AI Ethics Journal

Vol 1, Issue 1
Fall 2020

Opinion

What a Philosopher Learned at an AI Ethics Evaluation

James Brusseau, Ph.D. [1]

[1] *Department of Philosophy, Pace University*

DOI: https://doi.org/10.47289/AIEJ20201214

Abstract

AI ethics increasingly focuses on converting abstract principles into practical action. This case study documents nine lessons for the conversion learned while performing an ethics evaluation on a deployed AI medical device. The utilized ethical principles were adopted from the Ethics Guidelines for Trustworthy AI, and the conversion into practical insights and recommendations was accomplished by an independent team composed of philosophers, technical and medical experts.

History

Received 31 August 2020
Accepted 7 December 2020
Published 14 December 2020

Keywords

Artificial Intelligence, Ethics Evaluation, Case Study, Medical

Contact

James Brusseau
Department of Philosophy,
Pace University,
1 Pace Plaza, New York, NY,
10038
Email: jbrusseau@pace.edu

Acknowledgements

None

Disclosure of Funding

None

AI Ethics Journal

Introduction

AI ethics is undergoing two transformations. First, sprawling sets of ethical principles compiled by academic organizations and private companies (Peters et al. 2020, Whittlestone et al. 2019) are merging into a fragile consensus about what AI ethics means on the theoretical level. The loose consensus is reflected as the European Council's Ethics Guidelines for Trustworthy Artificial Intelligence (Floridi and Clement Jones 2019), and it yields what Morley et al (2020: 2147) call "the second phase of AI ethics: translating between the 'what' and the 'how.'" This essay contributes to the conversion of abstract principles into concrete artificial intelligence applications by documenting learnings acquired from a robust ethics evaluation performed on an existing, deployed, and functioning AI medical device. A nondisclosure agreement with the manufacturer will be honored here, but in broad terms the device generates a proprietary analysis of electrocardiograms that filters for anomalies and patterns associated with impending coronary disease. The evaluation was invited by the device-maker.

Our ethics evaluation formed part of a larger inspection involving technical and legal aspects of the device that was organized by computer scientist Roberto Zicari (2020). This document is limited to the applied ethics, and to my experience as a philosopher. These are nine lessons I learned about applying ethics to AI in the real world.

1. Start at the End: What is the goal?

An ethical evaluation helps AI-intensive companies conceptualize their work on the human level, as opposed to the technological, financial, and legal. Because the AI medical device we were investigating generated predictions of limited certainty, and only about the probability of impending coronary issues, there arose technologically embedded statistical questions. There were also financial concerns about how the predictive uncertainty could affect sales, and legal liability worries surrounded the possibility of misdiagnosis. As distinct from those sorts of difficulties, our evaluation initiated with ethical values and proceeded to their application. Departing from the value of human autonomy, one question we asked was: Does the diagnosis increase, or actually *decrease* patients' ability to direct their own lives? If the machine reports the possibility – not the certainty – of imminent problems, it may turn out that it is the diagnosis more than the disease that ultimately limits a patient's vigorous activity.

There is no perfect solution to this problem of imperfect information, but there is a difference between obliviousness to the dilemma and engaging it in clear, precise language. More than any solution, engagement is the goal of an AI ethics evaluation, which means illumination of what is at stake in terms of lived human experience.

2. Assemble a Harmonious Team

Our team included philosophers, AI engineers, and domain experts which, in this case, meant medical doctors. We consistently had several of each as we worked, and a total of eight to twelve functioned well for online meetings. The device belonged to a German company, and the team was composed overwhelmingly of northern Europeans, with an emphasis on Frankfurt.

Our project proved demandingly theoretical in two senses: the transacted concepts were sophisticated and specialized, and working together required exchanging them across the humanistic and scientific sides of knowledge. So, besides having people from the required backgrounds and in the right numbers, our team's congealing depended on members embracing interdisciplinarity and wielding significant cognitive power.

AI Ethics Journal

One obtrusive question about participants: Should the evaluated company provide a team member? In our case, the study was wholly independent, which protected against conflicts of interest. But, implicit in the independence is a judgement that our understanding of the technical AI and medical patient experience was sufficient to proceed without recourse to an internal company expert. Different circumstances may lead to a rebalancing of the independence versus expertise tradeoff.

3. The Evaluation Happens in the Real World, Not a Classroom

Kant's idea of dignity entered one of our discussions, and to elaborate the concept I introduced his famous example of executing prisoners, just as I typically do in my classes (Kant 1797: 102; Ak. 333). It did not go over well. The intellectual ambition was reckless, the context was a European mindset, and the result was our Zoom call devolving into emotional appeals for the end of all capital punishment before collapsing into a stilted and fruitless hour. The lesson is that the role of the philosophers in the group is to maintain a disciplined and intellectually rigorous process, but an ethics evaluation is not a classroom thought experiment, and preserving comity occasionally requires rounding the edges of philosophical purity in the name of getting something done.

In another case, one of the team's medical doctors contributed an extended monologue on the difference between patients who are asymptomatic and those who display only negligible symptoms. Here again, a distinction that centers intense discussion *within* a profession had the effect of hindering the larger, interdisciplinary ethics evaluation.

Where should the line be drawn? When should the expertise of the ethicists or the domain experts be curtailed to keep the process advancing? The question I learned to ask myself was: Is this philosophical truth going to affect any downstream corporate decisions? If the answer was *No*, then the truth became expendable.

4. Get Your Hands on the Equipment

We could have worked from a written description of the medical device, something like: *the AI analyzes electrocardiogram heart rhythms for anomalies indicating coronary disease, and reports a positive or negative finding.* But, palpable experience was invaluable. From hands-on experimenting we learned the human starkness of the machine's output, which was only two colors, green and red. Rationally, the dichotomy made sense since it was easy to register in some mathematical sense as high and low risk. Actually seeing the color, however, *looking* at it, that vision provoked questions on the flesh and blood level. Can those scoring green start eating Doritos for dinner? Should the red-scorers accelerate the bucket list? The truth sits between the extremes for most patients, but the larger point is that these critical human questions arose after – and to some extent because – our evaluators tangibly lived the user's reality, and so confronted for themselves the divergence of a two-colored output.

Further, the company asserts that the AI increases patients' quality of life which, in the abstract, seems obvious: getting notified before a heart attack instead of *by* one is an improvement. When a patient is faced by only a green or red result, however, ambiguities surge. What level of risk does red actually imply? Why? When it comes to living, is it sometimes better to *not* know? Does the addition of a yellow, intermediary score – which, in fact, the company did add – resolve any of these questions and dilemmas, or just make them worse?

These questions go to the core of the imperative that the AI benefits patients, and one reason they rose so forcefully was that our evaluators saw the colors with their own eyes.

AI Ethics Journal

5. Decide the Ethical Principles First, and Stick with Them

Aristotle, among others, notes that the beginning is more than half the whole, and he has rarely been more right than in the case of deciding the specific and limited set of ethical principles to structure an ethics evaluation. Our group frittered away hours modeling concerns in terms of one collection of principles, only to decide that a different set should be employed to better manage an aspect of the device's performance, and then another.

For example, one consequence of nearly any inexpensive and convenient AI diagnostic medical test is an increased patient demand for further medical testing: if you make it easy for people to find health problems, they probably will be found. This implies increased expenditure on healthcare and, to at least some extent, money subtracted from other worthy recipients, perhaps including schools, or parks. So, there is an obvious social concern here, and the question for ethics evaluators is: How should it be framed? Initially, our team employed a set of ethical principles that included beneficence, and we used that to structure the spending tradeoff. However, we subsequently changed our core principles to a group derogating beneficence ("justice" now filled an analogous role), and so we had to rebuild our discussion of the concern. Later, another shift again left us recommencing.

Defining which principles will structure the ethics evaluation is the cornerstone task: they guide the effort to locate and describe ethical issues and determine in the first place what can possibly *count* as an ethical issue. Everything the evaluating team does happens inside the principles that are first established.

Which ethical principles should be established? There is no shortage of candidates. More than eighty significant proposals have been registered recently in academic journals, by foundations, institutes, and by private companies (Hagendorff 2020). Our group finally settled on one of the most, or perhaps the most representative proposal, the *Ethics Guidelines for Trustworthy AI* authored by the European Commission's High-Level Expert Group on Artificial Intelligence (AIHLEG 2019). Besides the institutional sanction, the set worked for us because it was supported by four pillars, including the imperative to do no harm, which connects well with traditional healthcare ethics.

Every team will need to select or assemble its own set of principles, and do so in the midst of an inescapable Catch-22: you cannot know which ones work best to model a particular AI until you employ them, but you cannot begin employing until you have chosen the principles. So, there will be some initial heuristic experimentation, but practical reality eventually demands a decision that endures through the realization that no single set will perfectly cover any one machine.

6. Work in Two Directions

To maximally utilize the practical medical experience of our domain experts, as well as the theoretical expertise of our philosophers, our team located and described ethical issues by working in two directions.

Starting from the empirical, we asked the healthcare team members to describe problems they saw – or expected to see – arising from the AI's use. One of our team's medical doctors, for example, introduced the sensitivity/specificity distinction in testing: a diagnostic may correctly locate all those facing coronary risk (high sensitivity) but also falsely label a significant fraction of healthy patients as diseased (low specificity). The sensitivity and specificity measures move independently, though in practice there may be a tradeoff. Regardless, there is a dilemma of false positives here: How much needless testing and anxiety should be suffered by how many patients falsely identified as suffering coronary disease, in order to identify and treat

AI Ethics Journal

those facing true risk? And perhaps more significantly, how should the dilemma be weighed? In terms of monetary costs? As days lost to needless treatment versus days gained by heart-attacks avoided? Or possibly the answer lies in a different direction: saving lives is so important that non-fatal factors like money and time should hardly be considered. No matter the response, it is instigated from the ground up by domain-specific experience.

Starting at the other extreme, abstract ethical questions can open insights downward into tangible, lived experience. The *Guidelines for Trustworthy AI* lists requirements and sub-requirements which can serve as valuable interrogatory prompts. One that yielded discussion in our group revolved around solidarity, the social inclusiveness imperative that no one be left behind. In AI medicine, because the biology of genders and races differ, there arises the risk that a diagnostic or treatment may function well for some groups while failing others. As it happened, the machine centering our evaluation was trained on data which was limited geographically, and not always labelled in terms of gender or race. Because of the geography, it seemed possible that some races may have been over- or underrepresented in the training data. The solidarity question is: should patients from the overrepresented race(s) wait to use the technology until other races have been verified as fully included in the training data? A strict solidarity posture could respond affirmatively, while a flexible solidarity would allow use to begin so long as data gathering for underrepresented groups also initiated. Limited or absent solidarity would be indicated by neglect of potential users, possibly because a cost/benefit analysis returned a negative result, meaning some people get left behind because it is not worth the expense of training the machine for their narrow or outlying demographic segment.

There are no absolute responses to these questions, which is acceptable insofar as the ethical survey's goal is to To locate and describe them, it helped to work from the domain experts up to the ethical theories and, separately, from the ethical theories down to the functioning AI machine.

Note: As our group's work finished, the European Council published its *Assessment List for Trustworthy* AI (ALTAI 2020), which contains checklists of ethically pregnant questions that may be posed to specific technologies. It promises to be a good future resource.

7. Locate Tensions

Some ethical dilemmas are best conceived as tensions between positive outcomes. In our evaluation, we articulated a specific divergence between privacy and performance (or accuracy). Ideally, patients' privacy would be optimized, meaning they would maintain full control over their coronary data: after AI analysis, the electrocardiograms would be deleted or returned to the patient. It is also ideal, however, that every patient's case be added to the database to improve performance. More, maximizing long term healthcare quality would require monitoring the patients and their symptoms continuously for downstream effects of the AI diagnosis and responses to it. There is no right side to this split, but there are better and worse understandings of the tension between individual privacy, and AI performance along with social welfare.

Topically, debates about Coronavirus tracking applications share this structural tension. In an ideal world, only individuals would be privy to their own health status. It is equally ideal, however, that others be aware of nearby infection risks.

Regardless of the specific tension, our team located two broad types: practical, theoretical. Practical tensions can be resolved with more data, processing, money, work, and time. It may be possible, for example, to anonymize data

AI Ethics Journal

to protect individual privacy while also serving the public good of improving coronary diagnosis. Theoretical tensions, by contrast, are irresolvable. In AI, there will always be a conflict between privacy and personalization. On one hand, privacy is control over access to our personal information, and maintaining it requires limiting the release of data about ourselves. On the other hand, the power of artificial intelligence in healthcare – and elsewhere – lies in customizing outputs for accurate medical diagnoses (and Netflix movie suggestions, LinkedIn job recommendations, OKCupid romance matches). Inescapably, privacy and personalized convenience pull against each other across AI reality.

Finally, one resource we found helpful was *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research*, published by the Nuffield Foundation (Whittlestone et al. 2019b). It elaborates and exemplifies a set of ethical tensions that recur in AI ethics.

8. Don't Fear Rubrics (Especially in the Face of Paralysis by Analysis)

For philosophers, the idea that ethics can be checklisted is like an artist being told to paint by numbers. Still, a rubric worked for us, and it did because of a specific problem: it wasn't that we needed help doing the ethics evaluation, instead, we needed help to *stop* investigating. Left to our own devices, we could have discussed and debated interminably.

Our team's formal, written work initiated with a list of ethical issues surrounding the AI medical device as they were compiled by one of the philosopher members. Next, a pair of Zoom meetings joining the entire group yielded a second, expanded set of ethical concerns and flags, along with a number of distinct theoretical approaches to them. For example, the problem of false positive diagnoses may be conceived in terms of patients' autonomy and agency if the aroused health anxiety causes a needless restriction of activities. The problem could also be conceived in terms of social wellbeing if public money goes to superfluous medical testing instead of some more solid public good. In any case, we were simultaneously debating *what* counted as an ethical issue, and *how* it counted.

Then the other philosopher went back through and re-organized the set, combining some issues and finding others redundant, and that led to still another round of Zoom exchanges again lacking firm conclusions. There was a moment during this stage where the threat of terminal incompletion seemed real.

Ultimately, to force a consensus we had each team member commit their personal thoughts to a short rubric. First, we narrated each discussed ethical dilemma and tension in our own words. Then we mapped each one onto our ethics principles and tensions. Concretely this meant taking the four pillars of the *Ethics Guidelines for Trustworthy AI* (Respect for human autonomy, Prevention of harm, Fairness, Explicability) and selecting the one we individually found the most apt. Then, each of those pillars supports a number of requirements, from which we selected, and then each requirement contains sub-requirements which we also selected. The idea was to capture the dilemma in structured ethical terms. Here is an example of an entry from my rubric, with the description modified to protect the nondisclosure agreement:

Description: Black-box algorithm complicates the attribution of accountability. Accountability for a particular diagnosis is not clearly defined for patients or for doctors.

Map to Ethical Pillars/Requirements/ Sub-requirements (*Guidelines for Trustworthy AI*): Fairness > Accountability > Auditability

Identify Ethical Tensions (Nuffield Foundation publication): Accuracy *versus* transparency/explainability

AI Ethics Journal

Finally, we gathered the contributions, and their clean articulations allowed easy comparisons across the group and so enabled a brisk move to agreement on final issues, mappings, and tensions.

In the end, rubrics usefully facilitated a conclusion by forcing stark descriptions and categorizations that funneled our team toward consensuses. Were ethical nuances lost along the way? Yes. The other option seemed to be interminable nuances and no conclusions.

9. The Value Added is the Engagement

After ethical issues have been captured and described, a response is written and presented to complete the evaluation. It may include concrete solutions to identified problems. A medical device trained with data from a limited geographical area raises a fairness concern: Does it function equally well for Japanese patients? Central Africans? And, there may be a solidarity resolution: the company provides greater transparency about the training data already employed, and commits to including a wider range of patient information in continued development. It is also true, however, that in many cases problems like this are going to be solved on their own from the business side. Manufacturers want to maximize product performance and sales, and that can lead to solidarity ethics results without the ethics intervention.

Since many of this essay's readers, I suppose, are philosophers, it is redundant to note that the only questions genuinely worth asking are those without answers. For that reason, companies that commission an ethics evaluation will find that the primary value added lies outside of solved problems, and instead in questions that succeed primarily because they get posed. The AI medical device we investigated promised to conveniently, efficiently and inexpensively improve the quality of patients' lives by discerning the probability of impending heart problems. One of our mapped ethical responses traced through the principle of autonomy to ask whether a diagnosed, middle-range probability of coronary disease actually increases or decreases patient vitality and self-determination. Is it true that the machine serves patient lives if it *creates* that kind of health uncertainty? How do we know? And, what *counts* as quality of life? How is it measured? These are the kinds of tangles that connect manufacturers with their own products on the human level as opposed to the financial or legal, and this engagement is the real value of an ethics evaluation: it is facing artificial intelligence dilemmas that won't be solved by more data and faster processing, and that will probably be rendered even more acute.

References

1. (AIHLEG) Artificial Intelligence High Level Expert Group. (2019). *Ethics Guidelines for Trustworthy AI, European Commission*. Accessed 20 October 2020: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

2. (ALTAI) Artificial Intelligence High Level Expert Group. (2019). Assessment List for Trustworthy AI, European Commission. Accessed 20 October 2020: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

3. Floridi, Luciano and Tim Clement-Jones (2019). The five principles key to any ethical framework for AI. New Statesman, 20 March. Accessed 21 October 2020: https://tech.newstatesman.com/policy/ai-ethics-framework

4. Hagendorff, Thilo. (2019). The ethics of AI ethics—an evaluation of guidelines. arXiv:1903.03425 [Cs, Stat].

5. Hagendorff, Thilo. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines* 30: 99–120. https://doi.org/10.1007/s11023-020-09517-8

6. Kant, Immanuel. (1797. 1965). The Metaphysical Elements of Justice, Part I of *The Metaphysics of Morals* translated by John Ladd. Indianapolis: Bobbs-Merrill. Quotations and parenthetical references are from this translation and edition, followed by the standard AK pagination.

7. Miller, Catherine and Rachel Coldicott. (2019). People, power and technology: The tech workers' view. *Doteveryone*. Accessed 21 October 2020: https://doteveryone.org.uk/report/workersview/.

8. Morley, Jessica; Floridi, Luciano; Kinsey, Libby *et al.* (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics* 26: 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

9. Peters, Dorian and Rafael Calvo. (2019). Beyond principles: A process for responsible tech. *Medium*, May 2. Accessed 21 October 2020: https://medium.com/ethics-of-digital-experience/beyond-principles-a-process-for-responsible-tech-aefc921f7317.

10. Peters, Dorian; Vold, Karina; Robinson Diana and Rafael Calvo. (2020). "Responsible AI—Two Frameworks for Ethical Design Practice," in IEEE Transactions on Technology and Society, vol. 1, no. 1, pp. 34-47, March. DOI: 10.1109/TTS.2020.2974991.

11. Whittlestone, Jess; Nyrup, Rune; Alexandrova, Anna and Stephen Cave. (2019). "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions," in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). Association for Computing Machinery, New York, NY, USA, 195–200. DOI:https://doi.org/10.1145/3306618.3314289

12. Whittlestone, Jess; Nyrup, Rune; Alexandrova, Anna and Stephen Cave. (2019b). "Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research." Nuffield Foundation, Leverhulme Centre for the Future of Intelligence, University of Cambridge. Accessed 25 August 2020: https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf

13. Zicari, Roberto (2020). Z-Inspection: A process to assess trustworthy AI. http://z-inspection.org/