Practitioner Perspectives on AI Ethics

Derrick Zhen
dzh1@swarthmore.edu
David Gray Widder
dwidder@cmu.edu
James Herbsleb
herbsleb@cmu.edu
Laura Dabbish
dabbish@cmu.edu

School of Computer Science
Carnegie Mellon University

Motivation
Discussion on software ethics in recent years have centered around the pitfalls of AI powered applications and tools. As a result, guidelines like FAT (Fairness, Accountability, Transparency), have reached widespread consensus; and a variety of fairness toolkits and explainable AI modules are being used to mitigate harms produced by models.

Theoretically, these technical advances empower practitioners (engineers, data scientists, designers) to produce more ethical software. However, we observe that certain classes of harm go beyond FAT, occurring at a business logic level, outside of a practitioner’s sphere of influence. We address this conflict with a broad inquiry.

What do practitioners regard as an ethical concern? And how, if ever, do these concerns ever get resolved?

Methods
• We surveyed 122 software practitioners and performed 20 follow-up interviews.
• Our sample was ~72% male. Respondents had a median 12 years of programming experience.
• We iteratively coded the transcripts and card sorted the survey results.

RQ1: What is an ethical concern?
The issue was more than the company as a whole, the business model, as a whole of that company was pretty fraught. It was, you know, payday lending.
• Ethical concerns arise when practitioners believe that the products they work on have potential for harm.
• Concerns differ in scope. Some practitioners take issue with their organization’s very reason for existing, others are concerned about specific features / vulnerabilities.

RQ2: How are concerns resolved?
And I looked that manager in the eye and I said, Look, I said I’m not doing and I said you are going to have to write me up or fire me, but I’m not doing it.
I think a lot of projects are not in as good of standing is, as people like to think. So if 100% of your workforce leaves a project that can give the client cold feet on the project. Then they’re likely to just cancel the project completely.
• High risk actions practitioners took include switching jobs or explicitly refusing to work on projects. Low risk ones include reducing productivity or implementing technical solutions.
• Sometimes, practitioners who resign collapse their projects; other times they are replaced. Sometimes, escalation is successful; other times, nothing happens.

RQ3: What factors influence resolution?
So I didn’t own anything, and I didn’t have any debt. So I did not did not feel any compulsion to stay at that job, or to play nice, you know, to put up with, to put up with being asked to do unethical things.
• Personal factors like Visa status, savings, parental support, dependents influence their ability to take high risk actions. Political ideology, religion, and personal identity determine what a practitioner perceives as harmful.
• Workplace culture, like commitment to company values influence whether practitioners see hope in escalation. An organization’s incentive structure is also used strategically by practitioners during resolutions.