AI and Ethics (2021) 1:205–208
https://doi.org/10.1007/s43681-020-00030-3

OPINION PAPER

Who pays for ethical debt in AI?

Catherine Petrozzino[1]

Received: 3 September 2020 / Accepted: 17 November 2020 / Published online: 3 January 2021

Abstract

Many Artificial Intelligence (e.g., Machine Learning) tools are being developed with ethical debt. They are created and deployed without fully examining and addressing the potential ethical consequences. This paper looks more closely at the concept of ethical debt in Artificial Intelligence and its consequences. The analysis focuses on two prevalent use cases: Artificial Intelligence as a vehicle for screening job applicants and Artificial Intelligence as a predictor for those patients who will require extra healthcare services. The analysis also compares and contrasts the similarities and differences surrounding the concepts of ethical debt versus technical debt. The paper concludes with a discussion on the misalignment between those who decide to incur ethical debt and those who end up paying for that decision.

Keywords Ethical debt · Technical debt · Fairness · Equity

Too often, AI systems (e.g., autonomous vehicles, machine learning predictive analytics) are being developed with ethical debt. The models that underly the behavior of AI systems are implemented with little explicit effort to identify and address ethical concerns such as fairness. As a consequence, the organizations and broader systems that rely on these models could unwittingly inflict potential harms affecting health, opportunity, and general welfare for classes of individuals.

AI ethical debt is incurred when an agency opts to design, develop, deploy and use an AI solution without proactively identifying potential ethical concerns. Some have drawn parallels between ethical debt and technical debt. Technical debt occurs when an organization opts for an easy, sub-optimal software solution in order to economize on resources and time in the near term, with the vague notion of spending time in the future to fix it. An example could be the decision to address scalability issues in a future release. Given investor pressures and business drivers, the decision to defer addressing scalability could be the right one. The key is to analyze technical debt decisions thoughtfully and responsibly.

Conventional wisdom holds that technical debt is owned and ultimately paid for by the organization that opts to incur the debt. This is debatable. Arguably, Zoom took on technical debt when they developed their software without adequately addressing cybersecurity controls. Zoom paid a price in terms of negative publicity and ultimately having to issue a ‘feature freeze’ in order to focus on fixing the Zoombombing and other security-related problems. However, Zoom customers also paid a price by having their personal video meetings compromised, and, in extreme cases, being submitted to hate speech and pornography. This price is more difficult to calculate. What is the impact of anti-Semitic hate speech and nudity on a 14-year old Jewish girl? How about a class of adolescent Jewish girls as individuals and as a group attending a Modern Orthodox high school? To look at technical debt only from the perspective of the organization is myopic technical thinking and misses the socio-technical context in which the solution resides. In some cases, technical debt may have ethical repercussions that need to be factored in the thoughtful and responsible analytical calculus. Casey Fiesler and Natalie Garrett also observe how technical debt can morph into ethical debt in their Wired opinion piece, which also looks at Zoombombing and emphasizes the need for tech companies to think about potential misuse when they’re developing products [1].

However, ethical debt in AI is even more insidious than technical debt. Unlike technical debt where typically an explicit decision is made to opt for the faster, less technically desirable solution, ethical debt results not so much from a decision as an assumption that the AI solution is ethical.

✉ Catherine Petrozzino
cmp@mitre.org

[1] The MITRE Corporation, Bedford, MA, USA

After all, organizations often have a code of ethics or ethics policy that applies to employees. Beyond that, employees who are responsible for creating and deploying AI models are generally ethical, so why would an AI solution suffer from challenges with established ethical principles such as fairness, autonomy, beneficence, privacy and the public good? Add the market-driven pressure to deliver AI-based solutions quickly and the lack of established norms and regulations, and ethical debt becomes the de facto standard. Exactly what emerges as the resulting ethical stance of a given AI solution is a mystery, but one that if discriminatory, usually skews against minorities and other vulnerable populations.

Indeed, there are numerous publicized examples of well-intended initiatives that went ethically awry, two of which are presented below:

- Through an independent study it was discovered that a popular algorithm used by hospitals to determine level of care was far less likely to identify very sick Black patients as needing extra care than very sick white patients [2].
- An article in the Harvard Business Review summarizing an Upturn study that researched AI-based hiring models which were designed to combat discriminatory practices concluded: "Unfortunately, we found that most hiring algorithms will drift toward bias by default [3]."

It's important to note that these problems were discovered by organizations that had no affiliation with the development or deployment of the AI solutions. Both examples are associated with domains in which AI tools are prevalent; they are far from niche cases. In the healthcare example, the authors estimate that similar risk-predication tools are applied to 200 million people in the US annually.

The first example illustrates how an AI solution may be designed to be 'fair' from a technical perspective (i.e., based on healthcare costs, Blacks and whites are equally likely to be flagged as needing extra care) but still leads to outcomes that are unethical in terms of disparate impacts to classes of individuals and contrary to the larger goals of the healthcare ecosystem. Thoughtful and responsible ethical analysis must embrace the full socio-technical context before, during and after deployment of AI tools. This is the front line of ethics: what are the harms and benefits to individuals, groups, and society when an AI solution is deployed by an organization and used for consequential decisions? What are the range of outcomes that should be monitored to validate that the solution continues to contribute to an ethical ecosystem and the public good?

Both examples highlight ethically-relevant explainability challenges that are associated with machine learning AI solutions. Traditional software solutions are deterministic – it's relatively straightforward to understand why an answer was generated from given inputs for an individual. AI tools, especially those based on deep learning can generate results that are difficult to comprehend by impacted individuals as well as the tools' users and even their developers. This lack of clarity can make AI solutions seem capricious and obfuscate the creation of a meaningful appeals process, reduce transparency and individual agency and autonomy, and contribute to ethical debt.

Analogous to technical debt, ethical debt accumulates over time as AI models and results are repurposed or reused. Human judgment and decisions are part of every phase of the AI lifecycle. Many of these decisions have unexplored ethical connotations. Additional judgment and decisions will be made when repurposing or reusing the AI solution, often without understanding the initial ethically-relevant decisions, resulting in increased ethical uncertainty. Even if an organization proactively addresses ethics as part of the repurposing or reusing exercise, challenges remain with the opacity of prior AI solutions.

It's hypothesized that as technical debt accumulates, so does the degree of software entropy; the software becomes more complex and difficult to 'fix'. Likewise, the concept of AI tool entropy may apply to accumulating AI ethical debt. Combining AI tools that have unaddressed ethics may compound ethical complexity to an extent that makes it difficult or impossible to address.

The two examples also demonstrate the most publicized ethical challenge with AI solutions: fairness. Unlike technical debt, which mainly harms the organization and possibly users as a whole, ethical debt in AI has the potential to result in systemic harms to categories of individuals. The first example demonstrates bias against Black people in healthcare while the second example demonstrates the broader challenge of bias against women and minorities in hiring. These represent societal challenges that have been well documented; the latter particularly in the technology sector. Although there can be an array of reasons for unwanted bias, data is often the main culprit.

Data, the oxygen on which AI depends, is inculcated with decades worth of past and on-going discriminatory behavior, both explicit (e.g., data that sorts individuals into categories–ethnicity, gender, socioeconomic status, etc.) and implicit (e.g., data that does not record gender or race, but the results nonetheless reflect historic societal biases). Non-representative training data has been identified as the root cause for differential results in facial recognition systems where minority women have much higher failure rates than white males.

However, even in those cases where some effort is made to address fairness, the end result to individuals may be far from equitable. With the goal of being fairer, models have been developed that incorporate adjustments to correct statistical bias or do not include explicit data on race or gender. However, whether race or gender exists in proxy data, or the model relies on data that is infused with decades of racist and sexist societal policies and practices, the end result can still lead to people being treated unfairly when it comes to the broader mission, whether it be related to healthcare or hiring.

Unlike technical debt which is an a priori organizational decision, ethical debt is exacerbated by the reality that some ethical problems with AI solutions can only be detected after they are deployed. By definition, the AI drift problem is a problem that occurs over time, typically in the range of months or years. Drift is detected after the AI solution has made faulty determinations – and individuals are wrongly flagged or not flagged. Likewise, problems with fairness are associated with trend analysis which requires a substantial amount of data. How this impacts unsuspecting individuals can range from lost opportunity of a relatively innocuous flavor (paying more for a service than others with a similar profile) to upending one's livelihood (lost job opportunity) and health (being denied essential health services).

This after-the-fact determination is troubling on several dimensions. By the time the Upturn study came out, how many women and minorities had been denied jobs for which they were well qualified? How many very sick Black people were incorrectly not identified as requiring extra healthcare? Will and how does an organization make amends for inequitable treatment after the fact, especially when health or hiring is at stake?

Some organizations utilize the 'human in the loop' control as a check on AI fairness. However, this comes with its own troubling problems. Humans are subject to implicit biases which have been well-documented for years. There is also the phenomenon known as 'automation bias' in which AI tools that were designed to assist decision makers become the decision makers. Humans tend to trust technology over their own judgment. In situations where AI is being deployed to supplement an under-resourced organization, it is faster, easier and (arguably) more defensible to agree with the AI result.

The expectations that organizations have regarding the ability of a human in the loop to detect a problem may also be unrealistic. The ability of a domain expert to identify extreme cases where an AI result is inconsistent with the norm is reasonably high. However, noticing patterns of discrimination or subtle shifts over time requires a sustained organizational commitment to review and analyze the AI results against independent data. In practice, a human in the loop can be more of a placebo than a reliable control to address ethical debt.

There can be further problems with how an AI algorithm performs over time besides drift. As COVID has all too readily demonstrated, sometimes society has to accommodate an abrupt seismic shift. According to an MIT Technology Review article about the pandemic's impact on AI: "Machine-learning models trained on normal human behavior are now finding that normal has changed, and some are no longer working as they should" [4]. The article further discusses interesting challenges and approaches with trying to 'correct' AI in the age of COVID. Less clear is what's being done to address the erroneous results that could potentially further harm classes of society that are already struggling with the disproportionate impact of the pandemic. While an organization is fixing its AI problem (assuming it's aware that it has one), what is it doing to identify and fix collateral unjust human impact? What remedies do individuals have if a wayward AI solution incorrectly misclassifies them?.

Ethical debt can be a heavy burden that weighs not only on individuals, but on their families, group (e.g., ethnicity and gender) and society. In both of the above examples, unlike technical debt, those paying for the debt can least afford it. This is perhaps the biggest challenge with AI ethical debt: the misalignment of who incurs debt and who ultimately pays for it. Ethical debt can be profitable for those in the AI industry but very costly for those who lack agency in the decision-making process and don't even know that they are the ones paying.

The irony that tools which are advertised as being able to reduce discrimination instead reinforce it should be lost on no one. Organizations that generate AI solutions and the institutions that use them can only claim the high ground by avoiding ethical debt. AI redlining is no more acceptable than the physical redlining of the past. Nor should good intentions be a substitute for actively addressing ethics upfront during the ideation phase of a project which is considering using AI tools, and throughout all the phases of the AI tool lifecycle.

There's much excitement over AI's potential to generate novel solutions to difficult problems – and AI has demonstrated its value in a variety of applications from retail to healthcare. However the concerns surrounding AI ethical debt have been growing for years with seemingly little inclination by the AI sector to address the complex sociotechnical challenges it presents. One wonders at what point does AI ethical debt become 'ethical hubris', especially when practiced by an industry comprised predominantly of individuals who are least likely to be harmed by it.

Funding This paper was generated without funding.

Compliance with ethical standards

Conflict of interest The author declares that she has no conflict of interest.

References

1. Ethical Tech Starts with Addressing Ethical Debt, Casey Fiesler and Natalie Garrett, Wired Ideas (16 Sept 2020)
2. Dissecting racial bias in an algorithm used to manage the health of populations, Ziad Obermeyer, Brian Powers, Christine Vogeli, Sendhil Mullainathan, Science, (25 Oct 2019)
3. All the Ways Hiring Algorithms Can Introduce Bias, Miranda Bogen, Harvard Business Review (06 May 2019)
4. Our weird behavior during the pandemic is messing with AI models, Will Douglas Heaven, MIT Technology Review (11 May 2020)