-v1.0-os Accessed: 2024-7-22

Shumailov, I. et al. 2024. AI models collapse when trained on recursively generated data. *Nature* 631: 75-759. doi: 10.1038/s41586-024-07566-y

UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence. https://www.unesco.org/en/artificial-intelligence. Accessed: 2024-7-24.

Uszok, A., et al. 2011. Toward a flexible ontology-based policy approach for network operations using the KAoS framework. *MILCOM 2011 Military Communications Conference*: 1108-1114. doi: 10.1109/MILCOM.2011.6127447