DOI:10.1145/3633455

Nicholas Berente, Cameron Kormylo, and Christoph Rosenkranz

Opinion

Test-Driven Ethics for Machine Learning

Encouraging organizations to adapt a test-driven ethical development approach.

MACHINE LEARNING (ML) applications and the organizations that develop them should be accountable. Proposed regulations require impact assessment and there are calls to strengthen enforcement of regulations for ethical business practice regulations.[a] Responsible organizations should implement a "test-driven ethics" development approach rooted in pragmatist discourse ethics and lessons from test-driven development.

This approach extends the popular "principles" approach to ethics seen in industry, government, and the academy.[2] Adopting ethical principles will not guarantee ethical actions or outcomes. Principles make values clear, but they are difficult to apply, vary in levels of abstraction, and require judgment when choosing among operationalizations.[9] For example, the principle of fairness is good, but assessing fairness requires operationalization. Common operationalizations (for example, "equalized odds" and "demographic parity") can contradict each other.[b] In sum, *ethical principles espouse values*, but are poor at guiding ML application development.

IMAGE CREATED BY ANDRIJ BORYS ASSOCIATES WITH THE ASSISTANCE OF SHUTTERSTOCK.AI

Principles describe "final" goals[15]—the effect of ML applications on the world. But ML's effects in social systems cannot be entirely anticipated.[19] Simon[15] advocated for a pragmatic, procedural approach generating intermediate goals to set initial conditions for development in complex situations. For ethical ML, this involves operationalizing desirable principles, then continually testing to see if the outcomes align with these principles. *Treat each development iteration as a hypothesis about ethical implications of the application, and then subject this hypothesis to strict tests, while being open to unforeseen outcomes.* Treating each iteration as a working hypothesis is key in pragmatist or "fallibilist" approach. Fallible humans cannot foresee the future; they must conduct complex development activities with humility, monitoring and adapting using tentative goals. This makes it possible to adapt toward better outcomes because every iteration is a tested hypothesis.

In discourse ethics, this testing process involves communication and discourse among stakeholders.[8] Developers, managers, employees, shareholders, customers, users, and others are stakeholders with different perspec-

[a] For an argument for accountability, see Martin.[5] See Metcalf et al.[7] and Slaughter et al.[16] for examples of regulations.

[b] For examples of different operationalizations of fairness, see Mehrabi et al.[6] For an assessment of the contradictions and how to deal with them, see Teodorescu et al. 2021.[17]

tives on ethics. Situations change, and unforeseen ethical issues arise. Discourse ethics anticipates this, involving continuous communication among stakeholders without assuming a correct answer. Different voices come together to continually assess and weigh alternative actions. There is no ideal solution to complex ethical problems, but through communication things move toward better outcomes. Communication should encourage constructive, deliberative discourse. Applying discourse ethics in developing ML applications extends the paradigm of test-driven development toward ethical directions, but it takes significant resources and commitment to implement test-driven ethics. We sketch out a procedure for test-driven ethics that manages these costs.

Test-Driven Development and an Ethical Test Procedure

Organizations need procedures that anticipate ethical issues and guide development and systems for resilience of an ML application in use. A test-driven development process for ethics is an intentional and continual audit that reflects on the process, scans for change, and updates judgements. This means treating application implementation as the start of a continual process of developing tests for different scenarios and reconciling inconsistencies as new requirements, industry best practices, and new goals are identified.

A test-driven ethics procedure would involve three activities at each iteration: Generate tests with multiple operationalizations for each focal ethical principle; set conditions for a deliberative discourse; and audit the process for changes and new information.

Generate tests. Test-driven development (such as in DevOps approaches) uses a test library to continuously improve software quality.[1] Ethically oriented test-driven development would develop a test library for ethical standards, such as fairness, privacy, interpretability, bias, contestability, and so forth. Addressing all ethical issues at once would be daunting, so ML development teams could proceed incrementally—initially focusing on one principle and expanding with each iteration.

Test-driven ethical discourse requires multiple tests for each focal ethical dimension. Ethical principles cannot be measured with one metric or testing strategy. For example, a credit scoring system's fairness could be operationalized as demographic parity of gender, improving conditions for the least fortunate, data acquired without deception or fraud, or just use. The "Fairness Indicators" library from TensorFlow[c] calculates common metrics and scales for large datasets and models for each of these operationalizations. Further, the Aequitas Python toolkit provides a "Fairness Tree" that allows developers to determine the most appropriate metric(s) to utilize in their context and subsequently conduct audits across multiple metrics.[11]

Using the Synthetic Data Vault (SDV) library,[10] tests could utilize synthetically generated data, custom-built to the organization's use case, to test a model's performance across different demographics or user groups. Tests could build on practices such as adversarial testing and using pre-trained datasets designed to identify bias such as StereoSet.[d] The Alibi library[4] can be used to generate high-quality counterfactuals for an organization's ML to aide with model interpretation.[18]

Tests can also extend beyond quantitative analyses. Ethnographic methods can be crucial in identifying how AI systems are used by individual actors and larger institutions, allowing a more thorough understanding of their societal impact. Together, these libraries, frameworks, and datasets provide guidelines, tools, and approaches; the key is to include multiple tests for each ethical goal.

General goals and principles can motivate development, but organizations must operationalize each principle through multiple tests. Organizational stakeholders would determine the ethical tests for a given decision context and for their current intermediate goals. This differs from simple requirements testing because the aim is to generate discourse through discussion of their results.

Deliberative discourse. Testing is not a panacea. Testing is the catalyst for discourse among stakeholders and informs those discussions. The array of tests will likely identify unforeseen

[c] See https://bit.ly/3TPMots
[d] See https://bit.ly/3vrzVTg

Starting small is a good idea, and development teams should experiment, learn, and build on successes.

ethical violations and provide mixed results. This will spark generativity in identifying improvements. Effective discourse ethics relies on hearing from all stakeholders, including marginalized groups, in designing and testing requirements for any ML applications so that powerful stakeholder interests do not dominate. Diverse voices need to be incorporated to reflect differing priorities. Engaging unheard voices in discussion of the ethical test results during the application development process can establish a shared language and common ground and enable a defensible compromise at each iteration. Deliberative discourse is difficult and requires procedures involving civility; rationality; and reflective, communicative action. Development teams need to learn to deliberate with a variety of stakeholders and should start small. Deliberation systems can help.

Ethical standards may also change over time. Goals could shift from demographic parity to equalized odds, definitions of protected groups can be adjusted, or a conflicting goal could become more salient. For a given use, organizations must understand the important ethical concepts, and their developers must keep these in mind as they design the application and the testing for subsequent iterations.

Continual audit. After release, the organization documents decisions; reflects on how well the process worked; and scans for new ethical requirements test libraries and datasets, or contextual changes to inform subsequent improvements. Audits using the test suite can verify that ML applications comply with corporate policies, industry standards, and regulations. Organizations also need to observe the use and outcomes of the ML application and report relevant metrics.

The test-driven ethics approach requires that organizations generate and apply tests on ethical issues for each iteration, discuss test results with diverse stakeholders, and audit the process through reflection and integrating new information. Starting small is a good idea, and development teams should experiment, learn, and build on successes.

Discussion

Development organizations have been advised to incorporate an "ethics first" approach,[2] to follow best practices and norms,[12] and include frequent testing.[13] This does not tell developers how to incorporate ethics first, what norms to follow, and what developers should be testing. Fortunately, studies of discourse ethics during development provide some guidance. Value-levers open up conversations about application design decisions.[14] There are approaches for generating deliberative discourse around datasets used in practice (for example, Gebru et al.[3]). Reflective design, participatory design, or value-centered design can be leveraged to understand how ML applications can be designed for resilience with continual ethical testing as a design goal.

We advocate for a test-driven ethical development approach rooted in pragmatist discourse ethics. We encourage organizations to construct and share open source ethical test libraries, datasets, and best practices to advance this approach and harness the power of ML to contribute to a better world.

The implementation of this approach introduces significant costs to the development process. Espousing principles is easy, but operationalizing, monitoring, and auditing outcomes in line with those principles is not easy, nor is it cheap. Conducting an array of tests introduces an array of non-functional requirements that add burden to the development environment. Discourse among participants can be time consuming and require extensive coordination. Additional costs are unavoidable, but the alternative is to fail to effectively engage ethical issues in ML and later be held accountable. Test-driven development has added burden to developers, yet they have adapted, and they can do it again. At the Notre Dame-IBM Technology Ethics Lab, we are developing an appropriate test-driven ethics and auditing process, set of standards, and curriculum integrating both quantitative and qualitative auditing methods to make this idea a reality, while mitigating some of the associated costs. We invite others to help. Together, we can drive down the costs of implementation to ensure some ethical checks on ML advancements.

References

1. Beck, K. *Test-Driven Development: By Example.* Addison-Wesley Professional, 2003.
2. Floridi, L. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence 1*, 6 (June 2019).
3. Gebru, T. et al. Datasheets for datasets. *Commun. ACM 64*, 12 (Dec. 2021).
4. Klaise, J. et al. Alibi explain: Algorithms for explaining machine learning models. *J. Machine Learning Research 22*, 1 (Jan. 2021).
5. Martin, K. Ethical implications and accountability of algorithms. *J. Business Ethics 160*, 4 (Apr. 2019).
6. Mehrabi, N. et al. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR) 54*, 6 (June 2021), 1–35.
7. Metcalf, J., Smith, B., and Moss, E. New proposed law could actually hold big tech accountable for its algorithms. *Slate* (2022); bit.ly/3Vah75w.
8. Mingers, J. and Walsham, G. Toward ethical information systems: The contribution of discourse ethics. *MIS Quarterly* (2010).
9. Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence 1*, 11 (Nov. 2019).
10. Patki, N., Wedge, R., and Veeramachaneni, K. The synthetic data vault. In *Proceedings of the IEEE Int. Conf. Data Science and Advanced Analytics (DSAA)*, 2016.
11. Saleiro, P. et al. Aequitas: A bias and fairness audit toolkit, 2018; arXiv preprint arXiv:1811.05577.
12. Salge, C.A. and Berente, N. Is that social bot behaving unethically? *Commun. ACM 60*, 9 (Sept. 2017).
13. Schneiderman, B. Responsible AI: Bridging from ethics to practice. *Commun. ACM. 64*, 8 (Aug. 2021).
14. Shilton, K. Values levers: Building ethics into design. *Science, Technology, and Human Values 38*, 3 (Mar. 2013).
15. Simon, H. *Sciences of the Artificial.* Third Edition, MIT Press (1996).
16. Slaughter, R.K., Kopec, J., and Batal, M. Algorithms and economic justice: A taxonomy of harms and a path forward for the Federal Trade Commission. *Yale J. Law and Technology 23*, (2020).
17. Teodorescu, M.H. et al. Failures of fairness in automation require a deeper understanding of human-ML augmentation. *Management Information Systems Quarterly 45*, 3 (Mar. 2021).
18. Watcher, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard J. Law and Technology 31*, 2 (Feb. 2018).
19. Winter, S.J. and Butler, B.S. Responsible technology design: Conversations for success. *Perspectives on Digital Humanism.* C. Ghezzi et al., Eds. Springer Nature (2021).

Nicholas Berente (nberente@nd.edu) is a professor of IT, Analytics, and Operations at the University of Notre Dame, Mendoza College of Business Notre Dame, IN, USA.

Cameron Kormylo (ckormylo@nd.edu) is a research associate in IT, Analytics, and Operations at the University of Notre Dame, Mendoza College of Business Notre Dame, IN, USA.

Christoph Rosenkranz (rosenkranz@wiso.uni-koeln.de) is a professor of Information Systems at the University of Cologne, Koln, Nordrhein-Westfalen, Germany.