Sarah Vahlkamp
INST888
F2020

Mapping My Theoretical Framework

Introduction

In an effort to appropriately scope it to fit the assignment, this paper will start with a brief overview of the AI field in general, before pivoting toward the subdomain of ethics of AI. While ethics, particularly as related to the field of AI, is itself made up of many facets, for purposes of this paper, it will be discussed holistically, acknowledging that this will overlook some important nuances of the various dimensions. In the relatively near future, I would like to conduct a small research project in the subfield of ethics in AI. Together, these fields create an exciting combination for theory. In this paper, I'll look primarily at ethics in algorithms, and at the concept of consciousness.

Artificial Intelligence

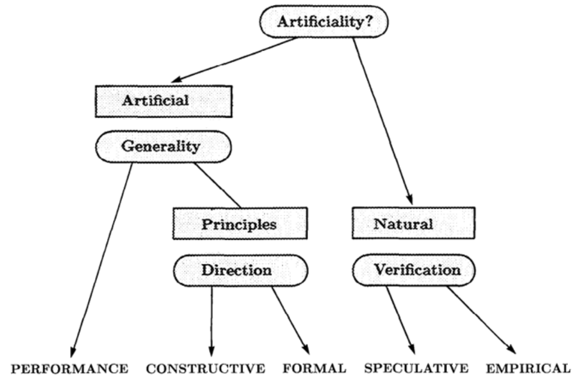
An interdisciplinary field

Artificial intelligence, in general, is a field that spans computer science, philosophy, logic, math, psychology, cognitive sciences, and more. (Doyle-Burke, 2020) Influential scholars within the field include Alan Turing, Ray Kurzweil (discussed below), Fei-Fei Li, Stuart Russell, Nick Bostrom, Peter Norvig, Marvin Minsky, Douglas Hofstadter, David Chalmers, Daniel Dennett, John Searle, Paul Allen, Allen Newell, Rodney Brooks, and John McCarthy. While this isn't an entirely demographically homogenous group, the list does reinforce some of the critiques directed to the field, and perhaps points to a need to introduce and support a variety of viewpoints into the discipline in an effort to prevent groupthink and appropriately represent voices across the spectrum of power dynamics.

Epistemology and approach to philosophy in general

There has been a noted anti-philosophical stance in parts of the technology world, though. For example, Peter Sweeney says “Yet, formalisms are embraced within analytic philosophy so this requirement in itself doesn’t explain the leap. Russell and Norvig stress their short history is necessarily incomplete, but this isn’t a symptom of brevity. Rather theirs is an accurate portrait of a historical split, as if the foundational work in AI marks a break with modern philosophy. While that’s a broad brush, it does capture the popular sentiment. The dismissal of philosophy is widespread, particularly among technical people.” (Sweeney, 2019) Relatedly, despite dabbling in his own philosophical theorizing, Stephen Hawking was known to say more than once that “philosophy is dead.” (Norris, 2011) This seems odd, given how intrusive technology is in our lives and our worldview, and how much it claims it changes the paradigms of society. In return, Dennett suggests a more collaborative approach, “It is true that there is a crowd of often overconfident scientists impatiently addressing the big questions with scant appreciation of the subtleties unearthed by philosophers and others in the humanities, but the way to deal constructively with this awkward influx is to join forces and educate them, not declare them out of bounds.”

I have pages of notes on the philosophical schools of thought embraced by various scholars closely associated with the field of AI. With this as perhaps the subject of future research, for now I will mention only a few key overarching ideas. Computer science tends toward a positivist or empirical interpretation of the world, while AI as a discipline, positions itself as a bit of an outsider in this field. Its future orientation and firm preoccupation in the minds of its scholars and the culture at large demand a more pliant method of understanding. There are still diverging opinions within AI, though, and have been for quite a long time. In 1985, Hall and Kibler produced a paper that summarized the theoretical perspectives of various researchers within the artificial intelligence field. As depicted below, this figure shows the diverging tracks of artificial and natural AI researchers.



PERFORMANCE	CONSTRUCTIVE	FORMAL	SPECULATIVE	EMPIRICAL
MACSYMA	DENDRAL	A*	SAM	GPS
CHESS 4.5	MYCIN	Resolution	PAM	EPAM
R1	ARCH	Fuzzy logic	ARIES	ACT*
	AM/EURISKO	TMS		
	Feigenbaum	Nilsson	Schank	Newell
	Buchanan	McCarthy	Wilensky	Simon
	Winston	Marr	Carbonell	Feigenbaum
	Lenat			Anderson

A typology formed from differing intentions and methods of AI researchers. Exemplary research projects are listed for each perspective along with well–known participants.

Figure 1.

This figure from Hall and Kibler's paper shows diverging epistemology and methodology in AI research

AI Ethics

Algorithms and the people they effect

Much of the current focus on ethics in AI is on the use of algorithms for various potentially life-altering decisions. Through their research, ethicists have found disproportionate effects across race and gender in hiring, facial recognition technology, mortgage decisions, and criminal policy. In order to appropriately address these issues from the lens of the different groups, researchers have broadly undertaken their work from the lens of feminism and critical race theory. While I will go into more specifics when discussing the scholars later in the paper, these points of view have sometimes put them at odds with their colleagues in the more traditionally-focused disciplines in computer science.

Social-robotics and ethics

A paper on ethics and AI would be incomplete without mentioning our future robot overlords. Despite its many forms and benefits, science fiction writers can't seem to help themselves from addressing AI through the lens of the impending robot uprising. They offer keen insights into the inhumanity of man, and present cautionary tales, urging us to establish strong but sensible preemptive laws and norms to reign in the hubris of man and the maniacal lust for power of robots. Famously, in his book *I, Robot*, Isaac Asimov provided a framework to have control over our metal cousins, known as Asimov's Three Laws of Robotics (seen below.)

> *Asimov's 3 Laws of Robotics*
>
> *1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.*
>
> *2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.*
>
> *3. A robot must protect its own existence as long as such protection does not conflict with the first or second laws.*

These takes, and the larger sci-fi obsession with utopia and dystopia, inspire discussion of ethics from a multitude of perspectives. What do we consider to be human? Who holds fault when AI does something illegal? How far is too far to go in the name of safety and security? These questions reflect a far wider analysis of the human condition and roles of ethics and moral philosophy therein. Dr. Liao discusses this, coming to the conclusion that: "On the assumption that moral laws are objective, one might think that they are also universal in the sense that they apply to all moral agents, whether genetically engineered or not."

Now we *finally* talk about me or: Why this framework supports research goals

Although I can philosophize to some remarkably unrealistic ends, when I re-center my focus, I feel like I generally follow a constructivist philosophy most closely. This is supported by the influence Kurt Lewin has had on my field. It has also been qualified fairly strenuously, in a reflection of my battle with myself to determine a worldview. Unfortunately, I can talk myself in and out of many ideas (it's why I prefer essay exams to multiple choice). Interestingly, my debates center around very foundational characteristics. I can and tend to argue from a relativist perspective. I see how different a person's experience can be relative to established norms. However, when I think about empirical knowledge and conducting research, I start to lean more strongly toward a positivist perspective, or at least methodology. Just to make sure I hit all three, it seems some combination of positivist and interpretivist feels most comfortable. Despite these seemingly incompatible or contradictory viewpoints, I do see value in mixing qualitative and quantitative methodology to try to see a full picture. Yes, we are influenced by our experiences, and this plays out most clearly in interpreting and presenting findings, but there are some ideas that, while I might not be able to go so far as to call them true, are repeatable enough as to be generalizable.

Rationalism offers an enticingly ordered perspective; a neat and clean viewpoint that can be understood and comforting in a world of ambiguity. It falls short, though, in its myopic understanding of context. Surely there are universalisms, but we can lean on empirically proven phenomena, particularly in relation to the mathematics and sciences to identify these.

For this next part, I need to be clear that I am referring to ill-informed and deliberately obtuse viewpoints, rather than diverse experience and knowledge. I proffer general anti-vaxx sentiment as the archetype for this short digression. I have at times discussed my discomfort with the potentially incongruent ideas of expertise and inviting broad interpretation. Despite my firm commitment to a focus on diversity, and true belief that often marginalized voices carry a closer interpretation to the truth, I do understand and value the input of established specialists. I worry that in a quest to be inclusive of all valid voices, we are watering down the weight of diligently researched opinions. I strive to be inclusive, while still acknowledging the commensurate value of conscientious study, understanding that inclusive and expertise don't imply distinction. While I might find myself seduced by the openness of Nietzsche's paraphrased assertion that "There are no facts, only interpretations," (Ludovici, 2016) this indulgence is overwritten by the stronger pull of philosophies that allow some room for debate with acknowledgement of the possibility of shared experience, say constructivist's falsification leanings mixed with a dash of abductive reasoning or occasional aggregate Bayesianism for good measure.

Further, I have some interest in learning more about Theory of Mind, and earlier this semester, reached out to Dr. Georges Rey, here at UMD. I don't believe computers can have thought, but do enjoy the thought-games that come with chasing down these ideas. There is clearly a reason I wish there were a test that could tell me which theoretical viewpoints fit my beliefs. I am also lucky there are catch-all epistemologies like pragmatism.

Consciousness and Perception

Underpinning the relativist viewpoint is the concept of qualia. I was so excited to see this source of rumination had a name. In basic terms, qualia is the word that describes the sense that we are only privy to the information we can see in our mind, in a somewhat Cartesian way. (Tye, 2018) Basically, this word sums up the idea that I will never know what your concept of the color red actually looks like. More eloquently, "Let us consider Johannes Kepler: imagine him on a hill watching the dawn. With him is Tycho Brahe. Kepler regarded the sun as fixed: it was the earth that moved. But Tycho followed Ptolemy and Aristotle in this much at least: the earth was fixed and all other celestial bodies moved around it. Do Kepler and Tycho see the same thing in the east at dawn?" (Hanson, 1958)

Depending on your interpretation of this term, an adherence to this line of thinking can imply subjectivity in a way that leads to constructivism. That I've taken it to mean that the term describes the nature of this concept in terms of application to both internal and external phenomena is just further evidence that despite my ideological conflict with myself, constructivism seems to address the nuances of the field more directly. For a path down the rabbit hole to explore similar ideas, a trip examining direct/naive vs. indirect realism offers some fun, if a bit far-fetched, thought experiments, but that's a topic for another paper.

Approaching AI and ethics from a constructivist mixed-methods approach serves my research goals in several ways. It allows me to implement my study with methodological nods to the prevailing theoretical lenses across this multi-disciplinary concept. Abductive reasoning is already employed in the AI field. Problems related to AI ethics and other research interests that may have many unknown causes benefit from adductive inference. It allows for many possibilities, in order to go down many roads. It is also useful for looking at phenomena that is hidden, whether malignantly or benignly. Moreover, abductive reasoning has become more aligned with creativity, establishing influence in an area of intense interest related to AI.

Scholars

These people might be important, too

There are a large number of scholars who are influential in this field, and it is growing daily. After reaching out earlier this semester to several UMD faculty who will be teaching courses related to AI and/or AI ethics, I was able to create a more robust list of important scholars to discuss. It's important to represent the breadth of the field, while still retaining the brevity of the assignment. In this paper, I'll only directly discuss three scholars, but each brings either a different theoretical point of view, or a different application of that view. The interdisciplinary nature of the field means that there is a breadth of research backgrounds. Given the focus of the topic, several of the scholars involved with ethics in AI come from a feminist critique and critical race theory lens. There is additionally influence of normative ethics, deontology, phenomenology, and relational ethics. I'll discuss each in more detail.

Matthew S. Liao

I found quite a lot of information about Dr. Liao in his Ask Me Anything (AMA) for r/philosophy. (Liao, 2017) His extensive research in normative theory and neuroethics extends to touch on AI ethics. (Liao, NYU bio) His book, E*thics of Artificial Intelligence*, is one of two books UMD philosophy faculty Dr. Dwyer is using in her new undergraduate ethics and AI class next semester as indicated in her email on 12/1. Liao has keen interest in normative philosophy, and further suggests the benefit of a foundation in philosophy of mind and social psychology. (Liao, 2017)

In this "interview", he describes his own philosophy. "This is going to be too brief, but I tend to be more sympathetic to realist, cognitivist, externalist positions. So with respect to you questions, I would a) be skeptical of the Humean theory of motivation; b) be in favor of moral cognitivism; c) question motivational internalism; and generally think that one's beliefs ought to conform to the world." (*ibid*) The discussion, which crossed many disjointed, but interesting topics, also hit on the idea of consciousness in a discussion on the threat of AI and internal will. In it, he suggests that the question might not be about will, but about the use of the technology, arguing that "For AI- is an internal will necessary for it to pose a threat to us? Is will seen as an

It probably doesn't need an internal will for it to pose a threat to us. Compare: nuclear weapons threaten our existence even though they don't have a will. But certainly, if AIs had wills, that could pose even greater threat.” (*ibid)*

In his AMA, he offered deep insight into his work and views. Rather than just re-print the rest of that, I’ll include a key point he included in his background: “In this book, I shall argue that consequences should matter in our moral decision making, but they are not the only moral inputs that matter. Considerations such as an agent’s intention, an agent’s rights, the fairness of an act, and so on, are also relevant for determining the permissibility of an act. I am also investigating the under-explored and under-theorized phenomenon of moral indeterminacy. For instance, many people believe that it is impermissible to kill one innocent person to save five other innocent people from being killed. At the same time, many people have the intuition that it may be permissible to kill one innocent person to save, e.g., one million people. My interest lies in whether there is a precise threshold when the act of killing an innocent person changes from impermissibility to permissibility or whether the boundary is fuzzy.” (ibid) This question is a fundamental question in ethics and morality of AI. AI, in any form, is programmed to make decisions. Prior to any large-scale implementation of AI, its designers and engineers are going to have to make moral and ethical decisions that have severe real-world consequences. Dr. Liao is addressing these questions head-on, citing the internet as technology that should have been thought through more deeply before introduction to the masses. Additionally, his discussions on who owns memories have surprising implications for future of work research.

Timnit Gebru

Dr. Gebru, formerly of Google (with a very recent and very controversial exit (Hao, 2020)), Microsoft Research, and Stanford, is hailed as a current leading voice in ethics and AI. She studied at Stanford under Fei-Fei Li, and her research into facial recognition technology and race shined a spotlight on a major area of concern in AI ethics.(Gebru, 2018, Buolamwini, 2018) She is cited as using critical race theory and feminist critique views in her research, and champions the voice of the marginalized and those with less power. (Doyle-Burke, 2020) A lecture she co-created in 2014 gives many insights into her philosophy, and supports this assessment of her views. Among many of the quotes in the slides, is one that seems to sum up her view on the researcher's role: "Our social positions in the world and set of experiences shapes and bounds our view of the world; this in turn affects the research questions we pursue and how we pursue them." (Gebru, 2020)

As shown by the voluminous reaction to her firing, both positive and negative, Gebru is a highly influential voice in AI ethics, which will continue to be used to take a stand that tech companies need to think about consequences of their innovation, rather than just implementation. Think of this as her answering the eternal 'you can, but should you?' question.

Ray Kurzweil

Ray Kurzweil is one of the world's leading inventors, thinkers, and futurists, with a thirty-year track record of accurate predictions. Called "the restless genius" by *The Wall Street Journal* and "the ultimate thinking machine" by *Forbes* magazine, Kurzweil was selected as one of the top entrepreneurs by *Inc.* magazine, which described him as the "rightful heir to Thomas Edison." PBS selected him as one of the "sixteen revolutionaries who made America." (About Ray)

Kurzweil's books have been read and cited by top AI researchers time and again. His arguments, while coming from a 'futurist'' perspective, include detailed looks at implementation and integration into society. He argues that "An overall strategy should include a streamlined regulatory process, a global program of monitoring for unknown or evolving biological pathogens, temporary moratoriums, raising public awareness, international cooperation, software reconnaissance, and fostering values of liberty, tolerance, and respect for knowledge and diversity." (Kurzweil, 2013) His existential views also reflect his futurist nature. "Kurzweil insists that once [the] Singularity arrives, the answers to the questions 'What does it mean to be human?' and 'What are the limits of human knowledge?' will be answered in ways as-yet-unimaginable. (Williams, 2011)

Kurzweil has outlined a future in which lines between humans and social robots are blurred, and in which "ultimately the Universe itself will become conscious and that we will have achieved nonbiological immortality." (Williams, 2011) In this same article, he outlines his view of AI oversight, as a balance between restrictive regulation and privatization.

AI ethics are further influenced by the voices of Cathy O'Neill, Safiya Noble, Jenn Wortman Vaughn, Eric Rice, Sarah Myers West, Miriam Sweeney, danah boyd, Emily M. Bender, Beth Singler, Deb Raji, John C Havens, Michael Madaio, Ruha Benjamin, Lilly Irani, Wallach and Allen, Bonnefan, and Mark Coeckelbergh.

Citations

About Ray Kurzweil. http://www.kurzweiltech.com/aboutray.html.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91).

Dennett, D. (2013, September 10). *Dennett on Wieseltier V. Pinker in The New Republic*. Edge.org. https://www.edge.org/conversation/dennett-on-wieseltier-v-pinker-in-the-new-republic. Archived August 5, 2018, at the Wayback Machine *Edge.org*. Retrieved December 14, 2020

"It is true that there is a crowd of often overconfident scientists impatiently addressing the big questions with scant appreciation of the subtleties unearthed by philosophers and others in the humanities, but the way to deal constructively with this awkward influx is to join forces and educate them, not declare them out of bounds."

Doyle-Burke, D. (2020, March 24). *Using Feminist Theory to Nuance Bias in Artificial Intelligence*. Medium. https://medium.com/@dylandoyleburke/using-feminist-theory-to-nuance-bias-in-artificial-intelligence-64ca6af33879.

Doyle-Burke, D., & Smith, J. The Radical AI Podcast. https://www.radicalai.org/.

Gebru, T., & Denton, E. (2019). *Fairness, Accountability, Transparency, and Ethics in Computer Vision*. Index of /slides/2020. http://cs231n.stanford.edu/slides/2020/.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Godfrey-Smith, P. (2009). *Theory and reality: An introduction to the philosophy of science*. University of Chicago Press.

Hall of Fame. 100 Brilliant Women in AI Ethics™. (2020, December 8). https://100brilliantwomeninaiethics.com/the-list/hall-of-fame/.

Hao, K. (2020, December 7). *We read the paper that forced Timnit Gebru out of Google. Here's what it says.* MIT Technology Review. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/.

Korb, K. B. (2001). Machine learning as philosophy of science. In *Proceedings of the ECML-PKDD-01 Workshop on Machine Learning as Experimental Philosophy of Science, Freiburg*.

Lawrence, N. D. (2019, December 2). *Perspectives on AI*. Neil Lawrence's Talks. http://inverseprobability.com/talks/notes/perspectives-on-ai.html.

Norris, C. (2011). *Hawking contra Philosophy*. Philosophy Now: a magazine of ideas. https://philosophynow.org/issues/82/Hawking_contra_Philosophy.

Pelillo, M. & Scantamburlo, S. (2016, September 19-23) *Philosophy meets machine learning: From* *epistemology to ethics.* [Conference workshop]. Machine learning and knowledge discovery in databases European conference, ECML PKDD 2016, Riva del Garda, Italy. https://www.dsi.unive.it/~scantamburlo/PhilML/ECML_Part1.pdf

Popper, K. (1976). The myth of the framework. In *Rational changes in science* (pp. 35-62). Springer, Dordrecht.

S Matthew Liao. S Matthew Liao | NYU School of Global Public Health. https://publichealth.nyu.edu/faculty/s-matthew-liao.

Search People Result. Academic Influence. (2020).

https://academicinfluence.com/people?text=&gender=&country=&discipline=computer-science&subdiscipline=artificial-intelligence&year-min=1800&year-max=2020

For a fun look at someone's idea of influence, here is an interesting site that drills down based on disciplines. According to their listing, the most influential woman in AI, Pamela McCorduck, is only the 42nd most influential person in AI.

Another woman doesn't appear until #78 – Ada Lovelace According to the site She is believed by some to be the first to recognize that the machine had applications beyond pure calculation, and to have published the first algorithm intended to be carried out by such a machine. As a result, she is often regarded as the first to recognize the full potential of computers and as one of the first to be a computer programmer. She is listed 26 spots behind Ken Jennings on this list.

Sweeney, P. (2019, May 21). Your AI Superpower. https://www.explainablestartup.com/2017/07/your-ai-superpower.html.

Tye, Michael, "Qualia", *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/qualia/>.

Williams, J. (2011). *The Singularity Is Near by Ray Kurzweil*. Philosophy Now: a magazine of ideas. https://philosophynow.org/issues/86/The_Singularity_Is_Near_by_Ray_Kurzweil.