SESSION 1: TAKING STOCK OF ETHICS ON AI (GENERAL DISCUSSIONS FOLLOWING THE 3 PRESENTATIONS)

DISCUSSANT: MALAVIKA JAYARAM - Executive Director, Digital Asia Hub, Assistant Professor (Practice), Singapore Management University School of Law, and Faculty Associate, Berkman Klein Center for Internet and Society at Harvard University

COMMENTS

[*Interpreters: Please note that if we are running short of time, I will leave out the sections of text that I have highlighted in yellow – I hope this helps you keep track of where I am*]

I'm very grateful to my fellow co-organisers, and particularly to our hosts in South Korea, for facilitating my remote participation at this important event. Those who know me will know that being unable to visit Seoul is a huge personal loss. It is a city that I love, and but for 45 students (and climate change, and time zones!), I would be there with you.

This session comes at an interesting moment, not just in terms of the AI trajectory, but the course I'm teaching this semester, on Ethics and Social Responsibility. Last week, we discussed various schools and theories of ethics – including metaethics, normative ethics, and applied ethics. In a few hours, we will be tackling our first site of applied ethics: the workplace. In other words, we're moving from Aristotle to Enron, and Immanuel Kant to Edward Snowden.

This trajectory, from the general to the specific, from norms and values to lived experiences and use cases, is something I mention because it seems to connect important points raised by our panellists: about the limitations of a subjective, partially representative study, coupled with its benefits as a heuristic, a beginning, and an approach.

Herbert used the analogy of Pandora and Gaia to caution us against moral stories: His statement that "There is an interest behind them usually either legitimizing new power distributions or questioning existing ones" could apply to many of the principles enshrined in the Berkman study. Are ethics codes being used to push us towards soft norms rather than enforceable rights? Are they displacing diverse, subjective realities with anodyne, seemingly objective norms? Are they universal when it's cheaper to have a one-size-fits-all corporate policy, and culturally relativist when it isn't? (For example, the argument that "Asians don't care about privacy, so we're only respecting local culture by not protecting it in certain countries, rather than imposing Western values"!)

In my previous class, we covered something that Herbert alluded to: that all the ethical codes in the world will not make people act, and will not shift behaviour. Knowing the rules does not – in itself - make people follow them. What does? We studied various explanations and motivations behind why people act ethically or morally: out of ego, altruism, fear, hedonism, emotion, reason, etc. If there is no wiring or connective tissue between principles and codes on the one hand, and motivation and behavioural change on the other, the plethora of values and aspirations codified in the Berkman visualization will remain inert, academic and dissonant from the lived realities of most of the world.

It is only through unpacking these principles, weighing them against the social and cultural contexts that they will be embedded in, and noting the gaps from principle to practice, that we will begin to appreciate their potential real world impact. I, for one, very much look forward to the sequel to this Berkman paper, which I have mentally titled "When the Rubber hits the (Ethics) Road".

I wanted to share 3 quick observations:

1. The map is not the territory: This formulation, popular in therapy, comes from Alfred Korzybski's 1933 work, “Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics”. It refers to the gap between perception and reality. I refer to it for 2 reasons: to highlight the gap between ethical codes and the reality of their adoption, but also to suggest a provocation and thought experiment.

Our mental maps are not as obvious as a printed roadmap (or Google Maps) because we use our mental maps to think our thoughts and feel our feelings. In Korzybski's words, we easily confuse them with the territory. But here’s the thing: Brains aim to please. Left to their own devices, our brains will accept whatever maps we give them and will use them again and again. It would be as if someone moved from Seoul to Singapore but continued to use her Seoul roadmap because she was familiar with it and liked it better than the Singapore roadmap. Therapists care about the emotional equivalent of this analogy, when our mental models don’t adapt to reality, and worse, when we don’t even realize that we're using a map at all.

But – and here’s the provocation – if we think of ethical principles and codes as maps, could we see them as ways to overcome the gap between aspiration and behaviour? If principles and aspirations are encoded in these documents, however well or poorly intentioned, could they serve as ways to orient people towards living with AI? Rather than seeing them as opportunistic PR exercises, or ways to avert liability, could they play a role in mainstreaming and disseminating values? Through repetition, could we hack the current wisdom that ethics are expensive, that they’re a compliance burden, and that human rights hamper innovation? Could we retell these stories, share these maps widely, and improve the territory?

2. The power of exceptions: I’m a lawyer, I’m always more curious about exceptions than rules, and points of deviance rather than compliance. (Yes, I know, that actually sounds like I’m more the criminal than the lawyer!) I was struck by what the Berkman paper refers to as “The boldest departures from the standard notice-and-consent model”: the Chinese White Paper on AI Standardization and Indian AI strategy.

“As the paper describes, “…The Chinese document states that “the acquisition and informed consent of personal data in the context of AI should be redefined” and, among other recommendations, states “we should begin regulating the use of AI which could possibly be used to derive information which exceeds what citizens initially consented to be disclosed.” The Indian national strategy cautions against unknowing consent and recommends a mass-education and awareness campaign as a necessary component of implementing a consent principle in India.”

I’m fascinated by what these two points of departure reveal about cultural context and the penumbra of consent; about highly specific attributes of their societies, such as informational asymmetries, digital illiteracies, and power imbalances. It is possible to be highly cynical about these principles given contexts in which consent can be meaningless, forced, obtained under duress or conditions of great poverty and desperation, but enshrining them in principles might just be the first steps towards eventual accountability. When that accountability is to the two largest populations on the planet, perhaps it’s even more important that these principles work in tandem with legal and regulatory means to operationalize them.

3. Which brings me to my final observation: In a thought provoking piece in the MIT Newsletter, titled “How Not to Teach Ethics”, Prof. Susan Silbey challenges the stories we tell about ethics. She outlines the recent push to force ethics into the curriculum of engineering and science schools, following the increasing crises of corporate and professional responsibility. Such courses are emerging at a moment when big tech companies have also been struggling to handle the side effects of Silicon Valley's build-it-first mindset. Prof. Silbey critiques the fact that most of these courses focus on getting students to reflect on their personal choices. She writes:

"This cycle of scandal and responsive calls for better training has been so often repeated that one can be surprised only by the paucity of models for providing that education. The standard model – required in law and medical schools now leaking into engineering and computer science programs with minor variations – teaches ethics as problems in individual decision- making, personal values, and choices. Training focuses on formalized rules of professional conduct, punctuated by appeals for social responsibility. It has not proved to be a successful regimen, if the repeated cycles of corporate and professional misconduct are any gauge.

Thus, when asked to interpret or explain social phenomena, including professional misconduct or inattention to competing interests, historical examples and possible precedents, the well-educated technologist as well as the popular pundit will more often than not offer accounts that rely on individual agency, choice, and personality. Unable to recognize or describe forms of social organization, many adopt a rationalist, often reductionist model of social action that in effect constitutes a powerful and unreflexive orthodoxy.

Such standard models fail because the diagnosis and cure share a basic misconception: that corporate and professional misconduct are problems caused by rotten apples; some few weak, uninformed, or misguided individuals making independently poor choices. She goes on to describe how even massive lapses like at Enron and Cambridge Analytica are narrated as the story of a few rotten apples giving the barrel a bad name. She asks "If we talk about ethics as individual decision-making without history, context, social structure and culture, we have not explained how the organization of apples in the barrel is part of why we see only an occasional bad apple and how those bad apples can infect the other apples. What are the mechanisms of infection and spread? This is the missing structural element that conventional accounts of ethics as bad apples usually miss, and an alternative approach to ethical education and responsibility might offer. "

You should read the entire article, which is available online, but I offer that as a way of tying together the 7 Deadly Sins that Herbert outlined, as well as the questions that he posed at the end. Particularly to link individual decision making to the aggregated moral culture, and more importantly, to structural factors, such as incentives, taxes, institutional corruption, bailouts, etc. Seeing ethics as more than a matter of individual morality is something that cuts through this first session, as well as the need for different stories about ethics.

Thank you.