IRIE
International Review of Information Ethics Vol. 28 (06/2020)

Jonathan Cohn

In A Different Code: Artificial Intelligence and The Ethics of Care

Abstract:

The following essay explores the intersection of *care* with ethical reflections on artificial intelligence (AI). The current debate around AI ethics focuses on questions of moral AI judgment and the general criteria for maximizing the fairness, accountability, and transparency of these judgments. While this discussion is important, it all too often obfuscates the actual purpose and intention behind the use of the algorithmic or AI technology. Where the rationale for developing these technologies focuses on increasing optimization and innovation, concern must be shifted to ensure that AI will be used primarily to address current inequities and harms, from exacerbating climate change to manipulating voters via social media to creating "better" weapons.

Keywords:

Artificial Intelligence, Care, Empathy, Ethics, Relationship, Technology

Outline:

Author(s):

Jonathan Cohn:

- ☎ number, ✉ cohn@ualberta.ca, 💻 website

ISSN 1614-1687

1. Introduction

This essay explores what happens when we place care at the center of our ethical reflections on artificial intelligence (AI). Presently, much of the debate around AI ethics focuses on questions of what moral AI judgment consists of and how to set up general criteria for maximizing the fairness, accountability, and transparency of these judgments. This discussion is certainly vital, but as Os Keyes et. al. argue, it too often obfuscates the actual purpose and intention behind the use of the algorithmic or AI technology; i.e. if you are using AI to decide how to most efficiently detain migrant children away from their parents, does the level of fairness, accountability, and transparency really matter? (Keyes et al.) So much of the rationale for developing these technologies focus on increasing scalability/optimization (i.e. efficiency) and innovation, such as finding more efficient ways to accomplish a task (Jiminez). The concern that AI will be used primarily to increase current inequities and for other unethical purposes from exacerbating climate change to manipulating voters via social media to creating "better" weapons are real and justified.

That said, for me, the question of how AI can be ethical in its judgments is relatively simple and straightforward. Following John Rawls' *Theory of Justice* (with a particular eye to his "difference principle"), if you are using AI to level the playing field and otherwise advantage the worst off, you are being ethical (Rawls). If you are not, then regardless of your AI's fairness, transparency, and accountability measures, your result will be at best amoral and at worst, completely unethical. For example, using AI to alert people of government funding and help they may be eligible for is ethical; using AI to relieve traffic congestion for suburbanites is more suspect (Baker). In that case, an ethical design choice would be to privilege emergency vehicles, public transportation, and pedestrians over all other vehicles in the hopes of helping the least well off and discouraging wasteful and polluting lifestyles; ironically, while AI is currently used to increase the speed of commuters, it may actually be wiser to use AI to disincentivize driving by slowing them down.

However, using AI to make blanket judgments about what is or is not ethical and who should be advantaged (or not run over) is extremely problematic. How many of those cars I am slowing down are full of precarious laborers or disabled people headed somewhere public transportation does not go? How can AI be employed in a way that does not inflict collateral damage to the least amongst us? With this question in mind, the principles behind the ethics of care (EoC) may provide the best model for creating and using AI for the betterment of those who need the most help. In contrast to most ethical approaches that center on whether certain generalizable acts and judgments are ethical or not, the EoC fundamentally opposes relying on abstract ideals and rationality as a basis for ethical decisions. Instead, it focuses on the specific context of each problem and the creation of caring relationships as the basis of a moral and ethical existence (Gilligan). In this system, the foundation of EoC is the principle that one must care for others by responding to their needs and helping as best they can.

In the following pages, I will address the potentials and limitations of programming AI to follow the EoC. I am interested in how AI can nurture, deepen and expand our relationship to each other and the world around us and how this process necessitates that we extend care not only through AI but also to it. In other words, what does it mean to be in an ethical relationship to AI itself and how can simply asking this question help us to become more ethical in general?

2. Ethics of Care (EoC)

In practice, Joan Tronto argues that caring consists first of being attentive and open to the needs of others; and second, of being responsive and taking responsibility for competently attending to these needs (Tronto). Caring does not imply a formal obligation, nor does it require reciprocity or even empathy. Rather, one must care in order to better understand and respond to the inequities and vulnerabilities that surround us and hopefully eventually make caring a habit. Under such a framework, evaluating or judging whether

another's problem is worthy of being cared for is as unethical as being inattentive, incompetent or unresponsive to their pain.

Carol Gilligan originally theorized the EoC as a response to morality, justice and judgment-centric ethical systems that too often privileged stereotypically masculine values of competition, individualism, and rationality (Gilligan). As many feminists and EoC theorists argue, imagining universal principles inevitably privileges the beliefs of those in power over the least among us (Noddings). Gilligan argued that much of the study of ethics (with a particular focus on the work of Lawrence Kohlberg) emphasizes the value of justice over creating relationships and caring for one another. Traditional ethics conundrums often frame justice in opposition to caring relationships and suggest being just is always more important than maintaining the relationship. In the process, this ethics overly values stereotypically masculine values over stereotypically feminine ones. Kohlberg's Stages of Moral Development went so far as to consider most women less ethical than men because they tended to value fostering relationships over justice. In response, Gilligan challenged this moral hierarchy and explained why an expansive sense of care—not justice—is the center of all ethical action. The EoC does not recognize the justice/care binary but rather focuses on how caring relationships will necessarily lead to a more humane form of justice (Taylor). Since Gilligan's early work, there has been an active debate over the gendered and/or feminist aspects of care and how distinct this model is from other more traditional ways of imagining ethics.

I bring this history up because it reminds me of current debates over what the ethics of AI should be. Just as Gilligan was reacting to a perceived focus in philosophy on abstract universalistic reasoning over the specificity of individual problems and relationships, I am here questioning the assumption that AI should create or instrumentalize a generalizable principle of ethical justice and universal objectivity at all (AI: Algorithms and Justice). As in every other case where ostensibly well-meaning philosophers have attempted this, from Confucius to Aristotle to Kant and onward, the result has celebrated the values, beliefs, and well-being of the most advantaged over everyone else (Ledzion). Presently, this tendency is most obvious in the popularity of the ethics conundrum: who should self-driving cars injure or kill in a collision? This question, with its focus on imagining universal rules for automated technologies and the conflation of ethical and actuarial value of different lives if anything illustrates our dominating desire for easily coded and clear-cut answers to fundamentally unanswerable vague ethical dilemmas (Maxmen). Starting from an EoC instead of justice-based ethics would begin from the premise that you must create a vehicle that would never be in this type of life or death situation (by adjusting their speed limit and investing far more on safety mechanisms than at first may appear necessary).

Furthermore, a care-centered AI would seek to help people form deeper relationships and better care for each other rather than simply attempt to replace these relationships so we no longer feel obligated to care. The self-driving car example suggests that many want the car to make life or death decisions so that we don't have to, but this type of AI infantilizes passengers/drivers by presenting them as not only unwilling but also entirely incapable of caring for those pedestrians and other drivers that surrounds them. Just as Cathy O'Neil has argued that algorithmic technologies often get used to make decisions that humans do not want to grapple with, AI too often is presented as a quick fix for getting around deeply problematic issues (who should I run over?) and impossible questions (i.e. what makes a good teacher?) (O'Neil). Instead, a care-centric AI would make us aware of the complexity of these situations and illustrate that there is no clear (ethical or otherwise) correct answer in order to make us better recognize our obligation to responsibly help others. i.e. It should make us think more deeply about the difficulties we face to better cope and care for ourselves and others.

When we try to use AI to avoid our problems, we might forestall them, but we also exacerbate them. Take for instance the continual desire to one day use AI robots to care for the sick and elderly. Articles describing this future use of AI continually imagine a dystopian future where bedraggled elderly drastically outnumber "functional" humans (Sanyal). In this future, films and news articles continually imagine that relationships with AI robots will become suitable replacements for relationships with other humans. To be clear, there is nothing wrong with being cared for or caring for a technology; expanding the realm of what we can and do care for—from humans to animals to technology to the world—is deeply moral even if it is often difficult

if not impossible to know how to best express this care. But at the same time, one must not use technology as an excuse to *not* care; at that point, technology becomes merely a fetishistic proxy for kinship and connection (Harvey).

While various media typically represent AI as a tool to move the elderly entirely from view there is no reason we couldn't instead use this technology to instead bring them and others who need care into the center of our culture. Indeed, AI technology should be quite adept at addressing the personal needs of individuals in ways that make us all better at caring. Most algorithmic technologies follow relatively strict and universal if/else style rules that tend to treat people and their problems in relatively generic ways: i.e., too often they crudely categorize people and then give out general advice that may work for some but certainly not everyone. In contrast, AI technologies are often designed to get around the strict rule-based logic of most algorithms (in both how they cluster people and how they decide what advice to give) and should, therefore, be more capable of responding to individual needs in a personalized manner. AI is, to some extent already sold using EoC-styled rhetoric as a response to an overly static, oppressive and racist sense of judgment explicit in standard algorithmic technologies (Noble). While current AI may not be great at this personalized care yet, this is the kind of innovation we should be encouraging.

But there are many hurdles in the way largely created not by the technology itself, but by the capitalist economic system that it enables. As in the car and robotic nurse examples, AI is too often presented as a tool for creating objective models and universal easy answers to complex problems. While AI (and algorithmic technologies more generally) are often now used to "personalize" user interfaces and experiences, this term is simply doublespeak that facilitates capitalism by leading people to purchase more products and view more advertisements more efficiently. How different would your Amazon or Netflix recommendations be if they privileged products and entertainment to make you care for others more deeply rather than to feed your libidinal desires? We must push for AI technologies that help us become more aware and more capable to address the needs of our communities rather than just our own desires. EoC offers a promising structure for getting us to this reality that we so desperately need.

At the same time, to truly incorporate an EoC into the design and use of AI, we must not only consider how we can best care for others *through* AI but also how we can care *for* AI itself. As in the example of robotic care workers for the elderly, we are more comfortable with imagining AI caring for us than us for it. Indeed, these representations encourage us to imagine these technologies primarily as throwaway commodities in need of continual upgrades and updates. The question of whether humans may one day care for—if not deeply love—AI is continually framed as a terrible fear and sign of a prurient mental illness or depression in films and other entertainment media. Some films like *Ex Machina* (Garland, 2015) or *Her* (Jonze, 2013) present these as prurient master-slave relationships while others including various *Black Mirror* episodes and *Blade Runner 2049* (Villeneuve, 2017) present them as caring but ultimately unfulfilling and impossible.

But importantly, the anxiety in all of these examples is less in the idea that one might deeply feel for AI, but rather that this relationship will fetishistically replace the relationship between humans, making it harder for us to care for each other. These representations ironically present caring for AI and other technologies in opposition to EoC. Yet, this does not have to be the case and I'd argue it is imperative for us to better care for AI in a way that extends the limits of what and how we care rather than simply replaces human relationships for technological ones. There is no reason the EoC need only apply to humans, animals, or even the animate. Indeed, given the problems of planned obsolescence, child labour, organized crime and warfare surrounding the mining and recycling of rare earth minerals, and the extremely high energy needs and the rapid destruction of ecosystems that this all causes, caring for technology is essential to caring for each other and our world.

At the same time, what would it mean to be in an ethical caring relationship with AI? How could we possibly know what AI needs or wants from us? Do we owe our technologies anything at all? It is as easy to imagine what AI wants from us as it is to imagine what a rock wants from us. We may imagine that like a rock, AI does not want to be ground down to dust or otherwise destroyed, but just like a rock, AI likely

doesn't care one way or another whether it is obliterated or not; there is no way for humans to know. Indeed, recognizing our limitations here along with the fundamental otherness and specificity of individual AI programs is perhaps the only way to be in an ethical and caring relationship with AI.

3. EoC and the Face-to-Face

For Levinas and EoC theorists, ethics itself is "a face-to-face encounter with a specific, irreplaceable other" (Taylor, 218). To be ethical in this paradigm, one must pay attention to the uniqueness and differences of others. Scholars and journalists alike continually treat AI unethically by not even attempting to do the bare minimum of attending to what AI is, how it is distinct from us and therefore what different needs it has; instead, AI is continually imagined as either humans, but better, or humans, but worse. Either way, these descriptions both simplify what it means to be human and what it currently means to be artificially intelligent. For example, arguments concerning whether AI is "intelligent" engage in unnecessary human-centric (and neoliberal) judgments over what intelligence consists of and severely limit the range of what can be considered intelligence. Often, these representations are used to bolster a particularly harmful representation of intelligence associated with a western imperial sense of rationalism. I.e., when we equate intelligence with efficiency and a hyper-individualistic desire to become completely autonomous, we are also arguing that those who do not support these values are unintelligent and expendable. This discourse does not recognize the vast diversity in how people think and express intelligence and is used to support only the most macho definitions of intelligence. Discourses that simplify AI by comparing it to humans can also be quite well-meaning, especially when they express a desire to not harm, take advantage of, or be violent toward technologies. For example, in Steven Spielberg's *AI* (2001), treating robots as prostitutes and destroying them for human entertainment makes this future society appear deeply immoral. This narrative and others like it discuss AI ethics as a metaphor for how to treat all humans and other living things. While noble, in the process it unfortunately also suggests that we should care for AI in the same way we care for humans when we cannot know what AI wants and needs.

This reductive logic concerning how we define AI has led to certain unfortunate decisions in how AI is currently programmed. Evolutionary algorithms, tensors, and other basic building blocks of neural networks, machine learning, and AI begin by generating a wide range of solutions to a particular problem, which then gets whittled down until the "correct" answer is generated based on test data; then, whichever methods resulted in generating this correct answer are kept, while all other methods are discarded. For example, if you scribble a 5 onto a piece of paper and then train an AI or machine learning program to identify that 5, the program will preserve any methods that resulted in identifying the scribble as a 5 and eliminate any that did not. Most of the time, such an action is hardly controversial, but what happens when the same basic principle is used to identify a person's race or sexuality? At that point, the programming, based on the "survival of the fittest," is not based on natural selection, but is instead more clearly based on social Darwinism and eugenic principles. While Darwin celebrated and marvelled at the diversity and variations of individuals and species, eugenic AI is entirely interested in eliminating variation in order to generate the ideologically agreed upon correct answer. In the process, we negate the possibility that there may be equally if not better ways of interpreting the world around us.

These design choices have led to an AI that far too often works on a logic of homophily to preserve the status quo and a sense of white bourgeois universalism (Hui Kyon Chun). Under such a paradigm, AI is utterly incapable of imagining a more just society and ethics that might generate it. This protection of current problematic social dynamics and beliefs is baked into modern conceptions of AI and is central to how it is sold. In other words, injustice and inequity is AI's feature, not a bug. It follows a long history of imagining that the key value of big data, algorithmic technologies, and AI is in maximizing the exploitation of workers and users alike.

4. Conclusion

If we start our discussion and programming of AI from an EoC perspective (and with a neurodiversity lens), we could recognize AI not simply as intelligent or not, but rather as representing a different type of intelligence, neither better nor worse than our own: just different. This approach has many benefits. Rather than designing generalizable AI with a one-size-fits all mentality and a universal white male subject in mind, we may push instead to imagine how different users and unique problems require equally differentiated and unique AI coding. As this design process would necessarily start with specific care-related goals in mind like creating more equitable labour conditions rather than general AI engines, this type of AI could be more straightforward and energy-efficient. In the process, by embracing a larger continuum of intelligences, it may also lead us to discover innovative approaches for these dilemmas (i.e. what AI is designed to accomplish but rarely does). We can use AI to better care for each other, but to do so, we first need to consider how to respect and care for AI itself.

5. References

"AI: Algorithms and Justice." *Berkman Klein Center*, March 28, 2019. https://cyber.harvard.edu/projects/ai-algorithms-and-justice.

Baker, Francesca. "The Technology That Could End Traffic Jams." *BBC.com*, December 12, 2018. http://www.bbc.com/future/story/20181212-can-artificial-intelligence-end-traffic-jams.

Gilligan, Carol. "In a Different Voice: Psychological Theory and Women's Development". Cambridge: Harvard University Press, 2016.

Harvey, David. "The Fetish of Technology: Causes and Consequences. Prometheus's Bequest". *Technology and Change* 13 (2003): 3–30.

Hui Kyon Chun, Wendy. "Queering Homophily." In *Pattern Discrimination*. Lünenberg, Germany: meson press, 2018. https://meson.press/books/pattern-discrimination/.

Jiminez, Javier. "5 Ways Artificial Intelligence Can Boost Productivity." *IndustryWeek*, May 22, 2018. https://www.industryweek.com/technology-and-iiot/5-ways-artificial-intelligence-can-boost-productivity.

Keyes, Os, Jevan Hutson, and Meredith Durbin. "A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry." In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, alt06:1–alt06:11. CHI EA '19. New York, NY, USA: ACM, 2019. https://doi.org/10.1145/3290607.3310433.

Ledzion, Michael. "Justice System Is Weighted in Favour of the Rich." *Financial Times*, October 31, 2018. https://www.ft.com/content/6064efde-dc46-11e8-8f50-cbae5495d92b.

Maxmen, Amy. "Self-Driving Car Dilemmas Reveal That Moral Choices Are Not Universal." Nature, October 24, 2018. https://www.nature.com/articles/d41586-018-07135-0.

Noble, Safiya Umoja. "Algorithms of Oppression: How Search Engines Reinforce Racism". NYU Press, 2018.

Noddings, Nel. "Caring: A Feminine Approach to Ethics and Moral Education". Reprint edition. Berkeley: University of California Press, 1986.

O'Neil, Cathy. "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy". Reprint edition. New York: Broadway Books, 2017.

Rawls, John. "A Theory of Justice". Cambridge: Belknap Press, 1999.

Sanyal, Shourjya. "How Is AI Revolutionizing Elderly Care." *Forbes*, October 31, 2018. https://www.forbes.com/sites/shourjyasanyal/2018/10/31/how-is-ai-revolutionizing-elderly-care/#2d41e0fae07d.

Taylor, Chloé. "Lévinasian Ethics And Feminist Ethics of Care." *Symposium Journal*, n.d.

Tronto, Joan. Moral Boundaries. New York: Routledge, 1993.