AI and Ethics (2021) 1:21–25
https://doi.org/10.1007/s43681-020-00013-4

OPINION PAPER

You cannot have AI ethics without ethics

Dave Lauer[1]

Received: 2 September 2020 / Revised: 2 September 2020 / Accepted: 4 September 2020 / Published online: 6 October 2020

1 Introduction

Artificial intelligence has emerged as the preeminent technology of the twenty-first century, infiltrating nearly every industry and impacting our lives in obvious, but also increasingly subtle ways. Each industry and company is grappling with how to leverage this new technology to optimize or personalize their products or offerings, understand their business or clients better, or to unlock new sources of revenue and opportunity. In the midst of this innovation, the concept of AI ethics is often overlooked, paid lip service, or simply ignores the idea that you cannot have AI ethics in isolation from a broader and all-encompassing ethical approach.

Industries are experimenting with AI in a very difficult environment. Widespread adoption of AI is a relatively new phenomenon. Outside a small circle of math experts, the nuances of different approaches and techniques are not well understood. Even the curricula for data science degrees and certificates are primarily focused on the application of these techniques, rather than the math that underpins such models. For most executives, and especially for legal and compliance professionals, AI remains a black box.

In this paper, I will examine the reasons that ethical deployment of AI has been so elusive for so many high-profile organizations, and I'll explain why there have been such egregious examples of unethical AI built and deployed into the world. I will draw on examples and lessons from other fields, such as medical ethics and systems theory, to demonstrate that AI ethics simply cannot exist without a broader culture of ethics. I will make the case that only organizations with a firm grounding in ethics, and an appreciation for the way complex systems behave can succeed at ethical deployment of AI.

2 Artificial integrity

Most AI projects fail to get out of the research lab, but many that do are soon embroiled in scandal. Let us start with a recent example, a new service called Genderify. Genderify set out to identify someone's gender based on their name, email address or username. In hindsight, this service was probably a terrible idea to begin with in light of the current cultural discourse on gender and identity. Surprising few people other than the founders, Genderify made predictions like "Meghan Smith" was 60% likely to be female, but "Dr. Meghan Smith" was 76% likely to be male. Needless to say, they shut down their service completely within hours of launching.

It would be easy to attribute such a failure to a lack of AI ethics, or a lack of an appropriate ethical AI framework on Genderify's behalf. But was AI ethics the failure here? What would the ethics of AI have told the principals of Genderify that any straightforward ethical framework wouldn't have? Can any framework take a fundamentally unethical objective, and somehow make it ethical?

Of course, this example comes with some obvious red flags. But there are far more insidious problems that were too complicated to have foreseen ahead of time and just as difficult to diagnose after the fact.

Examples of such ethical lapses abound: Uber's withdrawal from their autonomous vehicle development after killing a pedestrian; Facebook's rampant algorithmic spread of misinformation and disinformation; Clearview's illicit facial recognition surveillance and the backlash that followed.

In Microsoft's development of Tay, a Twitter chatbot, and Harrisburg University's attempt to develop technology to predict criminality, we have seen how racist training and racially biased data ultimately lead to racist AI models. Courtesy of Microsoft and Facebook, we have also seen how research image collections with sexist bias can be turned into AI models that link images of shopping, cleaning and cooking to women.

✉ Dave Lauer
dave@urvin.ai

[1] Urvin AI, Philadelphia, PA, USA

In each of these examples, we are confronted with ethical dilemmas. Some are more obvious and superficial than others. Each incident invites a series of questions. Why hadn't they defined the right set of policies and procedures to prevent such an outcome? Why couldn't Genderify see what, in retrospect, seems like such an obvious problem? In the case of Facebook (and many others), why are the steps that they are taking, and their vow to "fight the spread of false news," proving to be ineffective? Where is the broken part?

3 The broken part fallacy

The fallacy of the broken part is a well-understood principle in complexity theory. When there is a malfunction, the first instinct is to identify and fix the broken part. If a plane crashes, which part malfunctioned? If an autonomous vehicle misidentified a white truck as simply being part of the sky, and drove right into it, where is the problematic code? In these examples (and countless others), the broken part is only the most superficial problem. "Fixing" the broken part will often fail to prevent a future problem, because these types of problems are systemic, ordinarily involving cascading or multi-system failures. In some instances, these failures may occur "when no parts are broken, or … seen as broken." [2] Our impulse to "fix" a broken part is driven by our grounding in linear thinking and the search for the cause of an undesired effect.

Systems thinking, especially when it comes to system safety, demands that one examines the entire ecosystem in all of its complexity. "Systems thinking is about relationships, not parts." [2] In nearly every case, this search for a broken part leads to a band-aid solution that attempts to address the problem without consideration of the complexity underlying the causes.

Discussions over technology-driven problems typically turn to talk of "bugs" or programming errors. However, according to noted systems engineering and safety researcher Nancy Leveson, "[n]early all the serious accidents in which software has been involved in the past 20 years can be traced to requirements flaws, not coding errors." [5]

These errors often manifest themselves as coding errors or user interface flaws, but they are the consequence of poor requirements, poor governance and poor processes. This also describes the state of AI today. When AI ethics fail, we assign blame to inadequate or narrowly specified training data while looking past organization-wide ethical shortcomings.

4 Bad medicine

The alternative to a narrow ethical AI approach is a thorough examination of the entire environment that ultimately led to the problem or failure, including company management, legal and regulatory incentives, manufacturing practices, employee training, quality assurance, and so on.

In each of the aforementioned AI failures, the "search for the broken part" only served to obscure a more systemic ethical deficit. In each case, the failure to build ethical AI can be traced to an organization-wide failure of ethics. But how to go about overhauling ethics at the organizational level? Can a series of policies and checklists actually make an organization ethical? Can ethical AI exist in a vacuum separate and distinct from broader ethical questions? Can any narrowly defined ethical field exist in a vacuum from broader ethics? I hope to show why the answer to these questions is "Clearly Not."

Perhaps a short foray into a far more mature field could be instructive. The practice of medicine has been grappling with ethical and moral questions since before the first Hippocratic Oath was taken. While this is not the place for an exhaustive exploration of medical ethics, the "metamorphosis of medical ethics" as Edmund Pelligrino terms it, provides an instructive lesson for AI practitioners. The field has evolved from the Hippocratic Oath, and its broad and relatively subjective expression of "genuinely ethical precepts, such as the obligations of beneficence, nonmaleficence, and confidentiality, as well as … prohibitions against abortion, euthanasia, surgery and sexual relationships with patients." [7] The limitations of this approach were gradually recognized, especially as the approach became incompatible with a more modern, informed and equal society.

As such, a theory of "prima facie principles" was developed, and "adapted to medical ethics by Beauchamp and Childress' *Principles Of Biomedical Ethics*." [7] They settled on four principles for medical ethics—"nonmaleficence, beneficence, autonomy, and justice." [7] These principles should generally sound familiar to anyone with experience in ethical AI frameworks.

But the medical field has been confronted with the shortcomings and subjectivity of putting these principles into action. For instance, the emergent idea of autonomy "directly contradicted the traditional authoritarianism and paternalism of the Hippocratic ethic that gave no place for patient participation in clinical decisions." [7] Over time, this principle of medical self-determination has been largely accepted, especially in America. But today, we are facing a new contradiction. As unethical social media platforms proliferate misinformation and anti-science into the mainstream, such as the amplification of the anti-vaccination movement, the ethical importance of autonomy is in direct conflict with the ethical importance of truth. This underscores the key shortcoming of the *prima facie* framework. These principles are vague, and relatively static. They lack the dynamism to address major bioethical dilemmas such as "abortion, euthanasia and a host of other issues." [7] As Pellegrino explains, "[w]hat is required is some comprehensive philosophical underpinning for medical ethics that will link the great moral traditions with principles and rules and with the new emphasis on moral psychology." In other words, even in a field that has been grappling with ethical questions and issues for thousands of years, the attempt to define an ethical approach specific to the field, and divorced from broader ethical philosophy and questions, remains a moving target.

Much like biomedical ethics, AI ethics do not exist in a vacuum. Organizations that fail to grapple with basic ethical questions, or who have neglected to establish a culture of ethical and moral behavior, will not succeed. A fundamentally unethical organization, or the representative of an unethical industry, simply won't have the capabilities to build and deploy ethical AI.

5 Systemic safety

This is because companies are complex organizations. They exist within complex ecosystems inhabited by regulators, customers and partners. Those which neglect to incorporate complexity theory and systems thinking into their consideration of AI ethics are doomed to fail. There has been a significant amount of study and work done to better understand the complex interplays of such ecosystems, in particular the large body of work around safety systems in industries such as automotive and aerospace. In fact, the fallacy of the broken part is based on work in these fields, as well as the all-too-human desire to assign simple explanations to complex issues.

Unfortunately, in the world of complexity theory, there are few linear relationships and few simple answers. The field of AI bears much resemblance to the practice of system safety in these industries, which focuses on several areas:

- the complex interplay of incentives that are created from law, regulation, financial markets and for-profit businesses;
- the fostering of a "culture of compliance," powered by affirmative top-down leadership, bottom-up empowerment of the employees who are closest to the problem, and actual adherence to company policies;
- the training of front-line employees who are designing and building these systems and who have the most first-hand experience and ability to impact implementation;
- the sophistication and insight of empowered regulators who understand the industry, and co-evolve with the firms that they regulate;
- the avoidance of prescriptive top-down solutions in favor of principle-based guidance and appropriate transparency for policing and enforcement.

While the establishment of an ethical AI framework for a company is an excellent and important step, frameworks that fail to account for system-wide complexities will struggle with relevance as the world shifts and changes, and as decisions are made in the face of scarce resources and competing incentives.

6 Facebook: in control, absent responsibility

Let us dig into an example featuring everybody's favorite punching bag (and deservedly so). Facebook has played a critical and fundamentally unethical role in the mainstream proliferation of AI. The social media giant has profited immensely from the spread of misinformation, both during various elections around the world in 2015 and 2016, and since that time.

In response to broad criticism for their role in spreading such misinformation, Facebook has undertaken many different initiatives to address a problem that they acknowledge "is harmful to our community, … makes the world less informed, and … erodes trust." [6] These initiatives include, amongst many other efforts, using third-party fact-checking organizations, leveraging machine learning to detect fraud and spam, and the creation of an "independent oversight board" to help make "some of the most difficult and significant decisions around content." [1]

Will all of these steps work? After several years spent fighting the misinformation in its midst, and several months spent answering to an oversight board, the answer is, once again, "Clearly Not." Facebook remains a primary source for misinformation, and their most popularly shared posts are still extremist opinions masquerading as journalism and fact. Even as I write this article, Facebook is facing yet another scandal for failing to intervene as armed militias used its platform to communicate, organize, and ultimately descend on Kenosha, Wisconsin with deadly consequences.

Is this due to a lack of effort? A lack of ethics? Or are Facebook employees simply following their prime directive—maximize profits and shareholder value while thriving in a system and ecosystem that practically guarantee these problematic outcomes?

As long as we are focused on Facebook and the spread of misinformation, surely Facebook can exert some control over the content of paid advertising? This would, superficially, appear to be an easier, more tractable problem than managing user content. The question here is not whether Facebook *can* do anything about it. The real question is, is Facebook incentivized to do anything about it?

Facebook exists in a broader legal and regulatory ecosystem, one that would seem to supersede these ethical concerns. As Matt Stoller explains, “Section 230 of the Communications Decency Act immunizes Facebook from any consequences for the content of ads bought on their platform; they cannot be sued for facilitating fraud and counterfeiting, so they do not have any incentive to do anything about it.” [8]

Stoller goes on to point out that the motivations are put into even starker relief when compared to traditional media. Where the latter sell ads manually, Facebook uses an automated algorithmic system. Stoller explains the early twentieth century idea of Absentee Ownership, “which is when the locus of control and the locus of responsibility are different. Facebook isn’t legally responsible for the consequences of what goes up on its network, but it still has control over what goes up on its network.” [8] He makes this observation in the context of fraud and advertising, but it is equally applicable to our concern about incentives and ethics. Despite these clear regulatory incentives, we can imagine an ethical spectrum of reactions from companies operating in such an environment, and have seen such from various social media platforms. But Facebook has clearly fallen on the wrong end of this spectrum, as one of the least ethical players in the space.

7 AI ethics require actual ethics

So far we have introduced two broad ideas—that fundamentally unethical companies cannot simply deploy ethical AI through compliance and checklists, and that both ethical and unethical companies exist as part of a broad and complex ecosystem that influences their behavior in both obvious and subtle ways. This is not to disparage every company working across the digital or AI landscape. There are positive examples of companies who recognize this dynamic, respond to it with clear ethical principles and frameworks, and who work hard to ensure that such frameworks are part and parcel of their companies’ operations and ethos, rather than a box to be checked for marketing or sales purposes.

One clear example is found in the realm of search, advertising and privacy. In contrast with a search engine like Google that harvests as much information as possible and sells that to generate ad revenue, DuckDuckGo is completely focused on providing private search. The company markets itself as “ethical by design” [3] and is trying to differentiate itself from Google based on ethics and principles. Their ethos is that “the responsibility of making ethical decisions is not delegated to one or two roles, it’s in the fabric of how we work across *all* roles. It’s simply a natural byproduct of how we operate.” [3] In other words, it is an emergent property of their organizational principles.

This is exactly what Nancy Leveson describes when she calls system safety “an emergent property of the system that is achieved when appropriate constraints on the behavior of the system and its components are satisfied.” [5] Ethics can easily be seen as sharing the same characteristics of system safety in this context, and AI ethics can simply be seen as another branch of, or an emergent property of, a broad ethical approach.

This broad ethical approach is hardly specific to AI. When you review ethical AI frameworks and guidance, you see what is often described as “motherhood and apple pie.” These are tenets that few would disagree with. Most of the stated conditions have analogues in (or are simply a direct copy of) standard corporate ethics frameworks. For example, a review of 84 varying frameworks or guidelines found that the most common principles are: “transparency, justice and fairness, nonmaleficence, responsibility and privacy.” [4] The overlap with the aforementioned medical ethics principles of “nonmaleficence, beneficence, autonomy and justice” are clear. So is AI ethics special or unique in this context?

As the field of medical ethics has learned over decades of study and research, “[a] continuing dialogue with moral philosophers is requisite to assure that clinicians do not lose the benefits of a rigorous and critical analysis of their ethical decisions.” [7] In other words, principles are all well and good, but without a broader moral and ethical foundation, they can be disregarded in actual practice.

Once again, the field of complex systems has a lot to offer to better understand why this happens. These ethical principles are no match for the far more complex and intertwined world that is AI. Sidney Dekker puts it best at the beginning of his book, Drift Into Failure:

> The growth of complexity in society has got ahead of our understanding of how complex systems work and fail. Our technologies have gone ahead of our theories. We are able to build things - from deep-sea oil rigs to jackscrews to collateralized debt obligations - whose properties we can model and understand *in isolation*. But, when released into competitive, nominally regulated societies, their connections proliferate, their interactions and interdependencies multiply, their complexities mushroom. And we are caught short.
>
> We have no well-developed theories for understanding how such complexity develops. And when such complexity fails, we still apply simple, linear, compo-

nential ideas as if those will help us understand what went wrong [2].

Nancy Leveson defines such complexity more succinctly as “intellectual unmanageability.” [5] So to imagine that we can manage through this complexity with a series of checklists and policies is either naively optimistic, intellectually lazy, or the height of hubris.

8 Confronting the complexity of organizational ethics

Dekker admonishes that organizations must find “the political, practical and operational means” [2] to invest in ethics, “even under pressures of scarcity and competition.” [2] That is when such investments are needed the most, when the incentives to cut corners are greatest.

The problem with most ethical frameworks is generally not in their principles. The problem is when principles clash with a complex world beset by competing incentives and scarce resources. The problem arises when principles are adopted in a componential manner, rather than a systemic manner. The entire field of AI ethics is the perfect illustration of this. It may seem irrational to lament the inadequacy of ethical AI frameworks in the Journal of AI and Ethics. But I prefer to think of it as a strong imperative for the very existence of such a journal, and a prompt for some of the illumination we might achieve here.

Other fields have wrestled with the exact questions that we are confronted with in AI. There is much to be learned from their debates and study. AI presents unique challenges in many ways, but conversely, “there is nothing new under the sun.” Instead of trying to reinvent ethics, or adopt ethical guidelines in isolation, it is incumbent upon us to recognize the need for broadly ethical organizations. These will be the only entrants in a position to build truly ethical AI. You cannot simply *have* AI ethics. It requires real ethical due diligence at the organizational level—perhaps, in some cases, even industry-wide reflection. Until this does occur, we can look forward to many future AI scandals and failures.

Funding Not applicable.

Code availability Not applicable.

Compliance with ethical standards

Conflict of interest Not applicable.

Availability of data and material Not applicable.

References

1. Clegg, N.: Welcoming the oversight board.https://www.about.fb.com/news/2020/05/welcoming-the-oversight-board/ *(2020). Accessed 18 Aug 2020*
2. Dekker, S.: Drift Into Failure: From Hunting Broker Components to Understanding Complex Systems. CRC Press, Boca Raton (2012)
3. DuckDuckGo Blog: Ethical, by design: how we design with your privacy in mind. https://www.spreadprivacy.com/ethical-by-design/. Accessed 30 Aug 2020
4. Jobin, A., Ienca, M., Vayena, E.: Artificial Intelligence: the global landscape of ethics guidelines. https://www.arxiv.org/ftp/arxiv/papers/1906/1906.11668.pdf (2019). Accessed 12 Aug 2020
5. Leveson, N.G.: Engineering a Safer World: Systems Thinking Applied to Safety (Engineering Systems). The MIT Press, Cambridge (2012)
6. Mosseri, A.: Working to stop misinformation and false news.https://www.facebook.com/facebookmedia/blog/working-to-stop-misinformation-and-false-news *(2017). Accessed 18 Aug 2020*
7. Pellegrino, E.D.: The metamorphosis of medical ethics: a 30-year retrospective. JAMA **269**(9), 1158–1162 (1993)
8. Stoller, M.: Absentee ownership: how Amazon, Facebook and Google ruin commerce without noticing. https://www.mattstoller.substack.com/p/absentee-ownership-how-amazon-facebook. Accessed 28 July 2020