Whose Knowledge Counts in AI Debates?
By David Gray Widder (Cornell Tech)
Whose Knowledge Counts in AI Debates?
By David Gray Widder (Cornell Tech)

What counts as “doing AI ethics”? This paper draws on 75 interviews with those doing AI ethics in large companies, small startups, open source, academia and activist groups to examine what kinds of evidence counts in making AI ethics claims. I find that quantitative evidence counts, whereas personal experience of algorithmic harm does not. Using feminist and decolonial theory, I suggest that “Model Cards” and similar interventions may implicitly enforce the epistemological supremacy of quantification, and that intervening in epistemic power hierarchies is necessary to enable those directly experiencing AI harm to be included as valid participants in AI ethics discussion.

Black people in the US have long faced algorithmic discrimination from biased recidivism prediction and racist predictive policing systems, but it was a ProPublica investigation quantitatively demonstrating this discrimination that led to widespread public attention and gave credibility to these problems.

This discrepancy raises the question: what counts as legitimate evidence when discussing AI ethics? And therefore, who has access to this kind of epistemic power?

By examining 75 interviews with those in a wide variety of contexts, I examine what counts as “doing AI ethics”, and relatedly, what kind of knowledge is considered credible in each context. I draw theory from Diana Forsythe and Lucy Suchman together with the works of postcolonial feminist theorist Sara Ahmed and Black feminist theorist Kristie Dotson to make sense of what I find.

I show how, broadly, AI ethics work is seen as lower status than other AI work. Consequently, those doing AI ethics work attempt to legitimize it by constructing ethics work in the same quantitative, “objective” terms as other AI work. One participant even attempted to build a tool to
automate any human judgment out of the completion of model cards, so that it would be “objective”.

Conversely, when workers attempt to draw on their own experience of harm interacting with AI systems while discussing ethics, this personal experience is discounted. As an example of these “located” complaints, one woman of color reported to her team that it felt weird that their VR system projected white hands over hers, but her team ignored her concern. I then examine how we might make more space for “located” and “embodied” complaints, situated in our own experience, by reporting how activist groups examining AI harms do this in their own conversations and work.

I conclude by arguing that academic AI spaces, such as FAccT and related conferences, must attempt to legitimize forms of knowledge found outside of quantitative, academic frames. I sketch and advocate for humble technical practices: quantitative practices which explicitly lay out their epistemic limits, and integrate their resulting claims with those of other knowledge systems.

**Looking outward**

AI ethics is a project that often must defend itself against those who feel it is a detriment to progress in AI: one need only look as far as Marc Andreessen’s word vomit labeling “tech ethics” as “the enemy”, or how time and again, ethics teams are the first to be fired when tech companies do layoffs.

As a result, some AI ethics practitioners reasonably respond by seeking to legitimize AI ethics as a project by using the same kind of evidence which “counts” in AI: quantitative, objective knowledge.

But what will this response exclude? I argue it will exclude the kind of personal experience of harm that AI ethics ought to be particularly aware of: marginalized forms of knowledge, especially that held by those AI is particularly likely to further marginalize.