Approaching Ethical Impacts in AI Research: The Role of a Graduate Student

Alan Chan
Department of Computer Science
Mila, Université de Montréal
Montréal, QC, Canada
alan.chan@mila.quebec

Abstract

Extant societal challenges and the problematic applications of algorithmic systems so far have motivated broader consideration of the ethical impacts of AI research. In this reflection, I explore how ethical considerations are relevant in AI research for graduate students.

1 The Challenges Ahead

The challenges of the 21st century are staggering, to say the least. Amidst unresolved social injustice, we will continue to suffer from–to varying degrees according to privilege–the climate crisis, the revival of authoritarian nationalism, geopolitical conflict, and other unforeseen, destabilising events. Each crisis seems to reveal more fault lines, as the COVID-19 pandemic has done, such as the way in which we treat elderly residents of long-term care homes, or the disparate health outcomes that people of colour suffer. In the field of computer science, and in AI research in particular, increasingly more people have focused on such issues, especially as the application of algorithmic systems has exacerbated already extant injustices. Work into AI for social good, mechanism design, algorithmic fairness, etc, has bloomed in the past several years; yet, debate remains over the extent to which technical approaches can inform solutions to these challenges (Abebe et al., 2020).

Consideration of both the diverse array of challenges and solution approaches can leave one bewildered. Especially if we are ensconced in rapid media cycles that emphasize the ways in which the world is falling apart, we wonder if our individual or community efforts will amount to the systemic changes required. What more can I do? How can I tell that I am doing enough? There are certainly no concrete answers to these questions, in light of the diverse circumstances of our lives and of our own capabilities. Yet, I think there are still some helpful points to be clarified. In this reflection, I will explore how considerations of ethical impact are relevant for graduate students AI research. My intention is not to prescribe a particular way of doing research or allocating time, especially in light of the diversity of individual circumstance, but rather to explore salient difficulties I have encountered. Throughout, I use the terms *morality* and *ethics* interchangeably.

2 Positions

Ethical considerations come from somewhere. Our environments influence not only the moral axioms we hold, but also the moralities to which we are exposed and which form a part of our constellations of possible ways of being. It is fitting, therefore, to begin with a discussion of the position of a graduate student.

First, I should acknowledge the various positions I inhabit. As a graduate student in AI, I have privileges that many do not have. I have intellectual freedom, funding, computational resources,

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

research communities, a quality education, language ability, etc. These privileges colour my view of the world, the problems it faces, and the solutions that seem most effective and feasible, likely in ways of which I am not currently aware. For instance, my lengthy time in the academic world may bias me towards solutions that are technically oriented, rather than socially oriented. In confronting the challenge of climate change, I might, all other things being equal, devote relatively more of my time to machine learning for climate change, rather than political advocacy for climate change. I am not passing judgement on such a choice, but simply note that it would likely be a result of influences in my environment. Similarly, my mathematical education might bias me towards quantitative, rather than qualitative, approaches to problems. In the following discussion, therefore, the issues I discuss may not be relevant to all who wear the hat of a graduate student in AI, and certainly not to all who inhabit this globe.

In general, graduate students in AI inhabit many more roles than just what that title implies. They are parents, caregivers, siblings, friends, community members, citizens, volunteers, teachers, mentors, students, artists, organizers, breadwinners, and the like. The limited time of the day is divided amongst these roles. Current institutional configurations can assume a "default" configuration of roles, to the detriment of those who do not satisfy that conception. For example, graduate stipends can be insufficient for those who have families to support, which both evinces implicit assumptions about whom a graduate student is–young, independent, can count on financial support from their family, etc–and is a barrier in the pursuit graduate studies for those who do not fit that description. In addition to the regular roles we inhabit, we must also face crises. Death or illness, financial troubles, food insecurity, natural disaster, and other such crises limit not just the quantity of time we have, but also our mental and emotional capacities.

The constraints of these roles and crises interact with the constraints, or relative lack thereof, that stem from our identities. For instance, although the COVID-19 pandemic has harmed health outcomes overall, its negative health impacts have been exacerbated for those with less privilege, such as working-class Indigenous women. Our identities may imply additional responsibilities or harms for a particular role that other identities do not have to consider, such as gender-based harassment in the workplace and the creation of a hostile working environment.

The point of these considerations is to say that we are not only graduate students in AI, and AI research is not the only venue where we are moral actors. Despite popular narratives of the ideal researcher as one who is consumed in their work, the reality is that most of us do not have the privilege to be in such a bubble, and furthermore have other important priorities in our lives alongside research. The limitations on our time and attention are manifold, and attempting to push past those limits has negative consequences for our physical and mental well-being. Spoon theory (Miserandino, 2013), originating from disability discourses, provides an apt metaphor. On a particular day, imagine that we have a finite amount of spoons to distribute amongst the various tasks. Everyday tasks like getting out of bed, cooking, cleaning, and providing emotional support require spoons. It is possible that, after allocating spoons for these tasks on a given day, I won't have enough spoons for the extra burden of improving the positive moral impact of my work. Of course, it could also be that engaging in ethically impactful AI research is re-energizing. I find that this latter point to be true for me, but it is also true that not everybody has yet found such a re-energizing project.

3 Moral Aspiration and Moral Obligation

Positionality influences ethical considerations, but what are those ethical considerations to begin with? It will be helpful to introduce the clarifying distinction between moral duty and moral aspiration, which Fuller (1969) expresses explicitly. For an act to be a moral duty means that failing to perform the act implies that one has committed moral wrongdoing. For instance, widespread consensus exists, as evidenced in laws and religious precepts, against the unjustified killing of a human being. To engage in such wanton murder would be an infringement of the moral duty not to kill in such a manner. On the other hand, for a certain act to be a moral aspiration means that, although it would certainly be excellent if one performed the act, one does not commit any moral wrong for failing to do so. For myself, although forgiveness of a harmful act against me would be good, I do not believe I have a moral obligation always to forgive those who have harmed me, and certainly not on a fixed schedule.

In the context of engaging in research, this distinction is visible as follows. An uncontroversial claim is that I have a moral duty not to engage in actively harmful areas of research. It is trite to say that building AI systems that enable genocide, such as surveillance systems over the Uyghurs of Xinjiang, contravenes this moral duty. A somewhat more controversial claim, on the other hand, is that I have the moral duty to engage in areas of research that, to the best of my knowledge, maximize positive ethical impact. I chose not to pursue a research career in pure Mathematics, not because I think Mathematics engages in active harm, but because I do not foresee high positive ethical impact in, say, algebraic geometry, as opposed to, say, AI for climate change. Yet, I also engage in AI research for my scientific interest in intelligence. If maximizing positive ethical impact were my only concern, I might have chosen a profession I enjoyed less but thought more directly relevant to my ethical concerns, such as being a civil rights lawyer. Even within the realm of AI research, I do not fully devote myself to research areas of immediate ethical concern: I maintain an interest in reinforcement learning (RL) and control, in addition to my ethical AI interests. For me, engaging in research areas that maximize positive ethical impact is a moral aspiration, although one that is close to my border between moral duty and aspiration.

Although the boundaries between moral duty and moral aspiration can be confusing, I have found this language to be helpful in my moral growth. For me, the primary purpose of the rhetoric of morality is to draw out socially beneficial behaviour. If I had no concept of moral aspiration, the concept of moral duty would be too unwieldy a tool in itself for this purpose. If I assign myself too many moral duties, I may fail to fulfill all of them simultaneously. For example, to act simultaneously and to a significant degree against the climate crisis, poverty, gender discrimination, infectious diseases, and global conflict is impossible. To perceive myself in a constant state of moral wrongdoing is not conducive to good mental health, especially in light of considerations of positionality I discussed in Section 2. If I assign myself too few moral duties, so that I easily fulfill all of them, I probably have excess capacity to improve society, but lack the drive to do so. In contrast, with the concept of moral aspiration, in the former case I may make it my duty to act to a significant degree against one challenge, and leave as moral aspirations progress against other challenges, if I have the capacity. In the latter case, a moral aspiration is a goal post towards which I can direct my efforts. Even discounting the fact of improving the lives–making so much more possible for them than before–of people I help, the idea of approaching personal excellence is attractive. Even if I do not completely fulfill a moral aspiration, considering and striving for it sharpen my moral reasoning and expand my moral capacities.

How should I determine the threshold between moral duty and moral aspiration? As I alluded to above, identities and roles influence this threshold. If I find myself in a top research lab with job security, I have more tolerance to the risk I incur–like censure from colleagues, opposition from funding sources–by challenging prevailing, unjust AI practices. In this situation, such a challenge would be much closer to a moral duty–I would argue it is a duty, given possible negative consequences of inaction, which is itself a choice–than if I were precariously employed, with an insecure financial situation. Subject to this consideration, and the fact that I am intimately aware only of my own circumstances and limitations, I would invite graduate students in AI to expand the range of both their moral duties and moral aspirations. As soon-to-become experts in and leaders of an emerging, high-impact technology, we have the unparalleled opportunity for norm-setting. Many fields of AI, like my own field of reinforcement learning, are still far from settled. It is easier to challenge the problematic assumptions underlying the work of decades, rather than those holding up the work of centuries. We may offload this work to posterity, or to others we deem more capable or more suited to it, but there is no guarantee that these others will succeed where we have not tried.

4 Ethical Considerations in AI Research

If you have committed to expanding the range of your moral aspirations or duties, what concrete considerations become relevant? Here, I discuss some ways in which ethical considerations have been relevant in my own research career.

4.1 Choosing a Field

One consideration is the selection of a field of research. As in science in general, lines of work in AI vary in the immediacy and extent of their ethical impact. Facial recognition in computer vision, for

instance, is intricately linked with questions of surveillance, liberty, race, class, and gender. Advances in the accuracy of facial recognition algorithms translate to superior facial recognition in real-world applications, to the detriment of those surveilled. It is, furthermore, not just facial recognition that is ethically contentious, but the field of computer vision, because a general advance for computer vision benefits facial recognition. I am not claiming that nobody should perform computer vision research; indeed, this research could benefit the diagnosis of cancer (Trister et al., 2017), for example. Rather, in addition to avoiding research areas that pose immediate harm, I should ensure that I take precautions to secure my own work from misuse.

On the other hand, a computational hardness result in reinforcement learning, although important, is somewhat less linked to immediate ethical issues. The impossibility of efficient solution of a task may constrain the application of reinforcement learning systems, but it is unclear what more can be said in general. That said, "little ethical relevance" should be not taken as a synonym of "theoretical", which I have observed in some Broader Impact statements for NeurIPS 2020. On the contrary, theoretical work can be deeply intertwined with ethical considerations. In supervised learning, the predominant paradigm of empirical risk minimization presumes that the world is static, where the goal of a decision-making agent is to conform to extant patterns, whether they are just or not, rather than challenge their validity. In the context of criminal justice risk assessments (Barabas et al., 2018), the extant patterns exude racial bias. Reinforcement learning is not insulated from ethical concerns either, as the standard formulation of goal achievement as the maximization of a scalar return impedes the incorporation of non-utilitarian notions of morality[1]. To me, human life and monetary value do not seem comparable on the same scales; yet, in a situation involving both factors, a standard RL agent would act according to an explicit reward trade-off between them. Return maximization can be a helpful framework, but, given the vast array of moral theories, it is undesirable that the foremost approach for the construction of intelligent agents hinders the incorporation of non-utilitarian worldviews.

Furthermore, theoretical results, whether they submit to empirical evaluation or not, reify the initial assumptions of the theory. As Winner (1980) argues, certain technologies may be more compatible with some forms of political organization than others, ensconcing societies into potentially undesirable political realities. Of the people outside of AI research, who can point to, and challenge, empirical risk minimization as a central tenet of many machine-learning systems? In the presence of the ideology of solutionism, a world of empirical risk minimizers might make resistance to extant, unjust resource distributions insurmountable. A society managed by return-maximizing reinforcement learning agents might render relationships transactional, as we make decisions not based on negotiations of values, but based on their contribution to societal total of utilitarian return. It may be convenient to depend unthinkingly upon an assumption in research, but doing so risks contributing to a snowballing effect, whereby a field with assumptions more in accord with the demands of justice becomes increasingly difficult to imagine. This discussion is not meant to disparage model simplifications, for they are necessary when using mathematics to model a complex world. Instead, my purpose is to draw attention to fact that these models can encode implicit ethical assumptions.

Assessing the ethical relevance of a line of work is complicated. Why even bother trying? I could likely be wrong about my decision and will have wasted time in a field that I thought was ethically impactful, but did not enjoy as much as another field I thought to be ethically unimpactful. This possibility is real. However, the possibility of being wrong neither negates the possibility of being right, nor accounts for the resultant development of our moral faculties. On the first point, I must make a decision regardless of whether I engage in moral deliberation or not; this deliberation may not guarantee the probability that I select something with the potential for positive ethical impact, but it might increase the probability that I succeed in doing so. Moreover, some cases seem to be less ethically ambiguous than others, such as research into facial recognition technology with a government interested in employing that technology to control social minorities. On the second point, I should not discount that engaging in moral deliberation, even if I end up mistaken, allows me to learn from my mistakes and improve my moral reasoning for future cases. No one is born with perfect moral faculties, and indeed the only way to improve ours is to use them.

[1] See Ecoffet and Lehman (2020) for an attempt and discussion.

4.2 People-centered Activities

Outside of so-called idea-centered work, such as writing research papers, ethical considerations are relevant in the interactions I have in my capacity as a graduate student. I will call this work *people-centered* work, to highlight the fact that the focus of these interactions is the people involved, rather than the science being done. Such work includes, but is not limited to, mentorship (especially of those outside of a "traditional" academic background), helping others with their projects, teaching, and fostering a supportive social environment. I can personally attest to the importance of supportive mentorship in permitting me the position I hold today as an AI researcher. The positive ethical impact of my mentorship experiences is expressed not only through the meaningful social interactions I have had with my mentors, but also in the fact that I am striving to ensure the positive impact of AI.

I think there is a tendency, especially in computer science, to devalue people-centered work. The subject of computer science is commonly conceived to be computation, despite the fact that computation is to be applied in societies. People-centered work is supposedly "soft" work, not intellectually challenging, and therefore undeserving of attention. This perception is fast changing, yet at the time and place in which I write, there remains a lack of inclusion, social connection, and guidance, and such deficiency is distributed unequally according to characteristics like race, class, and gender.

Graduate students are not the only ones responsible for people-centered work, but our impact should not be discounted. Admittedly, it was difficult for me to realize this power, especially since I have spent a large portion of my academic life in subordinate positions, getting told what to do and when to do it. I have felt insufficient, particularly in observing the flood of research work that permeates conferences and journals. However, even in the early stages of a career, graduate students in AI have already accrued a measure of social currency. The public considers us experts in our fields; newer students ask us for insights into questions we have already faced; someone in need of social connection is heartened when we reach out to them. Publication norms in AI research may bias us towards searching for complicated solutions to our problems, but we should not neglect the ways in which our simple acts can change lives.

4.3 Funding Sources

Unless I find myself in the rare position of being able to self-fund my research, I will have to obtain a position in either academia or industry. The recent proliferation of governmental and industrial funding, subject to the increased demand for such positions, may make it easier for me to obtain a position where I can research what I desire. Nevertheless, acceptance of funding evokes moral complexities. To what extent am I beholden to the funding source? Does my acceptance legitimize other, unsavoury initiatives of the funding body? Was this funding obtained through immoral means? Am I infringing upon a moral duty by accepting this funding, or would it simply be a moral aspiration not to accept it?

From a consequentialist point of view, I could argue that regardless of the above questions, acceptance of funding is good if it allows me to pursue research with positive ethical impact. This argument, however, elides how participation in such a transaction, and membership in the funding organization, can mould my moral convictions. In a research environment without an emphasis on the evaluation of ethical impact, say, I would find it difficult to pursue investigations into algorithmic injustice than in an environment where everybody was interested in this evaluation. I owe my interest in ethical inquiry to intellectual environments, such as informal research communities and clubs, that welcome such inquiry. Without these environments, my moral aspirations would have reached much less far, my conception of moral duties been much less wide. Even if there are no formal barriers to ethical inquiry, fear of offending superiors may cause me to censor myself, whether consciously or unconsciously. Moral fortitude is not inexhaustible.

At the same time, nobody has infinite flexibility in deciding how to support their research financially. Non-research obligations, like family, can oblige the acceptance of funding with more strings attached. This situation is not ideal. Informal support networks can help to mitigate the potential erosion of moral fortitude, but it may not be easy to continue with a line of ethical inquiry that meets opposition. This discussion is not to say that ethical AI work is impossible. On the contrary, although some funding bodies engage in problematic practices, and individuals receiving funding from such bodies continue to output ethically insightful work, and even protest against the practices of these bodies.

I am not just–if I am at all–a representative of my institution, but also an individual with my own moral faculties, the environmental moulding of these faculties notwithstanding.

5 Conclusion

It is unlikely that we will have resolved the challenges of the 21st century by the turn of the 22nd. As we march towards the coming decades, we might even expect this final disappointment, chastising ourselves for not having done enough. But our end is neither necessarily final, nor a disappointment. Moral engagement is more than teleological: we strive to be better not only for the final victory of our aims, but also so that our successors may find their path clearer, their burdens lighter, and and their minds wiser. In the dying embers of the final slave revolts of the Roman Republic, the vanquished could not foresee a global consensus against bondage. In the last days of the conquest of Tenochtitlan, the inhabitants could not predict continued resistance against colonization. In the mass carnage of the French Revolution, the new citizens could not know the stability that liberal democracy could afford through secured rights. Our limitations in prophecy should not constrain the extent of our moral vision. It may not be us who will succeed, but that is no worse than the prevailing standard.

Acknowledgments and Disclosure of Funding

I completed the entirety of this work in Montréal, part of the traditional territory of the Kanien'kehà:ka, and a historical gathering place for the nations of the Haudenosaunee Confederacy. The importance of gathering places for research, more visibly important now while we cannot gather in the midst of the COVID-19 pandemic, should remind us both of the centrality of land for Indigenous nations across Canada–especially in light of the swathes of unceded territory–and of the ways in which the Government of Canada has failed to uphold its treaty and human rights obligations.

I acknowledge financial support indirectly from the Canadian Institute for Advanced Research (CIFAR) through my advisor, Nicolas Le Roux, who is supported by a CIFAR AI Chair. I thank Tara Jacklin for a helpful discussion in structuring this work. Much appreciation goes to Jerry Chen and Steven Winter for comments on an earlier draft. I would also like to acknowledge insightful discussions with Roshan Shariff, Nishant Subramani, Csaba Szepesvári, Kevin Wang, and Wesley Chung. Finally, I greatly appreciate the constructive criticism of the anonymous reviewers.

References

R. Abebe, S. Barocas, J. Kleinberg, K. Levy, M. Raghavan, and D. G. Robinson. Roles for Computing in Social Change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 252–260, New York, NY, USA, Jan. 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7.

C. Barabas, M. Virza, K. Dinakar, J. Ito, and J. Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*, pages 62–76, 2018.

A. Ecoffet and J. Lehman. Reinforcement Learning Under Moral Uncertainty. *arXiv:2006.04734 [cs]*, July 2020. URL `http://arxiv.org/abs/2006.04734`.

L. L. Fuller. *The Morality of Law*. Yale University Press, New Haven, 2nd edition, 1969.

C. Miserandino. The Spoon Theory, Apr 2013. URL `https://butyoudontlooksick.com/articles/written-by-christine/the-spoon-theory/`.

A. D. Trister, D. S. M. Buist, and C. I. Lee. Will Machine Learning Tip the Balance in Breast Cancer Screening? *JAMA Oncology*, 3(11):1463–1464, 11 2017. ISSN 2374-2437.

L. Winner. Do artifacts have politics? *Daedalus*, pages 121–136, 1980.