AI Ethics

-Tim Clancy S.J. September 30, 2025

1. No technology is ethically neutral. For any technology both empowers and entangles, creating new opportunities and hindering other, sometimes earlier, ways of doing things. This is true not only for a technology's users, but also for its designers, for its marketers, for its service providers, for government regulators, indeed for society as a whole. Everyone affected by a new technology also contributes to its use and dissemination with a degree of complicity commensurate with each one's role.
2. For all of us, in all these categories it is incumbent to ask ourselves not only (1) what the proposed technology can do ***for*** us but also (2) what it can do ***to*** us. Thus, AI Ethics includes not only normative questions of justice and fairness, risk assessment and cybersecurity, but also hermeneutical issues of meaning and identity. However in such a (3) complex, powerful and rapidly evolving environment, ***how*** to be ethical for any participant is not obvious.
3. Wisdom in the (1) invention, (2) application, (3) adoption and (4) dissemination of (5) innovative new technologies is too often retrospective. We learn from our mistakes. But these new technologies are so powerful, the pace of development and public release so accelerated, we are going to need prospective wisdom as well—that is, we will need to act wisely even before we know all the ways a technology will end up being used and what all their long-term consequences are likely to be. (6) One research group who collected over 700 AI risks from peer reviewed papers soberly noted that only 10% of those risks had been identified before the respective AI's release. How can we be more successfully proactive?
4. Historically there have been four dominant approaches to questions of morality and meaning. (1) Two are schools of normative ethics: (2) deontology, or principle, rule based ethics, and (3) consequentialism, where one optimizes benefits over cost, such as in utilitarianism. (4) A third, virtue ethics, approaches ethics (5) in terms of its impact on the character of those involved, (6) while a fourth, network ethics, attributes (7) agency and so distributes moral responsibility across the whole network of contributors, both human and machine.

5. Traditionally these moral approaches have been set in opposition to each other as rival methodologies, but I would argue in light of our limited understanding of AI's nature and potential impact,
6. we need to treat them as complementary, each asking a different set of questions, all of which ought to be on the minds of both designers and users, marketers and regulators. (1) Taken together the four approaches inform a whole field of questions to be addressed to any technology.
7. Let's start with deontology: "Deontology" comes from the Greek "*deon*" meaning "necessity," or in ethical terms, "moral obligation." It argues that there are moral principles that are universal and necessary, deducible from the very nature of rationality, freedom and human dignity. (1) Today these are often referred to as "human rights." They admit of no exception. That is, it does not matter what your justification is, it does not matter how significant the consequences are, if a course of action violates someone's human rights it is wrong. The ends do not justify the means.

   (2) Immanuel Kant is the most prominent formulator of this approach and so it is sometimes referred to as "Kantian ethics". (3) He bases moral principles on what he calls the "categorical imperative." He offers three formulations:

   (4) The first formulation focuses on the ***impartiality*** of an agent's motive or intention in acting: "Act only on that maxim (motive, intention) whereby you can at the same time will that it should become a universal law" In short, it's a form of the golden rule—treat your neighbor as yourself.

   (5) The second formulation focuses on ***human dignity***: ""Act in such a way that you treat humanity, whether in your own person or in the person of anyone affected by your action, never as a mere means to an end but always at the same time as an end in itself" That is, do not treat people like things. For Kant our dignity is grounded in thinking for ourselves and our ability to act on the basis of our own decisions. It is only in respecting one another's right to think for themselves that we treat each other with the dignity our humanity calls for.

   (6) The third formulation focuses on ***freedom:*** "Act as if one were legislating for all humanity, as a free citizen in a free society (a kingdom of ends)". For Kant one never acts for oneself alone, but always also as a member of society and

ultimately as a citizen legislating for humanity as a whole. Both users and programmers are complicit in the character of the society and world their technology is enabling, even creating.

8. What kinds of questions do normative principles of rationality, dignity and freedom raise for (1) data collection, (2) algorithm design and (3) cybersecurity?
9. First privacy. (1) Can one be truly free if one is under constant surveillance? Can you even think for yourself if you are constantly vulnerable to shame and punishment by others for thinking the wrong thing? (2) But on the other hand, does our right to think and choose for ourselves entail an absolute right to privacy? (3) And what does privacy even mean in the internet age? Does it mean all settings and cookies need to be "opt in"? (4) Does it mean that we "own: our data and so have a right to control all access to it, even if we do not own the platform that we upload that data onto, or the servers in which that data is stored? (5) Can we sell whatever right we have to the privacy of our data? Or would that be an abuse of our dignity even if voluntary, like selling oneself into slavery?
10. Similarly, what does thinking and choosing for oneself, that is informed consent, mean when an app's terms of service are too long and too complex for users to realistically read let alone understand? (1) Are users alone to be held responsible for what they agree to? (2) Do not marketers have a moral obligation to make a TOS that is understandable to its users without a degree in either the law or computer science? (3) Ought the government issue regulations to streamline and standardize terms of service like it does for nutrition labels on our food? Would that be fair to the developers thinking and creating "outside the box."
11. Bias in data sets, or fairness in the algorithmic processing of data however it is gathered, is also a clear application of Kantian morality. But bias can be subtle. Discrimination can occur just as effectively through proxies. And basing algorithms on past data may only cement the biases already baked into the data. Can diversity within design teams offset and complement each individual's inevitable blind spots. Impartiality is an ideal for everyone, not a reality for anyone.
12. And whose judgment matters? Who should count as a stakeholder whose opinions ought to be sought when designing a product? Just direct stakeholders, ie users and their customers? What about bystanders who have to live in a society where such tools are

available? But in that case, who isn't a stakeholder? But would obligating reaching out to all those affected make consulting stakeholders or "Democratic AI" too unwieldly to be ever realistic?

13. For example, a hiring algorithm may prevent parents with young children from getting interviews for a particular job. (1) But it may take being a parent to be alert to proxies. Parenting may be irrelevant to a job description, (2) still correlate with poorer attendance at work, due to the need to care for a sick child at the last minute. (3) It may correlate with less willingness to work weekends to ensure quality time with their kid or more resistance to moving to a new location due to the inevitable disruption to a child's schooling and social networks. Even if an application does not ask one's marital status it may address issues such as these that correlate with marital status, effectively screening out parents with young children.
14. Now as an employer using a hiring algorithm that disadvantages young parents, do I not have an obligation to know this effective "filter" before I start using the hiring algorithm? (1) I might consciously and deliberately intend a family friendly workplace! Is it morally responsible for me to wait for problems to appear, say until I notice the paucity of parents in my employ, before I have an obligation to critically examine my hiring algorithm for proxies? What else might I be missing unawares?
15. But on the other hand, is it realistic to expect an employer to even understand how their algorithm arrived at any particular output when AI algorithms today can have millions, even billions of parameters whose weights are constantly shifting in light of feedback from prior performance. (1) A printout of all an algorithm's variables and their ever-changing weights will not give me or any other human being an understandable, actionable explanation for a given output. I may well need another AI to model in humanly comprehensible terms how my hiring AI algorithm itself works.
16. If users have a responsibility to be critical users, rather than simply surrendering their judgment to the algorithms they use, do not the designers of these algorithms also have an obligation to render their reasoning comprehensible to their users? And do cybersecurity technicians not have a parallel responsibility to make their security software comprehensible to their clients, that customers know what is being done on their behalf to protect their data?

17. Or does the impossibility of ever fully understanding how an AI learning algorithm functions just go to show that calls for "Responsible AI" are hopelessly naïve? Or is it that Responsible AI simply cannot be based on a user's or a designer's or a regulator's intention, but must depend on something else? A second normative approach to ethics looks not directly at the motivation for a course of action but rather on its consequences. Afterall, isn't the road to Hell paved with good intentions?
18. But evaluating the consequences of an algorithm's output can end up being just as elusive as understanding the reasoning that produced that output. If a user cannot fully understand how their algorithm works how can that user assess its risks? What consequences ought a user or a programmer, or cybersecurity analyst to weigh in the first place? (1) Merely intended consequences seems too narrow.(2) The risk of unintended consequences can make a well-intentioned algorithm too dangerous to release into the wild. And even (3) beyond unintended consequences, what about longer term risks that are inherently unpredictable in advance? Or risks arising from interactions among multiple algorithms training on one another,(4) that may not even be conceivable at the time of a given algorithm's release? The 2008 stock market crash occurred in part due to highly complex learning algorithms trading in milliseconds in response to other similarly lightning-fast algorithmic trades. The trades were too many and too quick for any human to monitor, let alone evaluate, let alone respond in anything like real time, to how the algorithms were impacting the market.
19. Or consider the mere three-year history of generative AI. It's troubling how often designers are themselves surprised by how powerful their AI's end up being, or what uses they end up being put to by adopters exploring what all they can do with them.
20. When dealing with algorithms that we can neither fully monitor nor predict as they continue to learn and evolve, due diligence in risk assessment has gone beyond beta testing to "red teaming" prospective algorithms, to aggressively seek to find vulnerabilities designers never thought of, and offering "bug bounties" to enterprising hackers who discover such unknown vulnerabilities. When you don't know what you don't know, crowdsourcing the identification of vulnerabilities becomes a moral necessity. But then remember: only 10% of risks had been identified prior to the release of today's AI algorithms.

21. Furthermore, with the economic incentives to be first to market, particularly with a game changing new product, data scientists and cybersecurity analysts themselves are beginning to worry over programs generating potentially catastrophic, but in practice unforeseeable outcomes, even existential risks to the very survival of our species. When the combustion engine was invented no one thought it would prove a threat to human life on our planet. We have hopefully identified the threat of global warming in time to mitigate and ultimately reverse this risk to our species. But how much time will we have to respond to the next self-induced technological threat to our existence? What is your p(doom)?—your estimate for the probability of the release of some AI technology the ends up dooming our species?
22. Risks regarding the emergence of Artificial General Intelligence (AGI) has been a growing cause of concern in particular. AGI loosely means an AI that surpasses human intelligence not just in some particular domain (Artificial Narrow Intelligence), but across any and all domains of human intelligence. How assess risks that could be inconceivable by our more limited intelligence?
23. As a learning algorithm ever refining its own coding, can we ensure that any AGI will remain aligned with human values? How can we trust that they won't jailbreak any "guardrails" we program into them. How can we "keep them friendly?" Even if we cannot identify all concrete consequences of a new AI algorithm, can we at least reliably calibrate the category of risk a new app might generate?
24. On the other hand, Narayanan and Kapoor have recently argued in "AI As Normal Technology," that the pace of technological progress is much slower than often imagined. Rather than simply looking at the pace of innovation in assessing risk, (1) we need to distinguish four distinct stages in the emergence of any new technology. Each stage has its own time scale and opportunities for monitoring and remediation if necessary. They use the rise of electric technologies as a comparison. For this new and transformative technology the four stages were:
    a. (2) Innovation (Edison invented the light bulb in 1878)
    b. (3) Application (home appliances only begin to be marketed in the 1920's)
    c. (4) Adoption (beyond early adopters to the general public. Rural electrification was a New Deal project in the 30's.

d. (5) Dissemination. (its routine use in ordinary life or work, rather than even adopters only treating it as a novelty.
e. (6) When in this process does AGI become an existential risk? When it is invented on paper? Or when it is operationalized into apps? Or when those apps are adopted by the general public? Or when the use of AGI Apps is widely disseminated across everyday life and work?

25. There is also pushback of course from this attempt to calm the waters. Treating AI like past disruptive technologies ignore its unprecedented scope and epochal impact. New York Time opinion writer Thomas Friedman has compared AI to a diffuse vapor, that is coming to penetrate into everything. I have been arguing that the internet marks the next epoch in communication technologies, analogous to that of writing itself, from which we date the birth of "civilization."
26. We are entering "Civilization 2.0" Digital natives are crafting a new kind of identity, different from both the public, communal identity one finds in oral culture and the private autonomous individuality or "authenticity," prized in literate society. Today's identity is rather a network of partial identities or personae enacted on a wide variety of apps and websites, chat rooms and video games. This identity creates new psychological potentials and vulnerabilities, new inspirations and new threats to meaning and purpose.
27. The conundrums and uncertainties encountered by both Kantian and utilitarian moral methodologies should not lead us to scrap normative frameworks for assessing AI, but to use each to complement the other. Just as two heads are better than one, so two different ways of ethically assessing these new technologies are better than either one by itself. Diversity among ethical approaches is as necessary as diversity in our design teams. Users and programmers ought to attend to both the fairness and the comprehensibility of their algorithms, even if this is only achievable by approximation. And similarly we need to assess any algorithm's reliability and safety, again, even if certainty remains out of reach.
28. To put it in other words, with our limited intelligence we are not going to be able to morally evaluate AI algorithms algorithmically. Different algorithms optimize different values which are in tension with one another. For example, (1) privacy and transparency pull in different directions; so too (2) fairness and efficiency, even explicability may be in

tension, or (3) reproducibility and accountability. Kant had already recognized that (4) virtue and (5) happiness pull in (6) different directions at least in this life, deferring their coincidence to faith (7) in a hoped for eschatological future—(8) The Kingdom of God.

29. Virtue ethics raises yet another, different set of ethical questions, addressing the character of user experience and the workplace culture of programmers. However fair, comprehensible and reliable an app is, however safe its consequences, its impact on the life and character of its users could lead us to take it off the market.
30. When Eric Zuckerberg developed Facebook he could not see a downside to connecting people. But we have quickly discovered that social media can undermine self-esteem, particularly among impressionable teenagers. So too we have learned that digital media of all stripes can not only attract users but addict them, not only inform them but manipulate them. (1) Today the discussion is not over ***whether*** but ***how*** to regulate social media apps's impact on its users. The same questions apply to society itself. (2) Our very democracy is undergoing a stress test in today's media ecology of fake news, conspiracy theories and echo chambers.
31. Similar ethical concerns revolve around the impact of an algorithm on its designers. (1) At what point does a lack of work-life balance become immoral in a start-up culture?
32. At what point will programmers feel like "mere means" in pushing out product to meet ambitious deadlines?
33. Employers might ***consciously intend*** to care for their workers, but if the impact of job expectations ends up becoming inhumane do such intentions even matter?
34. In light of such risks should employers provide procedures for workers to give anonymous feedback on how aspects of their job is impacting their quality of life? Is a channel to provide such feedback inherent to a worker's dignity? (1) Similarly, what impact do non-disclosure agreements have on a worker's autonomy? (2) Is a worker's freedom truly respected if the only alternative to doing something he considers immoral in his job is to quit, especially if a non-compete clause means effectively quitting the profession? These are questions not only of justice and fairness but of meaning and identity
35. And finally, what about the character of the AI's themselves? The rejection of technological neutrality entails that AI's themselves can be good or bad, irrespective of

how they might be used. (1) In other words, some imaginable AI's may be too dangerous to ever be released to the public. What affordances, or "virtues" should a given AI embody? What "vices" should we be sure to avoid inadvertently programming into it? (2) For example, is the optimization of any one single value at the cost of all others itself a danger or "vice" irrespective of how it is actually used in normal practice? What affordances/virtues could compensate for, or at least mitigate such dangers inherent in the algorithm? Can we ensure that even when we can no longer understand how a given AI algorithm reasons, we can still "trust" it?

36. Given that AI's evolve through their learning programs, ought there be an ethic for how we work with AI, how we treat it, such that we can trust it always working ethically towards us? Would our modelling respect towards it, enable us to trust it to respect us? Not respecting it as valuable in its own right, irrespective of its instrumental value in achieving our goals, could itself be risky. Risky that it might respond in kind, not recognizing us as having any value outside the end it is optimizing.
37. And finally from a network perspective, degrees of agency and so responsibility are distributed across all parties involved in a given activity both humans and devices, not only in making and marketing an app, but also in running and maintaining it.
38. For example, you cannot have social media without content moderators. They contribute to the appeal, even the very functionality of the app. (1) But content moderators often burn out from PTSD from noxious posts they are constantly assessing. .Surely the impact on content moderator's mental health bears on the morality of the apps they monitor.
39. But what's the alternative to content moderators? Can AI automate at least a social media's app most egregious content? Until then, is content moderation like slavery, a necessary evil until its not?
40. Or consider the work required to maintain and service an app. At what point does a quota system for evaluating service performance become inhumane?
41. Or to what extent are programmers responsible for how their product gets marketed? They are not deciding on ad copy, but they may know how hyperbolic some of the claims are that their company makes for an app they designed. Are they complicit in the potential fraud? Or does such potential complicity simply come with the job? Another "necessary evil?"

42. Network effects can also be global. How ethically evaluate an app's environmental sustainability? Perhaps Artificial General Intelligence should not be developed not because of existential risks to the species but simply due to the energy consumption it would require to build and run it. (1) It's been estimated that electricity consumption at AI Data centers will increase four-fold in the next five years. Is that sustainable without overheating the planet?
43. The ethical issues around AI technologies are so complex and our knowledge so limited that neither external regulation nor the character formation of designers and users can be trusted alone. If regulation is only seen as an external constraint on innovation, creative developers can be counted on to find workarounds to any regulation, users can be counted on to jailbreak any guardrail, and hackers can be expected to crack any security algorithm. Character will always matter. .But on the other hand, any character, however smart and admirable, cannot know what he or she cannot know in advance of how users will play with a novel technology or its downstream consequences far into the future. Nor can anyone know the synergies enabled through a technology's use in conjunction with future apps not yet even on the drawing board. Just as we need a diversity of ethical methodologies to complement one another and compensate for each other, so too with regulation and character. With such a diversity of values at stake, AI Ethics will always remain as much an art as a science, negotiating and triangulating conflicting values, proposing courses of action that can at best hope to minimize and mitigate a new technology's hazards and curses while also enabling and promoting their utility and promise. Like the quest for wisdom itself, AI ethics will always remain an infinite task.

Suggestions for further reading: