Ethically Aligned AI: Applying ethics in developing AI systems for healthcare

Katrina Regan-Ingram

University of Alberta

Submitted to the Faculty of Arts

University of Alberta

In partial fulfillment of the requirements for the degree of

Master of Arts in Communications and Technology

August 2020

Acknowledgements

I have been fortunate to have the support of an amazing community on my academic journey. I am tremendously grateful to the following people:

First and foremost, my husband Ken, who has always been my biggest champion. Thank you for supporting me in every way.

The MACT 2018 cohort – you guys are the best! Thanks for the solidarity in learning.

The MACT team – Professors Gordon Gow, Thomas Barker, Rob McMahon, Mark Wolfe, Megan Lefebvre, Jason Daniels and Program Administrator, Eileen Crookes, thank you for all you do to deliver a high calibre learning experience.

My supervisor, Professor Geoffrey Rockwell, thank you for great discussions, valuable advice, opportunities to grow as a researcher and reminding me that it would all work out.

Professor Jonathan Shaeffer, thanks for inspiring my interest in AI and ethics, for supervising my self-directed course and your ongoing support of my work.

MACT alumni Hanne, Colleen and Sherrell, thanks for your friendship and being a sounding board on so many things.

Rose and Ken, thank you so much for being references, resources and friends.

MacEwan University, thanks for supporting me with professional development funding.

Professor Randy Goebel, Professor Russell Greiner and Laura C., I really appreciate your help in connecting me with researchers in the AI community – thank you.

The participants in my study, thank you so much for being part of my work and generously sharing your time.

My two AI ethics reading group members: Howard, Tugba, Emad, Jonathan, Marilene, Julie, Daniel, Fahim and Nicole. It has been such a gift to learn and unpack ideas with you.

Professor Kyle Whitfield, thanks so much for great chats about the research process.

Professor Hong Qu, thanks for providing background information about AI Blindspot.

All of my *Back to School Again* podcast guests, thanks for sharing your remarkable stories. You inspired me more than you will know.

Last, but not least, my parents, whose early work set the stage for this degree. Thank you!

Abstract

Artificial Intelligence (AI) is a technology that is quickly becoming part of our digital infrastructure and woven into aspects of daily life. AI has the potential to impact society in many positive ways. However, there are numerous examples of AI systems that are operating in ways that are harmful, unjust and discriminatory. AI systems are constructs of the choices made in their design. They exist within a socio-cultural context that reflects the data used in their training, the design of their mathematical models and the values of their creators. Ethics, the moral principles that guide our behaviour, can help us decide what values we want to uphold. There are opportunities to introduce a broader ethical perspective into AI development which might serve to mitigate harms and yield better outcomes in designing trustworthy AI systems. Ensuring that we have ethically aligned AI is important for achieving the trust and social license needed to apply AI in high stakes fields such as healthcare. This study explores the question:

> **How can AI researchers working on healthcare related projects use ethics tools to inform their research process?**

An exploratory, qualitative research design was selected to examine this question using a combination of individual interviews with AI researchers (PhD students) working on healthcare projects followed by a focus group. The subsequent findings and discussion highlight how ethics tools can play a role in helping AI researchers address ethical issues in the context of their work by offering a means to broaden ethical thinking in ways that go beyond technical fixes.

Keywords: artificial intelligence (AI), ethics, healthcare, data, privacy, bias, ethics tools

Table of Contents

Introduction

Artificial Intelligence (AI) is a technology that is being woven into the fabric of our daily lives and is quickly becoming part of our digital infrastructure. It can be a useful technology when it is working in ways that align with our values. Yet, there are many examples of AI systems that are operating in ways that are harmful, unjust and discriminatory. From hiring algorithms that screen out female applicants for certain jobs to systems that result in over policing minority communities to algorithms that determine eligibility for a bank loan but can't explain how that decision was made, AI systems are impacting people's lives in substantial ways (Bogan, 2019, Hao, 2019, Levin, 2019). Healthcare is a domain where the use of AI systems can have life and death consequences. Given the high stakes involved and concerns around the privacy of sensitive medical data, it is understandable that people are worried about the impact of AI systems within healthcare. There are inevitable questions about what we can do to protect people and how we can ensure these systems are accurate, fair, equitable and working to benefit everyone. In other words, how can we have more ethically aligned AI? This research paper seeks to address some of the ways that we can apply ethics in developing and deploying AI systems in healthcare.

First, the paper contextualizes the issues by exploring a body of literature which shows how AI, like other technologies, is socially constructed and how "values are instantiated" into these systems (Shilton, 2015, p. 2). AI is designed by people who use data as inputs to train models that they have built. The data itself is a construct of many choices that inform what is collected, by whom and in what ways – a "data assemblage" (Kitchin, 2017, p. 22). Similarly, the mathematical models used are the result of certain design choices and are not neutral (O'Neil, 2017). The development of AI models takes place within a greater sociotechnical and cultural context and is impacted by a wide range of issues including funding and the demographic make up of the AI community. The relationship between society and technology, as social construction of technology (SCOT) theory explains, is negotiated and framed by the actors involved in a cyclical process (Bijker, 2015). Thus, AI is a construct of the many choices that go into shaping its design. If ethics informs the AI design process, then we should be able to design more ethically aligned AI.

Next, the paper turns to exploring this question – how can AI researchers working on healthcare related projects use ethics tools to inform their research process? An exploratory, qualitative research design was selected to examine this question using a combination of individual interviews with PhD students working on healthcare projects followed by a focus group. The study was conducted online in accordance with COVID-19 social distancing requirements. The interviews focused on exploring how AI researchers working in healthcare think about ethics and how they apply ethics in the course of their work. Interview findings were used to inform a selection of three ethical tools that were discussed during the focus group. Data from the interviews and focus group were coded and analyzed using qualitative content analysis. A software program was used to apply a digital text analysis to the focus group data to yield additional insights and view that data from another perspective. These methods were selected to provide a rich data set to explore this question. Even given limitations surrounding a small, non-probability sample and the resulting lack of generalizability, this approach was deemed to be the best fit for this research.

The final chapter presents the results of my research. The findings section is organized in two parts: interview findings and focus group findings and it is followed by the discussion section. I stayed close to the data during analysis, as is the practice with exploratory research, highlighting excerpts from participants which were then grouped into relevant themes (Mayan, 2009). The focus group data is organized around feedback of the three tools and is followed by a visual text analysis. The findings indicate that participants would benefit from a broader perspective of ethics and that ethical tools can be effective in helping to weave ethics into an AI workflow. This was supported by the focus group discussion which led into an exploration of a topical ethical issue, explainability. Finally, the discussion relates the findings back to larger learnings, point towards two major issues: improved ethics education and revisiting how research ethics boards provide governance to researchers in an era of big data. While the limits of my research preclude drawing bigger conclusions, my findings suggest that future work in these areas, especially with respect to improved education, would be helpful next steps in the addressing applied ethics within AI.

Ethics can play a role in mitigating some of the challenges in AI. AI researchers, as the architects of AI systems, can use ethical tools in ways that improve their work. While ethics won't solve all of the problems in AI systems, it is a step in the right direction towards building trustworthy AI. To better understand the issues in AI and how ethics can help, let's turn to the literature to explore how AI systems are constructed.

Literature Review

Our world is shaped by the people who build our infrastructure. We understand this relationship clearly when we think about physical infrastructure. Urban planners, engineers and architects construct our transportation corridors and skylines, guiding the daily interactions of people within a place. Our digital infrastructure shapes our societal interaction in similar ways. Increasingly, the digital realm itself is informed by artificial intelligence. AI can be defined as machines that think and learn in ways that perform on par with, or better than, humans. As research advances and more AI technology is deployed, from autonomous vehicles, to the Internet of Things, to a wide range of AI enabled algorithms and robots, it will have an almost ubiquitous reach. Ample evidence suggests that "values are instantiated in sociotechnical systems" (Shilton, 2015, p. 2). AI systems are not an exception. They exist within a socio-cultural context that reflects the data used in their training, the design of their mathematical models and the values of their creators. Ethics, the moral principles that guide our behaviour, can help us decide what values we want to uphold. If AI systems are constructs of the choices made in their design, and AI researchers are the infrastructure architects, then proactively weaving ethical considerations into their work should result in more ethically aligned AI. Making ethics actionable in practice is a challenge. However, if we introduce and apply tools to examine ethical considerations and potential "blind spots" during the design process, we may be able to create better outcomes. Ensuring that we have ethically aligned AI is important for achieving the trust needed to use AI safely, especially in high stakes fields such as healthcare.

Healthcare is one area where AI can deliver tremendous benefits. AI technologies hold the promise to prevent disease through early intervention, enable personalized treatment plans and provide better access to care (Topol, 2019). It can also reduce costs and support clinicians facing increasingly unsustainable workloads (Topol, 2019). The potential for AI to improve healthcare is vast, but so are the consequences if we do not create AI that aligns with societal values. While technology has advanced, policy and ethical guidelines for the use of AI in healthcare is lagging behind, leaving the medical community "ill informed of the ethical complexities that budding AI technology can introduce" (Rigby, 2019, p. 121).

This paper explores how researchers can examine ethical considerations in the development of AI with a focus on healthcare. First, I will review the broad ethical issues in AI systems and the development of ethical codes. Second, I will take a critical look at the data used to train AI, the models used and the people who build AI systems, all of which intersect within a socio-cultural context. Next, I will look at the domain of healthcare and the areas where AI has proven beneficial as well as where it has fallen short. Finally, I will conclude with a proposal for further research to explore this question - how can AI researchers working in the healthcare domain apply ethics to inform their workflow?

Literature Review Methodology

Ethics in AI can be interpreted many ways, from the duties we might owe to a super-intelligent AI, to how we should engage with sentient robots, the morality of the singularity or how we should address the great displacement of human jobs by AI to name a few. My approach centres around how AI aligns with human centric social justice issues. This includes ensuring AI does not amplify bias or inequality, protecting against discrimination and making sure decisions are explainable. I've framed these within a socio-cultural context that informs the construction of AI, including how data is collected and stored, issues of privacy and consent, an examination of the people involved in building AI, the mathematical models used, as well as the deployment or use of the systems. I did not include issues of super intelligence, the morality of AI, AI rights or existential risk, which are part of the larger realm of AI and ethics, but beyond the scope of my approach.

Using these parameters as a starting point, I searched a number of databases, including Springer, Sage, SSRN, Phil Papers and ArXiv. I reviewed conference proceedings for the latest material, including Association for the Advancement of Artificial Intelligence (AAAI), the ACM

Conference on Fairness, Accountability and Transparency (ACMFAT) and IEEE's Xplore Digital Library. The MIT Technology Review newsletter provided breaking news and introduced me to key players in academia, industry and tech journalism. I also included reputable non-academic periodicals and newspapers.

I started with broad keyword searches such as AI AND ethic*, "build* AND ethic* AND AI", "artificial intelligence" AND ethic*, bias AND ethic*, fairness AND ethic*, transparency AND ethic*. As I focused my topic, I added other terms related to healthcare, privacy, big data and information ethics. The search process has been iterative rather than linear. I prioritized papers published in the past three years and reviewed 152 sources.

In addition to time spent online, I consulted people. MACT librarian, Patti Sherbaniuk, provided time saving suggestions and continues to send me articles of interest. I also joined a reading group and sought out other academics who connected me to new sources of information. Building and connecting to a community of practice has been incredibly valuable. It has made for a much richer experience in conducting research.

Theoretical Context

Social Construction of Technology (SCOT) is "a theory about the relationship between society and technology" used to "study technical change in society" (Bijker, 2015, p. 135). SCOT developed as a criticism to technological determinism, which claimed the technological impacts on society were both autonomous and determined social outcomes (Bijker, 2015). SCOT looks at technology through a relevant social group to understand how meaning is constructed and assigned, eventually converging to a stable state that is absorbed through a technological frame (Bijker, 2015). The process is cyclical, moving from "artifact, technological frame, relevant social group" to "new artifact, new technological frame, new relevant social group etc."

(Bijker, 2015, p. 137). This theory is pertinent to my topic and explains how society is navigating its relationship with AI, which in turn will shape AI technology. We are still in the early stages of this process, where we find differing opinions about the nature of AI as we move towards negotiating a "stable state".

For example, the first wave of ethics in AI has been focused on issues of bias, fairness and transparency. Its been concerned with fixing the technological aspects of AI and the conversation has largely taken place within the AI community. While there is still much work to be done on this front, a second, more radical critique is emerging which questions whether or not AI should be used at all in certain high-risk domains (Zimmerman, Di Rosa, Kim, 2020). This group sees the ethical issues as a "collective problem *for all of us* rather than a technical problem *just for them"* and calls for a much broader, more inclusive conversation involving all aspects of society, from corporations to government to individual citizens (Zimmerman et al, 2020).

This more radical "second wave" critique aligns with some of the potential limitations of SCOT. Political theorist Langdon Winner outlines a number of these issues saying that SCOT is too narrowly focused on immediate interests, fails to fully examine technology's place within the greater context of human experience, doesn't take a stand on moral or political principles, disregards the consequences of technology and maintains a certain level of elitism by silencing or simply not acknowledging particular voices as being relevant (Winner, 1993). I don't disagree with any of this. However, I see SCOT as a useful starting point to examine how artificial intelligence is being shaped, who is shaping it and how meaning is being negotiated in a socio-technical context. We can then use that knowledge and understanding as a basis to further "illuminate processes of technological design in ways that might serve the ends of freedom and justice" (Winner, 1993, p. 376).

Part One: Ethics in AI

Why Ethics?

We need to be cognizant of ethical issues as we build and deploy AI. Complex and high stakes decisions are being delegated to AI "such as granting parole, diagnosing patients and managing financial transactions" (Cath, 2018, p 2). Decisions made by AI technologies have adversely impacted people's lives without a clear indication or explanation of why and how these decisions are being made. AI holds many opportunities, but it also presents challenges for democratic societies to ensure AI is built and deployed in ways that are equitable, inclusive and do not amplify historical prejudices (AI for Society, 2018). If we accept the premise that AI will impact society then we need to decide how we want to shape it to ensure it benefits humans and aligns with societal goals. Ethics is a starting point to determine what values we want to uphold in the development, design and deployment of artificial intelligence.

Ethical questions are normative, seeking answers about how the world should operate. There are differing approaches to ethics itself, but in general there are "three major critical orientations: deontological ethics, utilitarianism (sometimes called consequentialism), and virtue ethics" (Goldsmith & Burton, 2017, p. 25). To summarize these positions:

	Key Features	**Key Figures**
Deontology	• Rules/law based • Moral duty • Laws can be universal • What are the right rules and how best to apply them?	Immanuel Kant Biblical (Ten Commandments)
Utilitarianism	• Greatest good for most • Concern for consequences • Favoured by computer science, fits well with a math model • What is the greatest possible good for the greatest number?	Jeremy Bentham John Stuart Mill
Virtue Ethics	• Moral Character • Good personal habits • Practical wisdom (phronesis) • Individual/localized • Who should I be?	Ancient Greeks Aristotle

Figure A (Goldsmith & Burton, 2017)

Ethical issues can be explored from various angles and it's useful to consider different approaches when confronted with difficult choices (Goldsmith &Burton, 2017). There are many difficult choices to consider in the development and deployment of AI.

Ethical Codes: Possibilities and Limitations

Ethical codes are historically rooted in the field of bioethics and the regulation of medical experiments (Boddington, 2017). Often "codes of ethics (and laws and other regulations) have developed in response to catastrophes or scandals" (Boddington, 2017, p. 49). However, this worst-case scenario approach is far from ideal. It is better to develop codes that anticipate possible problems, and in the case of AI ethics, that's the approach being attempted (Boddington, 2017). While AI ethical codes will build on other existing codes as it pertains to specific industries (i.e. healthcare), there are some ethical issues being raised that are distinctive to AI. These mostly centre around the "extension, enhancement and replacement of human agency and reasoning" (Boddington, 2017, p. 29).

In the past two years, many ethical codes for AI have been developed. These codes mostly reflect a western perspective and have been primarily developed by industry, professional associations, academics, think tanks and government bodies. On the surface, there seems to be a lot of agreement in principle. One assessment noted that ethical codes for AI only differ in one important area, explainability, and are similar to traditional bioethics in terms of general principles of beneficence, non-maleficence, autonomy and justice (Floridi Cowls, Beltrametti, Chatila, Chazerand, Dignum & Vaynea, 2018). Another review shows enormous overlaps (90% agreement) for certain issues such as accountability, privacy and fairness, which can lend themselves to technical fixes, while missing more nuanced areas like "contexts of care, nurture, help, welfare, social responsibility or ecological networks" (Hagendorff, 2019, p. 4). Given that most codes were developed by technologists, it's not surprising to see more technically oriented issues receive widespread agreement. The further development of ethical codes might benefit from a more diverse and inclusive perspective, such as public consultations, which could address some of the gaps. It also seems reasonable that putting codes into action, especially to address concerns with high levels of agreement, is a logical next step.

In order to make ethics work in practice, some researchers believe we might need to deal with conflicts or tensions that "prioritize one set of values over another" (Whitlestone, Nyrup, Alexandrova & Cave, 2019, p. 7). There might be trade-offs between values such as accuracy vs fairness, quality and convenience of services vs privacy (Whittlestone et al., 2019). It is not clear whether these trade-offs actually need to be made in general terms and circumstances may vary on a case by case basis. However, if trade-offs are needed, it is important to understand who gets to make those decisions, how they are being made and the resulting implications.

As we consider ethics from the perspective of protecting society in the design, development and deployment of artificial intelligence, it's also important to ask if ethics alone form an adequate response to these concerns. The construction of ethical codes is a form of soft regulation or self policing. It is one way to get "ahead of government in an effort to shape the regulatory framework that could eventually govern the use of AI" (Serebrin, 2019). Thus, it can be an appealing starting point for industry and the AI research community. This is not to dismiss ethics as a consideration, but rather, to point out that the ethics discussion itself may lean towards these interests, and therefore, may have limitations in fully addressing societal concerns.

The limitations around successfully translating ethical principles into action seems to be particularly challenging for the discipline of software development which includes artificial intelligence. Unlike professions such as medicine which have a long history of professional conduct, common aims and duties and proven legal and normative accountability mechanisms, the software industry has largely failed to achieve a similar ethical governance structures despite attempts made by bodies such as the Association of Computing Machinery (ACM) and Institute of Electrical and Electronics Engineers (IEEE) (Mittelstadt, 2019). While these criticisms are valid, they stem, in part, from a social construction of technology (SCOT). The profession itself, and society in general, is still navigating its relationship with these technologies. Medicine went through a centuries long process of negotiating its ethical norms and values before bioethics principles became universally "fixed" and translated into actionable ethical conduct by individuals and professional organizations (MacDougall, Langley, n.d.). Those universally agreed to ethical ideals helped inform regulations and laws which could then be enforced. Currently, the software profession, including AI development, is facing an "absence of proven methods to translate principles into practice" but this doesn't mean we are on the wrong path in attempting to translate ethical principles into practice, rather we are not far enough down the path (Mittelstadt, 2019, p. 503). Ethical inquiry emerges in practice from a mix of codes, professional culture, norms and decisions made by individuals (Ananny, 2016). We need to keep moving forward, anticipating how these "intersecting dynamics" will "matter for the future" while finding ways to hold them "accountable" to an ethical framework (Ananny, 2016, p. 96).

The literature suggests ethical codes can be a good starting point to align AI with important societal values. However, these codes "can only help close the AI accountability gap if they are truly built into the processes of AI development" and are enforced in ways that are accountable to public interests (Whittaker et al., 2018, p. 9). My research seeks to address part of this challenge through building ethics into the workflow of AI development in healthcare. Medicine was the catalyst for developing ethical codes. Just as doctors abide by the Hippocratic oath, AI researchers can benefit from considering ethical issues related to their work.

Part Two: Data, Models and People: Examining AI through an Ethics Lens

AI systems are comprised of data, algorithmic models and computing power. Each of these elements are constructed by people and exist within a greater socio-cultural sphere. By examining these elements in more detail, we can start to see how ethical considerations intersect with the construction of AI systems.

The Devil is in the Data

Big datasets are necessary inputs for training AI systems. They are also one of the key ingredients that can lead to bias, discrimination and unequal outcomes.

To understand data, we need to consider it within a bigger context. Data are abstracted elements or representations that seek to categorize and measure phenomena (Kitchin, 2017). While data is often seen as conveying the facts, data isn't purely objective. "Counting is political" and how data is collected, stored and constructed, for whom and for what purposes reflects the values of those compiling the data (Lohr, 2015, p. 91). Data are part of a "complex socio-technical system" that forms a "data assemblage" which includes "ideas, techniques, technologies, systems, people and contexts" that "evolve and mutate over time" (Kitchin, 2017, p. 22). All of this exists independent of AI; however, AI can amplify and codify these data agendas in ways that have profound societal implications.

Datasets are often incomplete, stuck in silos and contain historical biases. Examples of AI systems that discriminate against a particular group often have their roots in the use of data that contain a bias. For example, Amazon's hiring algorithm that discriminates against women for certain roles was trained on datasets that privileged men (Dastin, 2018). Convenience of access to data creates a more subtle bias. Researchers gravitate towards datasets that they can more easily access, especially if the data is in the public domain and requires little if any clearance or permission to use. For example, Twitter data tends to get used frequently (Ahmed, Bath & Demartini, 2017). This also raises questions about information shared for one purpose but used in another context (Ahmed et al, 2017). In addition, machine readable data is preferred as it is in a workable format and easily aggregated as opposed to non-digital data or data that is stuck in siloed systems (Kitchin, 2017). Finally, incomplete data sets can result in unequal outcomes in either being used, resulting in error, or not being applied, resulting in lack of inclusivity (Kitchin, 2017). It is important to look at data and where it comes from in a holistic way to understand different types of data bias.

The relative ease of gathering data, the low cost of storage, and AI's ability to process vast datasets have enabled a new business model, surveillance capitalism, that predicts behaviour by capturing and claiming human experience (Zuboff, 2019). "There was a time when you searched Google, but now Google searches you" (Zuboff, 2019, p. 262). Products and services are designed in ways that normalize and reward data collection and sharing (Zuboff, 2019). We are living through the results of this experiment and its impacts on how we work, shop, vote, stay informed, relate to each other and to ourselves – essentially, how we live our lives. The deployment of more AI systems, which are data hungry, will further intensify and drive the demand for extractive practices while continuing to concentrate power in the hands of a few large players (Zuboff, 2019). What choices do we want to make as we enable an AI empowered future? Some might say it's too late, that we live in a post-privacy era and that we've freely consented to these conditions of surveillance in exchange for convenience.

The concepts of privacy and consent underpin how our society functions. Privacy is a value that has evolved over time and has both legal and moral grounds. Generally speaking, people have an expectation of privacy that allows them to "restrict access" to certain information and to have that decision respected (Moor, 1990). Even if there is "nothing to hide", privacy is an important value for both individuals and societal cohesion (Solove, 2013). Consent is the mechanism by which we grant access to information and the gold standard is informed consent, a voluntary agreement between parties with equal bargaining power (Richards & Hartzog, 2019). Our laws and norms are based on this foundational understanding of privacy and consent, but the digital sphere is challenging this understanding. Online terms of use policies "often serve more as liability disclaimers" than real protections of informed consent (Kitchin, 2017, p. 8). Thus, while we may legally "consent", we actually have no real choice (Richards & Hartzog, 2019). Online consent raises particularly challenging ethical concerns as big data sets gathered from platforms like Facebook are increasingly informing research in sensitive areas such as mental health (Gomes, Pawson, Muriello et al., 2018).

Our data, including what we share publicly or within a selected online group, as well as the clicks, swipes, geo-location information and other "meta data" becomes the cost to participate in modern society. There is a blurring of public and private space that makes navigating the consent relationship more challenging (Ahmed et al., 2017). The increased use of biometric data, fueled by facial recognition, has led to calls to reconsider current privacy standards and stronger legal oversight (Kugler, 2018). Facial recognition technology used in law enforcement has recently become a hotly contested issue, leading some companies to bow out of this sector or limit its use for this purpose (Bajarin, 2020), while other sectors such as aviation are advancing new uses of facial recognition technology as a screening mechanism (Burt, 2020). Even if data is anonymized, many studies show how datasets can be aggregated to reidentify people (Heffetz &Ligett, 2013; Zang, Dummit, Lisker, & Sweeney, 2015; Henle, Matthews & Harel, 2019). Some researchers are working on technical solutions to solve this problem such as "k-anonymity", "differential privacy" and the use of synthetic data (Henle et al., 2019) or blockchain security protocols (Smith, 2019). There is increasing pressure to address these issues as the ethical choices of companies and governments involving invasive data collection practices have come under increasing public scrutiny in a post-Snowden era.

Data collection and storage practices raise ethical concerns that are part of the overall impact of AI. Online gig economy workers are often paid well below minimum wage to label data sets used to train AI (Semeuls, 2019). Research has also revealed the high carbon footprint created by running computations on massive sets of data. One study found that training a single AI model emitted the carbon equivalent of five cars over their lifetime (Strubell, Ganesh, McCallum, 2019). Anatomyof.ai is a visualization that maps an AI-ecosystem's full impact from the minerals mined as inputs for technology to the disposal of the e-waste at the end of its lifecycle (Anatomyof.ai). These externalities raise important ethical issues to consider as part of the total cost of producing and using AI. These are some of the reasons why data is a crucial element to consider in developing ethically aligned AI. However, data is not the only area that can present ethical challenges for AI.

Making Mathematical Models

It might seem surprising that mathematical models are not purely objective. In *Weapons of Math Destruction*, Cathy O'Neil makes the case that mathematical models are far from neutral, but rather "opinions embedded in mathematics" (O'Neil, 2017, p. 21).

"In computing, a model is the equivalent of a metaphor, an explanatory simplification" (Lohr, 2015, p. 160). There are different schools of thought as to which types of mathematical models to apply in AI as Pedros Domingos highlights in the *Master Algorithm*. His exploration of the five tribes of machine learning demonstrate how values and design choices inform mathematical models. Symbolists value logic and inverse deduction while connectionists reverse engineer the brain with backpropagation and neural networks (Domingos, 2015). Decisions are made about how much "noise" to eliminate, what is considered an outlier in the model, and to determine acceptable levels of trade-offs between accuracy, generalizability and explainability (Domingos, 2015). Popular algorithms are also reused across applications, further amplifying and encoding their reach. Underlying all of this, is tacit agreement that efficiency, prediction and optimization, the rationale to build AI, are worthy goals. Those higher order values both drive and are driven by the ability to harness big data. It enables a "data-religion" which sees data as scientific, objective proof by which to explain the world (Harari, 2016). Like religion, the tribes within machine learning carry ideological perspectives that are encoded within the mathematical toolsets they choose to apply.

AI cannot always explain its decisions. Part of the power of deep neural networks, which is the focus of much current AI research, is that they learn on their own, in a "black box" that even their creators can't fully grasp (Marcus, 2019). That does not instill a lot of confidence and trust in using AI, especially in high stakes environments like healthcare. It also poses problems from a legal standpoint in assigning responsibility and accountability. Recently, even deep learning pioneers like Yoshua Bengio, have been candid about the need to work on understanding the "why" behind deep learning (Knight, 2019). However, some researchers wonder if there is a double standard being applied to machines when in many cases, humans cannot fully explain their decision-making processes (Zerilli, Knott, Maclaurin & Gavaghan, 2018). Finding a reasonable balance for how much explainability is required from an AI system might be context dependent and require discussions about how and where it will be used and whether it's intended to support human decision making or replace it. Choices made in the model can lead to more or less explainability making this an important area for ethical discussion.

There is a need to design affordances into AI systems. Affordances speak to how we interact with objects based on their capabilities. We encounter affordances in the digital realm when we're faced with online forms that force answers to questions, categories in a drop-down menu or stereotypically gendered characters in a video game (Wittkower, 2017). These decisions may result in "disaffordances" that "fail to recognize differential embodied experiences" and include individual attributes like "race, gender, disability, and religion" (Wittkower, 2017, p. 2). IDEO, a leading design firm, thinks that human-centered design principles might be a way to address this problem in AI. They believe the design community can help "figure out how to apply the skills and abilities that data scientists typically have in service of people's needs" (Budds, 2017). Ensuring that AI is designed ethically will require data scientists to think about and design the affordances of their systems. The literature challenges the notion that math is neutral, impartial and objective. Instead, it illustrates how choices made inform mathematical models, which leads us to consider, who makes those choices?

Who Makes AI?

Beyond data and models, the people who build AI warrant consideration as they are the designers who's choices shape technology. Some ethical issues in AI stem from the very conception of the problem. How problems are framed and the corresponding solutions encoded within AI systems links directly to the people involved in the design process. A recent report by AI Now which looks at the issue of inclusion and diversity notes that "as the focus on AI bias and ethics grows the scope of inquiry should expand to consider not only how AI tools can be biased technically, but how they are shaped by the environments in which they are built and the people that build them" (West, Whittaker & Crawford, 2019, p. 6). According to a global AI talent survey, there are 22,400 people working in AI based on those who publish research, however 36,524 people self-report as AI specialists on Linked In (Mantha & Kiser, 2019). It's an elite community comprised primarily of highly educated white men.

The demographic backgrounds of those who develop technology play a role in how and what technology is developed. There are numerous studies and reports about the lack of gender diversity in the field of computing science. AI is even less representative with women making up only 16% of AI researchers (Mantha & Kiser, 2019). Silicon Valley likes to present itself as a meritocracy and has framed the gender diversity problem as a "pipeline issue" of not having enough qualified female candidates (West, Whittaker & Crawford, 2019). Yet, the masculinization of computing work was a long process that edged out women, as programming shifted from a "low-status, feminized task to work that was seen as central to control of corporate and state resources" (Miltner, 2019). Decades of research also shows that who makes a technology impacts what is constructed, and that gender plays a role in creating and reinforcing stereotypes through technology design (Cassell, 2001).

In *Algorithms of Oppression*, Safiya Noble challenges the notion that commercial search engines, fueled by powerful AI algorithms, provide an unbiased and value-free service. She presents numerous examples that highlight Google's racial and gender biased search results, particularly as they impact Black women. Search results seek to further entrench and protect the interests of those already in power, misrepresent marginalized groups and keep women and minorities locked out of participating in the creation of technology (Noble, 2018). Only "2.5% of Google's workforce is Black" (West et al., 2019, p. 3). "The diversity problem is…most fundamentally about power. It affects how AI companies work, what products get built, who they are designed for and who benefits from their development (West et al., 2019, p. 5).

In addition to individual traits, the socio-cultural environments in which people operate is influential. One example is the relationship between funding and the direction of research. Industry generally focuses on research with commercial applications which can mean that areas important to the public good, but not commercially viable, such as public health, may receive less attention (Fabbri, Lai, Grundy & Bero, 2018, p. e9). This has consequences, as we are seeing how chronic, systemic underfunding of public health infrastructure has contributed to a lack of readiness to address the COVID-19 pandemic at the expense of many lives being lost (Maani & Galea, 2020). Universities are not only looking to corporate sources for funding, they are also becoming increasingly interested in holding patents to commercialize research, leaving scientists with an escalating imperative to engage with market or economic considerations (Slaughter, Archerd &Campbell, 2004). This creates a host of ethical “quandaries” to be navigated such as conflicts of incentives between publishing vs patenting, financial conflicts of interest and ethical considerations around student labour and the use of public funds to advance private intellectual property (Slaughter et al., 2004). AI is attracting significant investment from both government and industry making questions about funding and the associated ethical quandaries timely. The social construction of technology is informed by personal demographics, mixed with factors like funding, that can impact not only how AI is built, but what gets built and who owns it.

Finally, it is also important to reflect on the overall project and ask is the purpose ethical? A controversial study of an AI system that uses facial recognition to determine sexual orientation raises questions about purpose (Burdick, 2017). Regardless of whether this system works or not (which is debatable) do we want to live in a world that uses of this type of technology? The authors of the study say they conducted this research in order to expose discrimination in existing facial recognition algorithms (Wang & Kosinski, 2018). However, their research received considerable backlash, particularly from LGBTQ communities who might be harmed by this technology (Burdick, 2017). Questions of purpose and stakeholders impacted should be central considerations for researchers. Working proactively with AI researchers to have frank discussions that surface ethical issues and "blind-spots" while acknowledging the socio-cultural context in which they operate can enable necessary adjustments to the development workflow or at least acknowledge where potential gaps exist.

Part Three: Healthcare and AI

Lower costs, improved patient outcomes and reduced workloads for clinicians – these are the promised outcomes that represent an AI "holy grail" in solving the global healthcare crisis. Companies like Babylon Health, whose mission is to make healthcare affordable and accessible to every person on earth, are tapping into the power of big data and algorithms to deliver better care by tracking and monitoring health information in real time (babylonhealth.com). AI is shifting both the approach to health research and healthcare delivery. "Instead of extrapolating from the data obtained from a small number of samples…we can now use clinical data at the population level to provide a real-world picture" (Ngiam & Khor, 2019, p. e263). Medical data sets include genetic information, imaging scans and real-time outputs from wearable sensors all of which needs AI in order to turn it into useful information (Topol, 2019). In the US, the Food and Drug Administration (FDA) has already approved the use of algorithms for a range of medical interventions in areas as diverse as monitoring heart health, early intervention of stroke and detection of wrist fractures (Mason, Morrison, & Visintini, 2018). In Canada, Radiology, Pathology and Dermatology have been identified as key areas where AI can play an important role as they align well with machine learning techniques focused on image detection and pattern recognition (Mason et al., 2018). The upside potential for AI in healthcare is enormous, but there are downside consequences that need to be addressed, as the following examples illustrate.

Discriminatory data and poor design. Small errors in AI systems can have life and death consequences in the realm of healthcare. A popular commercial healthcare algorithm was shown to be biased in recommending less healthcare coverage for Black vs White patients because of its use of healthcare spend as a proxy for the need for healthcare (Obermeyer, Powers, Vogeli & Mullinathan, 2019). This particular algorithm was rebuilt with new criteria that better represented need and it saw a jump in recommendations for care from 18% to 47% for Black patients (Obermeyer et al., 2019). This error and the harm caused may have been avoidable with processes that reviewed models with an eye towards ethical concerns.

Privacy and consent. The values of privacy and consent deserve special attention when it come to healthcare data. The sharing of 1.6 million patient records by a UK hospital as part of a partnership with DeepMind speaks to the ethical concerns surrounding the need for greater transparency and consent of data use for the development of AI healthcare applications (Hern, 2017). On the flip side, AI health researchers, such as Marzyeh Ghassemi, have argued that lack of big datasets for research impedes progress and can lead to biased outcomes as inadequate or incomplete datasets are used to train algorithms (C4E Journal, 2019). Determining how to best navigate the balance between preserving individual rights while enabling societal gains needs to be explored with respect to healthcare data. This is an ethical dilemma that deserves consideration from a broad set of stakeholders.

Impact on care relationships. Perhaps a more subtle issue is how AI impacts the clinician patient relationship. For example, a system can be used to either promote or deter certain ethical ideals, such as shared decision making in healthcare. There is a danger in deferring to a system, which may have one pre-determined way of ranking treatment options, not considering patient values or preferences (McDougall, 2019). This approach can undermine patient autonomy and participation in decision making, or conversely, if systems consider individual preferences as part of the design process, they could be used to enable and foster this ideal (McDougall, 2019). Design choices can be made in AI systems to promote values that are deemed beneficial.

Unanticipated uses and unintended consequences. There are also risks associated with using AI in medicine when the technical systems are working as planned. Adversarial techniques can be used to manipulate data in order to trick AI models, exposing healthcare systems which rely on these models to creative billing practices and insurance fraud (Finlayson, Bowers, Ito, Zittrain, Beam and Kohane, 2019). These same models might be deliberately fooled by clinicians for more honourable reasons. For example, a doctor might choose to key in special data that could over-ride a patient's algorithmic risk score that denies them a prescription for opioid pain killers if the doctor were convinced the patient desperately needed the prescription (Finlayson et al, 2019). These are just two examples of how AI systems can be subjected to misuse.

This is not an exhaustive list but a few examples that highlight ethical concerns involving AI in healthcare. Outside of clinical settings, consumers are monitoring their own health and wellness, through wearable devices like Fitbits or mobile apps that monitor everything from mental health to menstruation. These non-clinical devices are fueled by consumer demand and often receive little if any regulatory oversight, making them particularly vulnerable to misuse or unintended consequences. These consumer health technologies are compiling massive amounts of "granular personal health data" which is then "leveraged to inform personalized health promotion and disease treatment interventions" (Nebeker, Torous & Bartlett Ellis, 2019, p. 1).

Technology is changing the nature of how we think about and deliver healthcare. In Alberta, we are in the early days of implementing a new electronic health records system called Connect Care which brings together data from over 1,300 siloed software systems (Gerein, 2019). Connect Care is a one stop platform that houses all patient records, giving clinicians a single point of access to information, sharing information amongst a patient's care team, as well as providing decision support alerts about vital stats and other information directly from patients in real-time (Gerein, 2019). It is a big step towards aggregating the data needed to deliver personalized, AI-enabled healthcare, making a discussion of ethics particularly timely for AI researchers in Alberta working in the healthcare domain.

The collection and use of healthcare data accelerated in the first quarter of 2020 in the wake of the COVID-19 pandemic. In the name of public health and safety, actions that seemed unthinkable are now being quickly implemented, often with little or no regulatory oversight or opportunity for public discussion. For example, temperature checks (biometric data) are being implemented in airports, workplaces and retail settings in Canada as a means to screen people for possible infection (Steacy, 2020; Puckett, Eckhard, 2020). This has prompted privacy experts and human rights activists to raise concerns about overreach and discrimination (Burke, 2020). In Alberta, a virtual care app offered by Telus and Babylon Health, was endorsed by the provincial government, offering virtual doctor's visits paid for by Alberta Health (Bellefontaine, 2020). It was implemented without a privacy impact assessment in order to quickly bring it to market and provide healthcare services in keeping with physical distancing. Alberta's privacy commissioner has now launched an investigation (Bellefontaine, 2020). The Government of Alberta was also the first province to launch a voluntary contact tracing app called Alberta TraceTogether built on technology used in Singapore's contact tracing app (Mertz, 2020). While some countries such as China and South Korea have seen success in using mandatory contact tracing apps along with other surveillance tools to help control the coronavirus, critics question the value of these apps and whether they are "practical, accurate and technically capable" (Kelion, 2020; Zastrow, 2020).

AI is a powerful tool and while there are reasons to be critical and careful in our use of it, choosing not to apply it, especially in the domain of healthcare, also poses a moral dilemma. If AI can deliver on even some of its promises to save lives and improve patient outcomes, there is an opportunity cost in underusing it (Floridi et al, 2018). We need to be willing to explore how we might build and deploy AI in ways that reduce risks while delivering benefits. Having these conversations at the design stage with AI researchers and other stakeholders is a practical first step.

Summary of Literature Review

"We know AI is going to change the world, but who is going to change AI?"
-Dr. Fei-fei Li, Stanford Human-Centred AI Institute

The ethical alignment of AI is an important and timely issue. The literature illustrates that AI systems are currently shaping our world and all indicators point towards an increased use of AI technologies. People are already being harmed by certain implementations of AI technology. Often, it's the most vulnerable and marginalized communities who are subject to the greatest harms, but as the world relies more on AI, no one remains immune. The literature is filled with critiques about harms or potential harms but is light on solutions. Ethical codes can play a role in addressing and mitigating some of these issues. However, these codes are currently very high level and need to be made actionable. There is a gap between the ideals of ethical codes and the reality of building AI in practice. There is also a lack of inclusivity. AI researchers, industry, academics and policy makers are driving conversations about AI and ethics with little or no public input. AI will impact everyone, and more work needs to be done to integrate viewpoints from both experts and laypeople as we explore ethical considerations. AI researchers, as the designers of AI systems, have a key role to play in ensuring that AI is designed to align with ethical concerns. This is perhaps most urgently needed in high stakes domains such as healthcare. The risks of continuing to build AI without examining ethical consequences holistically are serious. Yet, not harnessing the potential of AI carries an opportunity cost in unrealized benefit. We need to explore how we can apply AI in ways that align with human values and advances societal good. Ethics that are made actionable can play a role in enabling that future.

There is little research taking place at the ground level to implement ethical practices within AI design and development, especially within healthcare. My research seeks to address this gap by working with AI researchers at the University of Alberta who are building systems in the healthcare domain to better understand and document ethical considerations in their workflow in order to make recommendations to strengthen ethical practices. I will also develop a toolkit that can become a resource. Instead of framing ethics in terms of a checkbox within a process, this approach seeks to inform the overall process. This toolkit will take a holistic approach to ethical considerations, something that is not currently being done with more general AI ethics frameworks. The toolkit will be informed by data gathered from this study.

> “As engaged intellectuals we understand that we are entangled within world systems of oppression and exploitation...Our choice is to stand alongside or against domination, but not outside, above or beyond it.” (Frey & SunWolf, 2009, p. 41).

I am inspired by these ideals of engaged research, to make a small contribution that pushes back on the world systems in which we are all entangled and perhaps helps reshape them in a beneficial way.

Research Design and Methodology

AI systems are constructs of the choices made in their design. They reflect the biases of the data sets and mathematical models used by their creators. They are also influenced by the greater socio-political environment in which they are created. If we introduce and apply ethical tools to examine blind spots during the design process, we can create better outcomes. Ensuring that we have ethically aligned AI is important for achieving the trust and social license needed to apply AI in high stakes fields such as healthcare. There is little research taking place at the ground level to implement ethical practices within AI design and development. My research addresses the question:

> **How can AI researchers working on healthcare related projects use ethics tools to inform their research process?**

My study employs a qualitative research approach involving semi-structured one on one interviews with four PhD students and one recent masters graduate at the University of Alberta, all of whom are working on AI research projects in the healthcare domain. I used a short pre-interview survey administered through a secure online form in order to gather basic background data that was used to help me familiarize myself with the researchers and their research projects as preparation for the interview. In addition to notes transcribed from the audio recordings of one on one interviews, I also conducted a focus group with the participants where we reviewed three ethics tools to examine their usefulness in assisting AI researchers with addressing ethical considerations. The focus group was also recorded and then transcribed into notes. Both one on one interviews and the focus group were conducted using an online video platform (Google Meet). Data was analyzed through qualitative content analysis and used to make recommendations for guidelines and tools that can help AI researchers working in the healthcare domain to address ethical considerations in the course of their work. This chapter outlines my research design, methodology, data gathering strategies, data analysis and research process in more detail.

Design

"Research is a process of making good arguments" and using claims, data and warrants to tie those arguments together in a logical and cohesive process (Merrigan, Huston & Johnston, 2012, p. 9). The focus of my research is to document ethical practices currently being followed in the design of AI systems for healthcare applications, identify gaps, develop ethics tools and to make recommendations for ethical guidelines. These objectives speak to a qualitative research design which allows for an individually tailored and nuanced approach to understanding the experiences and practical realities of my research participants. In addition, I've chosen to follow-up my one on one interviews with a focus group involving all participants in order to review a selection of ethical tools intended to help AI researchers. Focus groups are useful for "sharing and comparing" in order to better understand if there is agreement amongst the group members as to the usefulness of these ethical tools (Denscombe, 2010, p. 353).

While I briefly considered ethnography, or focused ethnography, as a method to truly immerse myself in understanding the AI researcher's workflow and how ethical considerations are addressed in the course of their work, the practical constraints of my time and the ability to gain the level of access needed to conduct either of these methods tempered my ambitions. That being said, my research is built upon a foundation of what might be considered a version of ethnography. In the year leading up to my research project I immersed myself within the local AI community. It started with a self-directed course on ethical codes supervised by University of Alberta AI researcher and professor, Dr. Jonathan Schaeffer, which gave me valuable insights into Alberta's AI research community. I signed up for the community mailing lists, went to the local meetups, talked with numerous computer science professors and students, networked with industry professionals, attended conferences, engaged with CIFAR and Amii, took a course about machine learning and made a podcast about it, toured the local offices of DeepMind, wrote articles about some of Alberta's AI research pioneers and joined two AI ethics reading groups. While all of this isn't objectively part of the study, it was important preparation for doing this work and enhanced my credibility as a researcher.

I also found myself conducting research during the COVID-19 pandemic which resulted in the need to limit my data collection methods strictly to online platforms that adhered with social distancing requirements. In considering my objectives which were exploratory and seeking to understand a phenomenon rather than explain it, qualitative description was the best method to approach my research. Though qualitative description is sometimes deemed a "less sexy" approach, it is a sound method that offers "a comprehensive summary of an event in everyday terms" and "an accurate accounting of events" (Sandelowski, 2000 p. 336). The outcomes of this study provide an important baseline and foundation for further research which is in keeping with a discovery paradigm.

Participants

Most AI researchers in Alberta are based at the University of Alberta. Using a purposeful sampling strategy to intentionally focus on a target group, I recruited four doctoral students and one recent graduate from the computer sciences masters program all of whom are conducting AI research in healthcare (Merrigan et al, 2012). I originally chose to limit my study to University of Alberta AI researchers because I had planned to conduct in person interviews. Though my data collection moved online due to COVID-19, which opened up the possibility of extending my geographic reach, for practical purposes, I felt that it would be easier to recruit participants based on the relationships I had already built with this community over the past year.

Sampling Method

A purposeful sampling strategy, which is a non-probability sampling method that targets a specific group who can offer the best insights into a particular topic, was best aligned with my intentions to study AI researchers in healthcare (Etikan, Musa, & Alkassim, 2016).

Inclusion Criteria. My inclusion criteria were affiliation with the University of Alberta as a PhD student working on an AI research project in the healthcare domain. Focusing on students provided a baseline level of group homogeneity in terms of shared experience. Students are also the future of AI and most likely to benefit from ethics training.

Exclusion Criteria. I excluded non-University of Alberta affiliated AI researchers as well as professors or industry professionals whose academic training and life experiences would differ greatly, thus further dividing an already small sample.

Given the small size of the group of AI researchers in total coupled with the very specific inclusion criteria of working on a healthcare project, attempting to meet quotas based on other factors, such as demographics, was not practical. Having said that, I was able to interview both male and female participants while not adhering to specific quotas.

I started with a purposeful sample, however, network or snowball sampling, whereby person A suggests talking with person B, played a role in my recruitment process, as did elements of convenience such as geographical proximity and willingness to participate (Etikan, Musa, & Alkassim, 2016). Leaning into these variables was a useful way to connect to this relatively small community, an approach that became even more important as recruitment during a pandemic proved more challenging than I had anticipated. Appendix A contains a sample recruitment email.

Setting

I used an online video platform to conduct interviews due to public health advisories that resulted in the closure of most public spaces, the need for social distancing and university policies to uphold these practices. After reviewing several online platforms for security and privacy concerns, I selected Google Meet, which is supported by the University of Alberta (Schneier, 2020). The interviews were recorded and transcription software was used to help generate text-based notes which formed the primary source material for a qualitative content analysis.

My original intention was to conduct in person interviews in a field setting, such as the AI research lab or a participant's office, to allow for informal direct observation of the participant's work context. In person interviews are the gold standard for qualitative research and conducting an interview remotely is typically "not a preferred way" (Johnson, Schietle, Ecklund, 2019, p. 2). However, in person interviews were simply not feasible given the timeframe I had allocated for data collection while adhering to the public health advisories put in place due to the COVID-19 pandemic. I understand that there may be a trade off to the data collected using an online video platform due to a "mode effect" that might have resulted in less rich or detailed data (Johnson et al, 2019). At the same time, it is possible to provide for a high degree of researcher and participant satisfaction with a qualitative research interview conducted through online video platforms. One study found Zoom was rated highly by participants for forming and maintaining rapport with the researcher (69%), convenience (56%) and user friendliness (56%) (Archibald, Ambagtsheer, Casey & Lawless, 2019). I suspect that given my participants were AI researchers with a high degree of comfort with using technology, mode effects from using an online video platform may be minimized.

Instrument

The primary source for data collection was self reported information gathered through one on one interviews. These were semi-structured interviews which allowed for some guidelines while also providing fluidity for participants to engage and lead the conversation in meaningful directions. I selected interviews as a data gathering strategy as it aligns with a qualitative approach and supports “claims describing group values, beliefs and practices” which fits with my research question (Merrigan et al, 2012, p. 68 Fig 5.2). Interviews allow for real-time dialogue which can provide a deep and rich data set. They are useful for understanding opinions, experiences or feelings and exploring sensitive issues or gathering privileged information (Denscombe, 2010). There can be challenges with interview data. As one critic notes, “what people say in an interview will indeed be shaped, to some degree, by the questions they are asked; the conventions about what can be spoken about; by what time they think the interviewer wants; by what they believe he/she would approve or disapprove of” (Alshenqeeti, 2014 p. 43). Any data gathering strategy will have some limitations and it is important to be aware of these when conducting research. However, given my research question which is centred around documenting an AI researcher’s personal experience in applied ethics, one on one interviews are the best fit for my project.

I designed my interview guide to be comprehensive, yet also allow for flexibility. Appendix B lists some sample questions. I used a brief, pre-interview online form in order to familiarize myself with the interviewee’s research prior to the interview and whenever possible, I also met with participants in advance to review the consent form as a pre-interview icebreaker. I took this approach because online platforms do not allow for natural small talk opportunities in the same way that in person interviews do and I wanted to foster trust and a comfortable environment for the actual interview. I structured the interview guide around these key areas: ethics in general, data, models and ethical codes, tools and training. I started with broad, general questions before moving towards more specific details as recommended in the literature (Mayan, 2009). Prior to finalizing the interview guide, I sent the questions to my supervisor and a computer science professor and adjusted it based on feedback as part of my process to "foster content validity" with the measurement instrument (Merrigan et al, 2012, p. 88). I ended all interviews with an open-ended question "Is there anything else that I should have asked you about?" which is a recommended best practice (Boyce and Neale, 2006).

In addition to the one on one interview data, I also conducted a focus group with the participants. I chose to conduct a focus group because I wanted to not only gather feedback about ethical tools, but also to better understand "the extent to which there are shared views among a group of people in relation to a specific topic" (Denscombe, 2010, p. 177). Since ethics are normative, group response is relevant to understanding what might be accepted as a useful ethical tool. I chose two ethical tools from the AI Ethics Applied Typology, a database of over 100 publicly available ethical tools (Morley, Floridi, Kinsey & Elhalal, 2019) and one tool that was developed by the Berkman Klein Institute and MIT Media Lab (Calderon, Taber, Qu & Wen, 2019). My selection criteria were informed by an initial analysis of the one on one interviews. I selected tools I believed would be relevant to all participants, that worked across a range of different types of projects (i.e. image data vs text), addressed three distinct ethical concerns and did not require lengthy technical explanations. The following tools were selected:

Datasheets for Datasets: A tool that documents data provenance based on the idea of electronic component datasheets (Gebru, Morgenstern,Vecchhione, Vaughn, Wallach, Lii & Crawford, 2018).

Principles for Accountable Algorithms and Social Impact Statement: A guide for writing an algorithmic social impact statement in order to provide transparency (Diakopoulos, Friedler, Arenas, Barocas, Hay, Howe, Jagadish, Unsworth, Venkatasubramanian, Wilson, Yu & Zevenbergen, n.d.).

AI Blindspot: A general purpose set of flashcards used to foster discussion of the whole AI workflow (Calderon, Taber, Qu & Wen, 2019).

A more detailed description of each tool is available in Appendix C.

During the focus group, participants were presented with a set of thought-starter questions (outlined in Appendix C) and encouraged to have a free-form discussion about each tool, after which the tool was rated on a Likert scale where 1 = not at all useful and 5 = extremely useful. The focus group was recorded, and notes transcribed for analysis.

Procedures

I submitted my ethics application for approval in early March and was granted approval by early April. Through a combination of leveraging network relationships that I had built with the AI research community over the past year, working through professors who specialized in AI research in healthcare and referrals from participants (snowball sampling), I was able to recruit five participants who agreed to take part in my study. I had initially hoped to recruit 6-8 participants, however, I found it more challenging to connect with this community during the pandemic and I prioritized my need to stick with the planned timeframe to complete the study over further recruitment efforts. The recruitment challenge also played into a decision to allow a recent University of Alberta graduate from the master's program in computer science, currently working on an AI healthcare application, to participate in the study. I also recruited one participant who met all of the criteria of the study but who was not a computer science major. I realized that I assumed that anyone working as an AI researcher at the PhD level would be a computer science major, however, this was not necessarily the case.

The study was structured in three parts: a brief pre-interview questionnaire administered through a secure online Google form which took approximately15 minutes to complete, a 45 to 60 minute interview using an online video platform (Google Meet) and a 90 minute online focus group (Google Meet). The pre-interview questionnaire was used to gather background data about the participant's work, such as links to their research website or published papers. By having some familiarity in advance with a participant's project, I was able to better engage with them during the one on one interview. Participants were directed to the secure Google form containing these questions upon their return of the signed consent form.

One on one interviews took place in April and May 2020. All one on one interviews were completed prior to the focus group interview which took place in late June 2020. The one on one interviews focused on gaining a better understanding of the ethical practices and considerations participants were already implementing in the course of their work and what issues they felt were important. The interviews also helped me to understand participant's perspective and knowledge level with respect to ethical considerations within AI. Recordings of interviews were turned into text-based notes through a combination of transcription software and manual editing. Notes were sent back to participants for their review and approval as part of the validation process (Denscombe, 2010). In keeping with exploratory research and the ideal of staying close to the data, I took care to keep notes in the participants own words while lightly editing transcripts for clarity and to remove any sensitive or identifying information (Mayan, 2009).

A preliminary analysis of one on one interview data was conducted in order to uncover common themes and areas for further exploration in the focus group. The participant interview data helped drive the selection of the three ethical tool sets to explore in the focus group. I sent all respondents a PDF copy of the three ethical tools three days in advance of the focus group so that they could print the information if they wished, given that the group was meeting online. This was also a chance, for those who wanted, to review the material in advance. I produced a PowerPoint presentation to provide a visual touchstone for the discussion (Appendix C). Following best practices, participants were reminded of their commitment to confidentiality and that given the nature of the focus group, their data could not be excluded from the study after the focus group concluded (Denscombe, 2010). I used a series of thought-starter questions (Appendix C) to foster group discussion, allowing me to moderate and facilitate, rather than ask a series of questions, which is a distinguishing factor between focus groups and group interviews (Denscombe, 2010). The 90-minute discussion was hosted and recorded on Google Meet. The recording was transcribed with the assistance of transcription software and text-based notes were used in the final analysis. All sensitive and identifying information was removed from the notes.

Analysis

I used qualitative content analysis to analyze my interview data which was captured as an audio recording and turned into text-based notes. “Research using qualitative content analysis focuses on the characteristics of language as communication with attention to the content or contextual meaning of the text to provide knowledge and understanding of the phenomenon under study" (Hsieh, 2005, p. 1278).

The process of content analysis involves the careful reading and coding of the texts to find themes and patterns that form relevant classifications and underlying patterns in the data (Mayan, 2009). Qualitative research interviews are typically analyzed using an iterative process of data collection, analysis, data collection, analysis, except in cases where many semi-structured interviews ask the same set of questions of each participant roughly in the same order (Mayan, 2009). Since I was using semi-structured interviews with questions asked in roughly the same order, I chose to collect all of the interview data and then analyze it. I used my questions as preliminary categories, refining the analysis to find higher level themes as part of an inductive process (Mayan, 2009). I was also able to apply an iterative approach by reviewing the one on one interview data in order to help guide the direction of the focus group. The focus group data was also analyzed using classic content analysis but using the group as the unit of measure (Onwuegbuzie, Dickinson, Leech, & Zoran, 2009). Similar to the interviews, I categorized the data into higher level themes. I also made note of any relevant group interactions. One draw back of the online format is that body language and eye contact between participants was lost, especially as most participants felt more comfortable turning their cameras off. I was able to note tone of voice, order of speakers, who engaged with whom and frequency of engagement as additional contextual information. Finally, I used text analysis software (Voyant Tools) to analyze just the focus group respondent data in the transcript in order to create a visual representation of the focus group. The texts for the three tools were analyzed both individually and together.

Validity and Reliability. Reliability, which is focused on the “consistency of measurement over time” and validity, which relates to the “accuracy of measurement” are important to ensure a research study could be replicated with similar results (Merrigan et al, 2012, p. 85 and p. 301). These can be challenging factors for qualitative researchers to address, as so much of qualitative research is context dependent. Qualitative researchers often aim, instead, for credibility, or the degree to which research is trustworthy, as a way to assess the quality of a study (Pandey and Patniak, 2014). There are several recommended methods to ensure credibility, such as prolonged engagement, persistent observation or triangulation, whereby either multiple sources of data, different theories or multiple investigators compare and contrast findings (Pandey and Patnaik, 2014). These methods were not feasible given the scope, resources and timelines for my project. Instead, I verified interview notes with participants as a means to validate my data through a “member check” (Merrigan et al, 2012). This was a practical way to address some of the issues surrounding reliability and validity given limited resources and is considered by Lincoln and Guba (2006) to be the most important way to strengthen credibility.

Limitations. In addition to the challenges of addressing credibility through recommended methods due to a lack of resources, my study is also constrained by a small sample size and selection bias with respect to the use of purposeful sampling, both of which limit generalizability.

Researcher bias was another factor that I was keenly aware of and while I tried to be consistent in my delivery of the interview questions, there were inevitable differences resulting from circumstances. For example, during one of my interviews, a fire alarm went off in the participant’s building, which forced the interview to an abrupt conclusion. I do not know how or if the quality of the data collection was impacted by conducting this particular interview in two parts, however, the incident reminded me that it is impossible to control for every variable. There is also debate in the literature as to whether a researcher should embrace contextual factors including researcher bias as an inherent strength of qualitative research (Mayan, 2009) or view these factors as problems to be mitigated (Merrigan et al, 2012). My own view is aligned with a pragmatic approach of "subtle realism" which aims for positivistic scientific rigor while acknowledging "the insights of post scientific concepts" (Denscombe, 2010, p. 298).

Given that very little research has been conducted about how AI researchers implement ethical practices in their work and there is not a history of prior theoretical approaches, an exploratory design using conventional qualitative analysis was the best fit. However, I am aware that this will limit "both theory development and description of the lived experience" and thus, the results might be limited to a first step in building a concept or model (Hsieh et al, 2005, p. 1281). Even with that limitation, I think this is an appropriate choice as my research aims to be descriptive rather than interpretative. Overall, I attempted to address known issues while respecting the limits of my resources and as such, this work is best viewed as preliminary inquiry into a topic that would benefit from further exploration.

Summary of Research Design and Methodology

Determining an appropriate research approach considers many elements. There is the research question itself, theoretical perspectives and personal attributes of the researcher, and practical considerations such as time, resources and access to participants. There are also ethical considerations. I tried to anticipate and account for all of these areas in my research design. However, research does not go exactly as planned. I had to make several adjustments to my originally intended process in order to react to world events, in particular, COVID-19 and the public health advisories preventing in person meetings. Under the circumstances these were necessary trade-offs, and I feel that moving to an online platform did not adversely impact my research design. In some cases, such as having the ability to record discussions, or finding a common meeting “place”, it made conducting the research easier.

Findings and Discussion

Introduction

My study used a combination of one on one qualitative interviews with five AI researchers working on healthcare related projects followed by a focus group involving these same researchers to investigate this question:

How can AI researchers working on healthcare related projects use ethics tools to inform their research process?

I used an exploratory approach in order to better understand both how researchers think about ethics and how they apply ethics in the context of their work. The individual interviews helped inform a selection of three ethical tools that were discussed during the focus group to understand the usefulness of each tool for AI researchers in the context of their work.

The findings are presented in two parts: Interview Findings and Focus Group Findings. The interview findings are further segmented into two major categories that address how the participants think about ethics and how ethics intersects with an AI workflow. Summarized feedback outlines major themes related to these areas and is supported by excerpts from the interviews. I also mapped participant feedback against a typical AI workflow.

The focus group findings are presented as summarized feedback along with a usefulness rating for each of the three ethics tools: Datasheets for Datasets, Principles for Accountable Algorithms and Social Impact Statements, and AI Blindspot. I also analyzed the focus group transcript using a text analysis software to create visual representations of this data that provide additional insights. Finally, key learnings, limitations and recommendations for further work are presented in the discussion section followed by a brief summary which concludes this chapter.

Part One: Interview Findings

How do AI researchers think about ethics?

During the interviews, I asked a series of questions designed to explore aspects of ethics in an effort to better understand how participants think about ethical issues. The discussion was wide ranging and included question such as:

- What is the role of ethics in AI?
- Who is responsible for ethical issues?
- What do you see as the biggest challenges or barriers to applying ethics?
- What are areas of ethical concern in your own research?

There was ample opportunity for participants to steer the conversation into areas and topics that they felt were most important to discuss. These are the major themes that emerged with respect to how AI researchers think about ethics:

- **Ethics is important but not necessarily pertinent**. There was unanimous agreement that ethics in AI is an important issue. Everyone shared examples of ethical concerns when ethics was discussed in a general sense. However, when asked about ethical concerns in terms of their own work, most participants said things like “I don’t think about ethical concerns too much at this point” or that they had no ethical concerns about their research. Aside from the potential for misuse, participants felt ethics in the context of their own work had been addressed.

- **A limited perspective of ethics.** The role of ethics was primarily framed in terms of “doing no harm” and promoting “the greater good”. Most participants noted techniques they used to address issues like bias in data from a statistical perspective. Both these comments tend to convey a narrow frame on considering ethical issues. In part, this could be the result of having received little or no formal ethics training. Only one participant who had ethics training was familiar with the principles of bioethics and named the GDPR when asked about ethical codes. Other participants had personal principles that they applied, such as a commitment to transparency. One participant said they were trying to education themselves more about ethics. Participants agreed that including ethics training in their PhD program would be useful.

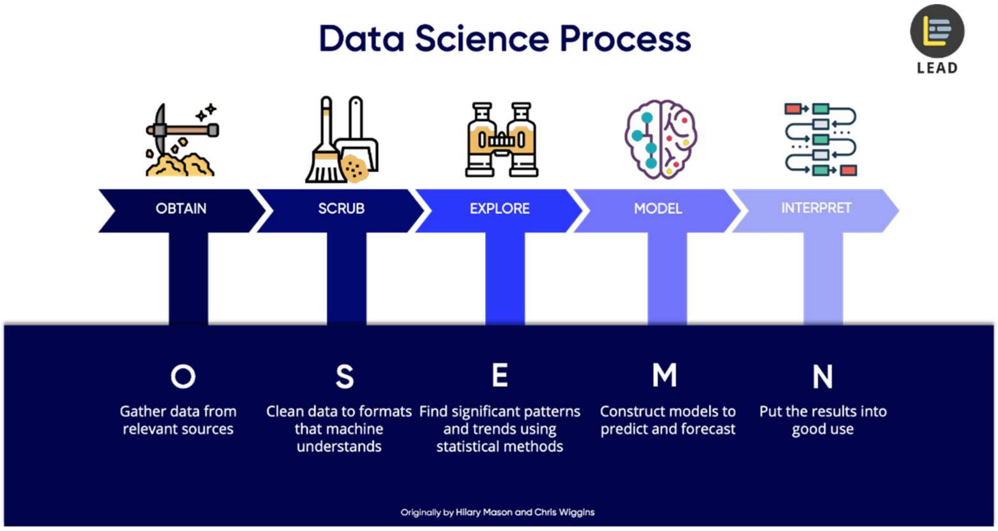
- **Commercial incentives compromise ethics.** When asked about challenges and ethical concerns, participants shared anecdotes about an evil company misusing technology: “let’s say some company would like to use this to track their customers, let’s say some insurance company.” There was a shared concern that commercial incentives compromise ethical decision making and a general lack of trust for “big pharma” or “big tech”. Some participants made a distinction between academia and industry. Academia was seen to have noble intentions, with people working in AI to contribute towards enhanced knowledge, not to “get rich off their work.” When asked about funding for their own research, participants shared that funders had no impact on the direction of their work.

- **Unclear who is responsible for AI.** The question of who is responsible for AI resulted in a version of “everyone and no one”. There were thoughtful responses that covered all the players, from individual developers to corporate executives, those who market or deploy these systems, university ethics boards, governments, and end-users. There was agreement that this is a complicated question.

These themes provide a high-level summary of how the participants think about ethics. Further details are available in Appendix D, in an "at a glance" chart of the summarized interview findings as well as interviewee demographic data from the pre-interview questionnaire.

How does ethics intersect with the AI workflow?

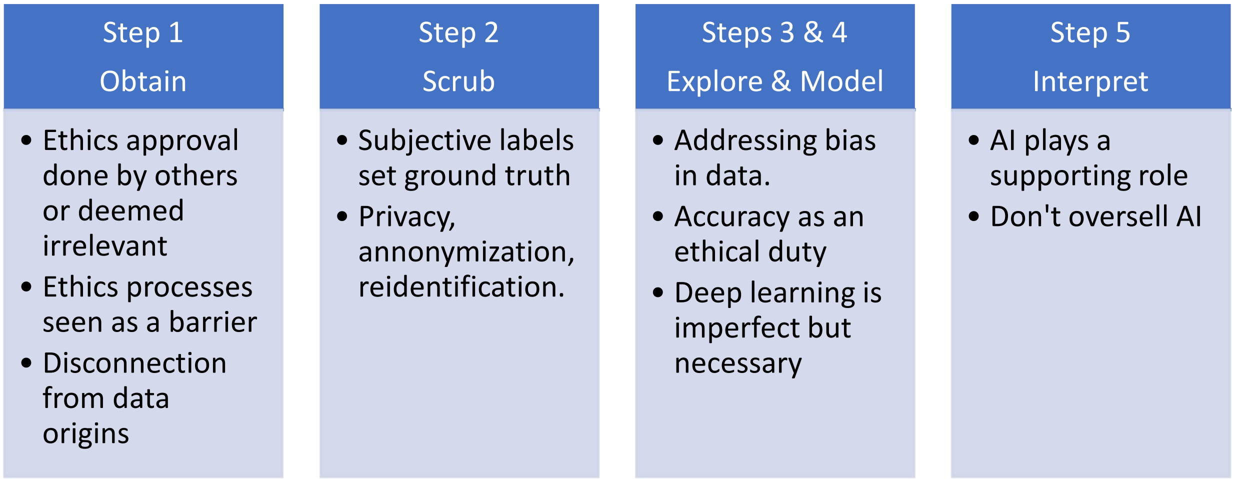
There are several variations on AI workflows. I used the OSEMN framework (Figure B) to think about a general workflow as obtaining data, scrubbing or cleaning the data which also includes pre-processing, exploring, modelling and then interpreting the results.


Data Science Process
LEAD
OBTAIN
SCRUB
EXPLORE
MODEL
INTERPRET
O
Gather data from relevant sources
S
Clean data to formats that machine understands
E
Find significant patterns and trends using statistical methods
M
Construct models to predict and forecast
N
Put the results into good use
Originally by Hilary Mason and Chris Wiggins

Retrieved from https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

Figure B

I grouped summarized participant feedback into themes and mapped it against the OSEMN workflow (Figure C) to show where ethics intersects with the AI workflow. I combined Explore and Model as these steps are often an interconnected iterative process in machine learning.



Step 1
Obtain

- Ethics approval done by others or deemed irrelevant
- Ethics processes seen as a barrier
- Disconnection from data origins

Step 2
Scrub

- Subjective labels set ground truth
- Privacy, annonymization, reidentification.

Steps 3 & 4
Explore & Model

- Addressing bias in data.
- Accuracy as an ethical duty
- Deep learning is imperfect but necessary

Step 5
Interpret

- AI plays a supporting role
- Don't oversell AI

Figure C

The following descriptions showcase high-level themes from the interviews as they align with aspects of the OSEMN AI workflow in order to understand how ethics intersects with an AI workflow. Not all data fit neatly into categories, and I will address that further in the discussion section.

Step 1: Obtain

Ethics approval done by others or deemed irrelevant. I started with the ethics approval process as set forth by the university research ethics board. This approval process both supersedes the workflow and intersects with the "obtain" phase. Most participants were not actively involved in directly obtaining ethics approval and some were slightly unsure if their project had gone through the university approval process. In cases where it did, they said that a supervisor or other colleague, such as a post-doc, handled that process. In many cases, ethics approval was not required for their project because they were using public datasets which fell outside the scope of the university's research ethics board process.

Participants said:

> *"this data has already gone through all the ethical approvals…we just take the data and try new things, different things, we ask different questions"*
>
> *"It's only about analyzing pre-collected data, so that's why we didn't have to go through ethics approval."*
>
> *"research approval is something I don't do directly…they have already done that part"*
>
> *"my research is kind of theoretical"*
>
> *"I'm not 100% sure how we get the ethical approvals and the access"*

One participant who was involved in the research ethics approval process said they were granted "a waiver for the research because we used data from care as usual and we did not change anything."

This indirect or non-existent relationship with ethics approval at the start of the research process means that students received little or no hands-on experience completing an ethics application or working through a research ethics board approval process. It's also worth noting that the need for ethics approval is framed primarily in terms of data collection and not in relation to other factors such as study design or purpose. That might be a limitation of the research ethics process itself which tends to exclude studies that use existing public datasets, particularly when they do not contain any personally identifiable information.

Ethics processes seen as a barrier. In cases where there was a direct engagement with an ethics process, opinions about the frustrating realities of moving through the process stood in stark contrast to praise for the general ideals of ethics. The ethics process was described in terms of hoops that needed to be jumped through to gain access to resources such as health data sets, to comply with regulations, to appease partners such as hospitals or to meet with the approval of advisory committee members. Participants felt the process took too long, contained too many barriers, and often involved "very conservative" non-technical people who "always wanted to be on the safe side." All of this could prove frustrating as one participant shared:

> *"I was just frustrated, and I just think that this is just delaying things because I don't understand what is the rationale behind all of this? What are we gaining by going through all of this process?"*

Other participants said that non-technical people often don't appreciate the practical aspects of machine learning when trying to apply ethical principles:

> *"Why didn't you get consent?...number one, it's publicly available data…number two….we are analyzing tens of thousands [of users]…with that many users, it's impossible to get consent from everyone!"*
>
> *"I have seen FDA procedures…say…"you didn't include any American people"…but would we except American hands [bone structure] to be different than the hands [bone structure] of everyone else in the world…Do we expect to have different bones depending on skin colour?"*

The experiences participants had with the ethics approval process was not a positive one, but rather a frustrating chore leading to a "spoilsport" notion of ethics (Boddington, 2017, p 8). This experience can play into a negative and limited perspective in thinking about ethical issues merely as hurdles to be navigated.

Disconnection from data origins. AI researchers almost exclusively use secondary data in their work. Medical data typically comes from medical research studies, patient records, hospitals or other primary care visits. One participant was using social media data that had been used in a prior study. Participant's data came from a variety of secondary sources:

> *"our data comes from a workers health assessment"*
>
> *"research concludes that Twitter, social media data is useful, and it is publicly available…the two datasets I used are basically a sample of these kinds of data sets"*
>
> *"usually there were academic institutions, the ones who collected the data…and then they publicly release the data"*

Another participant used synthetic data, which is data generated by the researcher. Synthetic data is useful for investigating problems where real-world data is not available such as evaluating two different outcomes for a patient if they were to take treatment A vs treatment B.

> *"the reason why we use synthetic datasets or semi-synthetic data sets…is the problem of counterfactuals…we never observed the outcomes…it's just impossible….in my research…I want to know the best possible treatment for this patient, so I need to know the outcome of treatment one and treatment zero – which one is better?[1]"*

All of the participants were highly cognizant of the "sensitive and personal" nature of healthcare data. They took care to ensure safe data handling practices and followed all privacy regulations to ensure that data was securely stored. Yet, once made anonymous, stripped of any personal identifiers, and abstracted as texts, numbers or images, it was easier to forget that the data connected to people. As one participant noted "it's just images". Once the data is anonymized, it is seen as "safe" to use and the people behind the data are abstracted away. This practice of reusing secondary data disconnects the researcher from the origins of the data.

[1] Synthetic data is used when real-world data is not observable. These are counterfactual events, hypothetical situation whereby we will never actually know what happened.

Step 2: Scrub

Subjective labels set ground truth. Preparing data to make it ready for use in a machine learning application is a very important and time-consuming step in the workflow. A recent report found that close to half a data scientist's time is spent loading and cleansing data (Anaconda, 2020). Though it makes up a significant portion of the workflow, preparing data is viewed as lower level work. "Data preparation and cleansing takes valuable time away from real data science work" (Anaconda, 2020, p. 12). Data labels are important because they set the concept of ground truth, the parameters by which the model learns about the data set. As one participant noted "machine learning is like a child…[it] takes things literally." Thus, the labels for data become incredibly important and the labels are subjective.

The "gold standard" for labelling is human annotation. However, it's difficult to have humans label enough data for machine learning which may need tens of thousands or even millions of instances for training. One participant explained how human labelled data sets can be used to automate labelling in a "bootstrapping process" to "propagate labels…from small data sets to even bigger data sets."

Some participants used pre-labelled data in their work, thus accepting ground truth from the inherited dataset with few modifications.

> *"the occupational health supplier we worked with, they already worked with quite a good program for occupational health assessments, and we looked at it together to determine, is it complete, is it not complete, what could we use more of, is there validated questionnaire measures, that validated the sustainability of a worker."*

One participant noted how labellers for image data are often not diligent in their labelling practices which leads to poor training data for the machine learning application and mistakes that get encoded into the model. They shared:

> *"there is a difference between labelling for a human expert and labelling for machine learning… I have an image of the brain and then somebody is going to identify where the tumor is…so they can put a circle around the tumor. And maybe the circle doesn't cover the tumor entirely. Maybe it only covers half, but the human expert only needs a very loose reference of where the tumor is. …machine learning, it's completely different. If you say whatever it is inside the circle, is a tumor and whatever is outside is not… for the computer, the entire circle is a tumor and everything outside of it is not and it is going to learn from that."*
>
> and
>
> *"There are many people who take the labels as they are and then they train the machine learning system with those labels without really understanding the problem. This is where we have many, many failures. We should be doing this interpretation, but not all of the people do it."*

In addition, this same participant also felt that medical data was particularly challenging to label because medicine itself contains judgement calls or subjectivity, saying "sometimes you present the same picture or the same data to two different physicians and the physicians will disagree and they cannot express why they disagree."

Another participant who also works with medical images of tumors said that despite a two step verification process involving labellers and clinicians, much data has to be discarded because it can't be verified properly and that "you can't really be 100% sure about ground truth" unless a biopsy had been done on the patient."

The participant who used synthetic data was in the position of having to "play God" using math to determine the ground truth saying "in a real-world data set, God knows that objective function. In synthetic data sets, I write that function…and draw samples from it."

In all cases, there is a subjectivity in the form of human judgement made along the way that is encoded in the labelling process.

Privacy, anonymization and reidentification. Participants were confident that the data they were using was sufficiently anonymized to the point that it could not be re-identified. One participant said they were "only using textual data" another that "it's just images" of hearts or lungs. One participant did share concerns about how data in general could be accumulated and aggregated by powerful monopolies, how "they [big companies] have the data at the end of the day" and can use it to construct profiles for their own gain. However, when asked if there was a similar concern when it came to healthcare data they said "not really…there's a lot of extra steps when people are dealing with medical data, thankfully…they are extremely cautious…it would be very, very hard to do."

Steps 3 and 4: Explore and Model

Addressing bias in data. All of the participants were familiar with the issues of attempting to address bias in their data. They explained the use of various statistical methods to address this concern.

One participant shared a bigger concern that "data in the medical field is very biased towards men" and there were few medical studies with female participants.

Another noted how the inherent bias in their dataset worked in conjunction with their research area, saying "in Twitter, 40% of users are young adults…and my research is focused on young adults, so the skew doesn't bother me a lot." They also said that bias could be overcome by "statistical distribution – the more data you have, the more generalization you will get."

A different participant acknowledged that the questionnaire data being used in their work contained bias because "questionnaires by definition are always biased a bit because people answer what they want to and sometimes they have preferable answers." They also shared that they hoped their biometric data might be more "objective and fair."

Finally, one participant had an interesting story about "technical differences in the quality of the data." They said that MRI images gathered using a Siemens manufactured machine are technically different than ones gathered using a General Electric machine and that there are problems when you mix these images into a single dataset:

> "*data is actually different depending on where it is coming from and it's not different because of biological reasons but because of the technical differences that happen because of acquisition*."

This was an interesting take on technical bias that did not surface in my literature review.

Accuracy as an ethical duty. Participants expressed an obligation towards technical accuracy of the system. If they build and test their models for accuracy, there is a sense that they are doing their ethical duty. One participant talked about building technology that they would feel good about having a family member use. Another expressed concerns about ensuring their technology worked for the patients saying "you have to be sure there is nothing wrong…92% accuracy means that in 8% you've got something wrong, and in the medical field that's not necessarily acceptable." At the same time, there are ethical considerations that fall outside a technical scope which can remain unidentified. If these issues are raised, they may be seen as someone else's job to fix or as things that they, as technologists, have little ability to impact.

Deep learning is imperfect but necessary. Most of the participants used deep learning methods in at least part of their work even as they acknowledged that these methods are not fully explainable. While many of the participants are in favour of explainability, it is seen as less important than accuracy. As more participant said:

> *"we started with simpler model, such as linear models, logistic regression models that are very explainable…however these simpler models are not as accurate when trained on the data. That's why people moved on to more complicated models such as neural networks that are less explainable but more accurate."*

This participant went on to mention the field of XAI, or explainable AI. This is a growing body of research aimed at trying to make current AI techniques more explainable to find the holy grail of "accurate, complicated methods that are also explainable." However, we are not there yet.

Another participant who said "one of the pillars of ethically aligned design is to ensure explainability in some way" also shared concerns about "cancelling" deep learning saying "these models are doing very good and just for the sake of transparency…it may not be wise to just cancel them or say that these models are not useful." This comment illustrates the tension between the desire for explainability, a pre-requisite for issues such as transparency, autonomy and legal redress, and the desire to continue using deep learning models.

Only one participant said that explainability was important enough that they chose not to use deep learning methods in their model, saying the reason "we don't use deep learning methods is because….we want to know why a prediction was given." Explainability was also a major topic of discussion during the focus group.

Step 5: Interpret

AI plays a supporting role. Most participants were not at the interpretation stage in their work. However, there was discussion about the role their work should play in this process especially as it relates to medical diagnosis or prescribed treatments. Participants unanimously said that AI should play a supporting role in the clinical process and that it should aid clinicians in decision making but should not be used to replace medical professionals. This is how they described their work:

> *"our model is not saying that you are now diagnosed with depression…it's to be used by the clinician…a companion tool for the clinician"*
>
> *"everything we are building right now is supposed to be used by doctors"*
>
> *"we develop a decisions support tool"*
>
> *"we're creating mostly measurement tools…not diagnostic tools"*
>
> *"AI should not replace healthcare professionals"*

While AI researchers are enthusiastic about the potential benefits of using AI systems in healthcare, they also clearly understand the limitations. They shared concerns about potential misuse of their tool by non-experts and about "function creep", the idea that "the algorithm could be used for other reasons than what it was developed for", such as excluding people from receiving appropriate treatment or insurance coverage.

Don't oversell AI. Participant are concerned that non-technical people, especially those with commercial incentives, tend to oversell AI's capabilities. Participants view part of their duty as technologists as ensuring that this overselling does not happen. As one participant stated, "I think this is also part of the ethical piece, we try not to oversell what AI can do in real settings right now." Another said, "we have people who do not really understand the limitations of the technology and want to go ahead and deploy it before it's actually ready."

Incentive structures related to commercialization were repeatedly identified as an issue. Some felt that companies were racing to deploy technology that was not market ready and that certain AI developers were only interested in "big salaries" and "getting rich off their algorithms." They noted how these commercial goals had resulted in the deployment of harmful AI systems and how that in turn hurt the AI community by creating mistrust and backlash. As one participant noted, “IBM trained models could not recognize Black women…and that was a big hit to the community”. Another said that Tesla was irresponsible in how they had conducted trials of self-driving cars. These comments illustrate participant’s concern to build trustworthy AI, not only for the good of users, but also because it is good for the AI community.

These selected highlights from the interviews serve to paint a picture of how AI researchers think about ethics and how ethical concerns intersect with the AI workflow.

Part Two: Focus Group Findings

During the focus group, the five participants were presented with a brief description of an ethical tool and given some thought starter questions to help generate discussion. At the end of the allotted time, participants were asked to rate the usefulness of each tool on a scale of 1 to 5, with 1 being not at all useful and 5 being very useful. These findings represent a summary of the group discussion for each tool as well as the average rating for the tool. I also analyzed the focus group transcript containing only the participant responses using text analysis software to create data visualizations and keyword findings.

Tool 1: Datasheets for Datasets (Rated 4.2/5)

Practical and useful. Participants felt this was a practical and useful tool which is reflected in the relatively high rating they gave it. One participant who was very familiar with electronic component data sheets noted how useful these were.

> *“I used to work with electronic components and I used to read all these datasheets - 90% of what I did was looking at the datasheet and seeing what each piece does and how*

> *I can use it in my circuit. So, I think if we could move in that direction, that would be very beneficial."*

Many participants offered ideas for how to improve the tool by adding new categories for things like anonymization of data or taking a closer look at the sampling strategy involved in the data set. One participant suggested that there should be a common repository where datasheets could be accessed by everyone and updated by those who were making changes. This way, everyone could transparently see what changes were made over time which would also address concerns around how data was pre-processed. Another participant noted that the datasheets could benefit from being standardized, just like electronic component datasheets are standardized by IEEE.

Issues with documenting data. There was some discussion as to whether datasets were as easily documented as electronic components because data can have multiple uses or be applied to different situations. As one participant observed:

> *"there's an important distinction between the datasheets for electronic parts and for datasets…electronic components are designed for one purpose and you can not use them for anything else…and that is not the case in the data sets. The electrical components, they don't have more capabilities than what they currently have…The data set is a different product….it is not something fixed, it is not something that it is even finished…it is just raw material."*

The discussion also delved into the idea that "the data is just out there" and could be used by anyone to do whatever they wanted, which another participant acknowledged was "a bit scary". There was also a sense that the tool would be most useful if those collecting the data completed it and those using the data could review the information. This would save time instead of "reading long papers" and other ways that datasets are currently documented.

Shared exchange of ideas. One interesting moment that occurred was when a participant expressed confusion as to why a question about funding might matter on the datasheet and another participant raised the idea of how funding might influence or create bias in a data set. That led to the first participant commenting later as to how they had learned something new:

> *"Who funded the data is also a very interesting thing that I just got introduced to, which I think is more important for the people who are doing research in the medical field directly, or pharmaceutical research."*

It was an "aha" moment for me as a researcher witnessing this shared exchange of ideas around ethics that led to an evolution in knowledge and thinking about these issues. This is exactly the type of dialogue these tools are meant to surface.

Tool 2: Principles for Accountable Algorithms and Social Impact Statement (Rated 3.2/5)

This tool was rated lower than the other tools in terms of usefulness which may be partially attributable to how the group discussion evolved. The tool is meant to address the issue of transparency through writing a social impact statement for an algorithm that is informed by using five principles. However, the group discussion became fixated on one principle – explainability. Explainability is a hotly debated topic in machine learning and came up as a controversial issue in some of the one on one interviews. The discussion was very animated, almost a debate at times and the polarization in the group was reflected in the rating of the tool. I made one gentle attempt mid-discussion to try and recentre the conversation back to the whole tool but that did not get traction. Rather than being more heavy-handed, I decided it was important to allow the group time to explore this highly charged topic.

Problems with explainability. The conversation started with one participant passionately sharing their views about explainability and why they felt this tool was "completely useless" and could "do more harm than good":

> *"I really "hate on" this field of explainable AI … I understand why people want it and I understand why this is important. Now in my view, everything that is going to be written about this is going to be made up stories…. is not going to be something real….and the reality is these are not quantifiable concepts."*
>
> and
>
> *"most of the explanations is going to be to go towards the non-technical people, right? So, if you tell them I have an algorithm that can do this, and then you explain it, you know that that is false, because chances are that is not exactly what the algorithm is doing"*
>
> and
>
> *"when you go to the technical part, you will see that many of these concepts [like explainability] will not pass any formal or technical evaluation or scrutiny….other than accuracy - accuracy, you can actually measure. And other than that, I don't see how any of these concepts do something useful for the technical research in machine learning."*

Some agreed that "we don't yet have very strong mechanisms to explain black box models" and that there was a vagueness or lack of a concrete definition for explainability and some of the other principles in the tool, such as fairness and responsibility. However, they also felt that despite these limitations, it "doesn't mean that we will never have it [explainability] or we should not do anything" to attempt to address explanations. In fact, "the user would be very interested to know why a machine learning model reached that decision" and that explanations are "largely dependent on the audience" and need to be made domain specific. This user perspective was echoed by another participant who said that even though there are things which can't be measured yet, they are still "very real and important to people."

One participant aimed to seek some middle ground and build consensus by acknowledging the divergent viewpoints of others while also expressing concerns that a lack of explainability can be used to shut down research, which is not a good thing either:

> *"As Participant A mentioned, in order to truly be able to explain what our algorithm needs to find the underlying causal mechanism…we really need to develop causal models. And that's the next step in AI and there are people who are working in this field. So, I think explainability is very important, but we are not able to do it yet. I think we shouldn't shut down the idea altogether, because we can't do it yet, as Participant B mentioned, we should try to achieve that high goal, as opposed to just shut it down."*
>
> and
>
> *"some people in healthcare and industry, they just shut down an idea because it's not explainable, and that's not good as well, because if they don't fund us, we won't be able to develop more explainable algorithms. So, I think, explainability is very important, but…it's not as definite as we want. For example, accuracy is pretty well defined. We know exactly what it means but explainablity is not like that…It really depends on the granularity of the explanation and the explainee."*

Another participant added:

> *"We can't completely explain to people, things like deep learning that we're not a hundred percent sure about, we can tell them what we know - why we believe that these things work. We can tell them our current explanation and if somehow we can come up with a better answer then we can update that."*

Problems with fairness. Moving on from explainability, another participant said they felt the principle of fairness was also concerning saying “I'm not a hundred percent sure how much fairness comes into play” and that the concept of fairness in this case was “vague”. They also questioned whether or not an overarching definition of fairness made sense or whether fairness should be more nuanced and contextual.

This sparked some discussion about how the tool might be more useful at provoking thought and debating these hard to define issues, rather than being a definitive authority on these principles. As one participant noted, it’s “wise to think about what we are doing and why are we doing it.” The discussion could easily have gone on longer, but I had to bring it to a conclusion due to time constraints.

In general, the primary issues with the tool seemed to stem from the vagueness of certain principles which were being used try to draw a conclusion or make a claim in the form of an algorithmic social impact statement. This proved to be very problematic for some of the participants and resulted in a lower rating for the tool overall but made for a very interesting and engaging discussion. In that regard, the tool may have proved to be more useful than acknowledged.

Tool 3: AI Blindspot (Rated 4.4/5)

The last of the three tools, AI Blindspot, was highly regarded by everyone. It is possible there may have been some effect in seeing the pendulum swing towards consensus for this tool after the more polarizing discussion around tool two. There was a notable change in the mood of the group as this tool was discussed.

A useful discussion and reflection tool. Comments were very favorable for this tool:

> *"It's obvious that I love it. I like it a lot. I like it a lot because it's a discussion and reflection tool. It's not judging. It's just a reflection tool to talk about and think about and I think that's important."*
>
> *"I also like this one and that the main reason is because this is a thinking process and it involves people from many different areas."*
>
> *"I also like the tool very much. I think it's a very good educational tool, first of all, which any researcher who is working on AI should take a look at these cards."*
>
> *"These will be things that you always kind of have in mind, and they become, something like a requirement for you to kind of be careful. And especially for people from the technical side, they think this is great."*

The idea of the cards being more about reflection and less about pointing in a concrete direction was a big part of the appeal. Some participants noted how the cards could be helpful to foster conversation not only with technical team members but also non-technical stakeholders, such as clinicians.

Small improvements. Two participants pointed out ways that the cards might be improved. One said that "there could be some room for more flexibility" in terms of expanding on some of the points on the back of the cards in order to customize topics to a particular project or domain. Another participant said it would be useful to have a companion guide to use with the cards in order to document conversations, somewhat like a reflective journal. This kind of documentation would help ensure past conversations would be not be lost and the team could build on prior conversations.

AI Blindspot seemed to be a tool that everyone saw value in using. Participants liked how it covered a range of issues related to various aspects of the AI workflow process. They also appreciated that it did not try to provide a concrete direction but was designed to stimulate thinking.

Focus group data visualization

Using Voyant Tools, I entered the transcripts from the focus group for each ethics tool containing only the participants discussion and no researcher comments to produce the following visualizations:

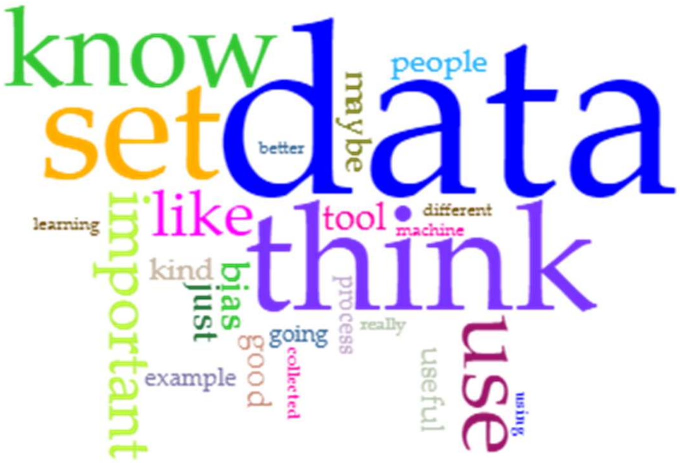


Figure D - Tool One: Datasheets for Datasets

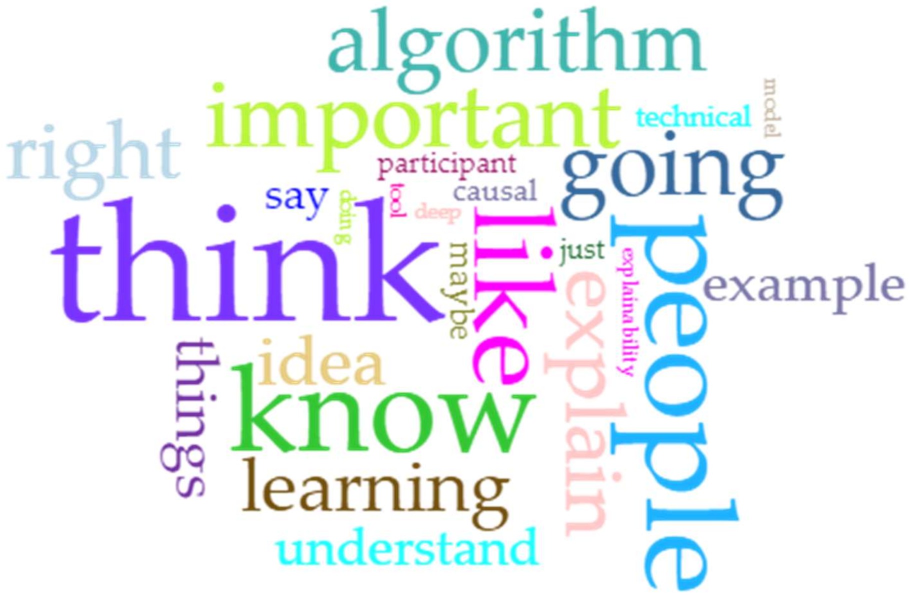


Figure E - Tool Two: Principles for Accountable Algorithms and Social Impact Statement

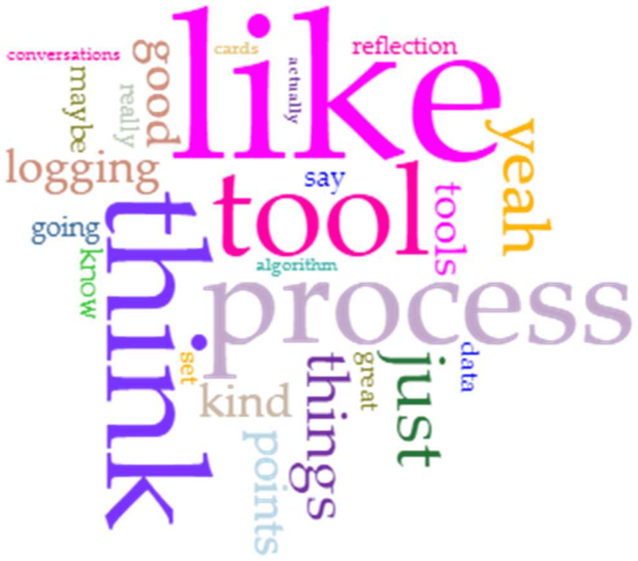


Figure F - Tool Three: AI Blindspot

Top Terms Overall: Think, Data, Like, Know, Dataset (See Appendix E)

Tool	Text Length	Distinctive Words
One	2665 (35%)	Bias, Pre-Processing, Purpose, Collected
Two	3679 (50%)	Causal, Models, Deep, Fairness
Three	1061 (15%)	Process, Logging, Reflection

Figure G

The data visualization terms align with key themes discussed in the focus group with data playing a dominant role in tool one (Datasheets for Datasets) and everyone "liking" tool three (AI Blindspot). Appendix E provides additional insights on the frequency of terms used as well as unique words in each transcript. It was interesting to note that approximately 50% of the text was generated discussing tool two (Principles for Accountable Algorithms and Social Impact Statement) which is consistent with spending more time on this tool as the group debated the concept of explainability. Explainability is also connected to some of the distinctive words for tool two noted in Figure G while Figure H shows the semantic web of explain and explainability connecting with other terms.

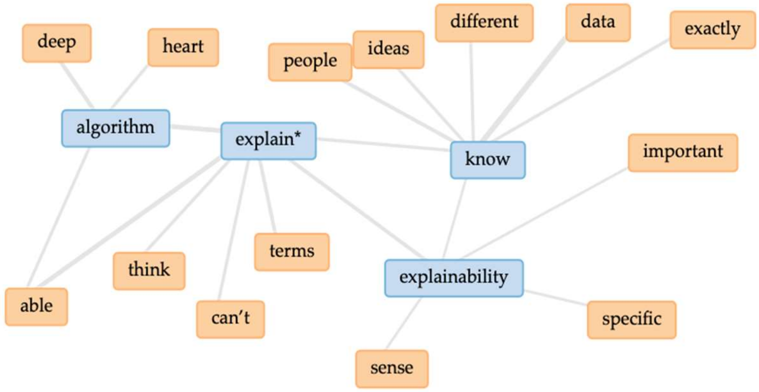


Figure H

The focus group was useful to see how AI researchers navigated various issues and ethical concepts and to note the areas where there was alignment as well as those where differing perspectives gave rise to debate. In general, there was consensus that ethics tools could be useful in the context of the AI workflow. There was also a shared sense from participants that discussing ethics was useful and there could be further utility in having ongoing discussions about ethics in the context of their work. The specifics of which tools may work best for which researchers will be framed by personal choice, project fit and context.

Discussion

To the best of my knowledge, there have been few if any studies which seek to understand how AI researchers working in the healthcare domain can apply ethics to inform their work. My research is informed by a social construction of technology (SCOT) lens which attempts to understand how AI researchers (the social group) are shaping the direction of AI technologies in healthcare from an ethics perspective. My findings also illustrate how participants are situated within a community (socio-cultural and economic context) which impacts how they think about ethics and how ethics is applied in their workflow. SCOT is an iterative process that informs how humans shape technology and is useful to highlight how ethical issues intersect within technological processes (Bijker, 2015).

Participants acknowledge that ethics is important. However, most participants were limited in their thinking about ethics and often do not see how ethical concerns connect to their day to day work. To the degree they think about ethics within their work, it’s primarily for pragmatic reasons, such as clearance for obtaining data, ensuring a statistically valid use of data and making sure their work is technically accurate. Those are good things; however, they represent a fairly narrow approach to applied ethics. This limited approach misses bigger picture opportunities, which were identified in the literature, to include more fulsome ethical discussions around issues throughout the entire workflow. Here are some of the bigger picture issues that could be addressed:

- determining who is responsible for an AI technology and how it will be governed
- discussing how to address systemic bias in a dataset
- designing affordances in the technology
- deciding how to account for fairness and transparency in technology design choices
- exploring ways to build more inclusive technology development processes
- discussing when not to use AI or particular AI techniques to solve a problem.

My study identifies numerous points where discussing ethical issues might serve to enhance decision making within the workflow of AI development. This played out in my research during the focus group. As part of the discussion, participants engaged in an extended dialogue around the issue of explainability and were afforded the opportunity to take a bigger perspective to broaden their ethical thinking. There are many tools that can pro-actively foster this type of engagement for AI researchers. These tools can be tailored to be domain specific. They can also be used to enable technical (computer science) team members to connect more effectively with non-computer science team members to encourage a more inclusive and holistic approach towards ethical questions. If we can broaden the ethical thinking of AI researchers, then we have an opportunity to drive better outcomes for AI technology.

A narrow perspective on ethics may be a result of and reinforced by socio-cultural factors within the research community. There are two factors that were raised in the findings which seem relevant to discuss: limitations in research ethics oversight as it relates to big data and a lack of education in ethics for computer science students. These areas are actively being explored, rethought and renegotiated by some members of the research community.

The exclusion of public datasets from research ethics board oversight seems to create a loophole for AI researchers whose core focus is the reuse of large, secondary data sets and it sends a message to these researchers that their work doesn’t require the same level of ethical oversight as other research. Yet, the literature suggests that in the era of big data, researchers need to be more accountable and take responsibility for their work (Berendt, Buechler & Rockwell, 2015). There are also researchers calling for a “reconsideration of fundamental ethical assumptions” by both individual researchers and research ethics boards in light of “pervasive data practices” resulting from big data (PERVADE, n.d.). Changes in the ethics review process for projects involving big data could have a ripple effect for individual AI researchers in practical terms, but more importantly it serves to highlight the relative importance of ethics in AI.

The lack of basic ethical training for computer science students, especially at the PhD level, was a surprising finding. The computer science majors who participated in the study all completed undergrad and graduate work within engineering or computer science which is not unexpected. The only participant in the study with formal ethics training was the non-computer science major. The approach this participant had towards ethical issues and how ethics informed their work was noticeably different. Computer science is a highly competitive discipline and as one participant noted, there is little time for anything outside of their program. One of the unintended consequences of a STEM focused approach might be the loss of a more broad-based outlook which humanities can provide. If ethics is deemed important, it needs to be woven into the fabric of the program itself, ideally well before a student reaches a PhD level.

A recent global survey found that only 18% of students receive ethics training (Anaconda, 2020). This same study found that “social impacts from bias in data and models” along with “impacts to individual privacy” were the top two issues in AI/ML that need to be addressed (Anaconda, 2020, p. 30), echoing concerns that were raised in the literature review. One way to start addressing these issues is to include ethics training within computer science curricula. The Mozilla Foundation has identified “teaching students how to ask and explore ethical questions” as an important element in changing industry norms and building trustworthy AI (Ricks and Surman, 2020, p. 22). Another team of researchers has documented ethics curricula in US computer science programs and makes practical suggestions for teaching ethics in AI in ways that are substantive, timely and relevant (Garrett, Beard & Fiesler, 2020). This lack of education and training is a gap that contributes towards a limited understanding and application of ethics by individual AI researchers.

Shifting paradigms at an AI research community level, as it relates to education and review boards, would reshape ethical thinking at the level of the individual AI researcher and consequently, impact technology development. From a social constructivist perspective, Bijker describes this process as the “technological frame” by which “the actions and interactions of actors…build up around an artifact and thus incorporate the characteristics of that technology (Bijker, 2015, p. 138). The frame becomes “the way we do things” and it causes those in a community to see things a certain way. Bijker notes the example of how Dutch family houses designed for two-parent, two-child families “dominated architectural thinking in the 1950s through the 1970s, and thus made it very difficult to conceive alternatives” (Bijker, 2015, p. 138). There are signs that the AI research community might be ready to “conceive alternatives” and reframe their approach to ethics in AI which could break the current “technological frame”, thereby changing the characteristics of the technology that is developed.

Ethics tools have a role to play in creating an alternative frame whereby ethical considerations are seen as necessary inputs towards building trustworthy AI. Trustworthy AI is AI that is informed by ethical principles (Wing, 2020). Ethics tools can be used to broaden ethical thinking and address the many bigger picture questions raised in the literature. These issues are not currently being addressed through the narrow lens of technical fixes. The focus group findings provide one small example of how this might work in practice. However, in order for tools to be useful, it’s necessary to learn how to use them. This is where the role of education might intersect with ethics tools.

Ethics tools could be used within an educational context in order to broaden ethical thinking and support the habit of building ethical considerations into AI development. The tools can help educators incorporate ethics training into a computer science curriculum by weaving it into more technical course work. By using ethics tools within a post-secondary context, future AI researchers are gaining the practice and habit of thinking about ethical considerations as it relates to their work. Additionally, weaving ethics into a discipline-specific context can help "convince students that ethics is a part of their profession and not just a public relations add-on" (Garrett et al, 2019, p. 2). The use of ethics tools might serve to reinforce the idea that AI researchers need to be responsible for the choices they make as professionals when designing AI systems.

Limitations. My study has several limitations that should be noted. As mentioned in methods, there are limitations around generalizability which come from conducting a qualitative study with a small, non-probability sample. While my study surfaced many interesting findings, as a single researcher with a constrained timeline and limited resources there are areas that I was not able to fully explore, including comments made in interviews that while interesting, felt tangential to the core study. Conversely, there were topics not raised in the course of the interviews that I believe are important to consider but were deprioritized based on time constraints. A list of these areas is noted in the conclusion section under future work.

Aside from my own biases, there is selection bias in my study. Participants who self-selected to take part in the study are likely more aware of ethical issues than their peers. Electing to take part in a study about ethics signals at minimum an interest in the topic. Participants are all working in healthcare, a high stakes domain with a rich tradition of ethics and better regulatory oversight of data and research practices than many other less data sensitive domains. The interviews and focus group discussion confirm that participants share an appreciation and concern for ethics. Yet, despite this selection bias, there are still ways that participants can improve their ethical practices within their work processes. It’s not too much of a stretch to think that other AI researchers would see similar, if not even more benefit, from taking an applied ethics approach in their work as well.

Validity and Reliability. Despite noted limitations, I believe my study makes a worthwhile contribution towards the question of how to apply ethical tools within an AI workflow. I ensured that the data collection instrument was sound by seeking input from senior researchers, applied the standard of “member checks” (Merrigan et al, 2012) to verify findings and stayed close to the data while following best practices for coding (Mayan, 2009). I am also taking care not to overstate my results.

Summary of Findings and Discussion

My study set out to explore how AI researchers working in the healthcare domain can apply ethics to inform their workflow. My findings show how some AI researchers think about ethics, where ethical considerations intersect with their workflow and how using tools to explore ethical considerations and support ethical processes throughout the AI workflow can be helpful. Certain tools in particular can contribute to broadening perspectives around ethics which can be useful in guiding better decision-making processes. The hope is that using these ethics tools can drive better outcomes for AI technology, though further work would need to be done to substantiate that claim.

Technologists bring a certain paradigm or perspective to their work. This was highlighted by the inclusion of one non-computer science major who participated in the study. This participant was an outlier in many ways, from their formal ethics training and knowledge of bioethics, to their design choices which gave preference to explainability over the use of deep learning. By having this non-computer science AI researcher in the study, I was able to see more clearly the different ways in which AI researchers, most of whom are from computer science, think about ethics and how it applies in the context of their work.

Lastly, the need for better ethics education and training within computer science programs seems like a practical next step that universities are well positioned to address. This could have tremendous impact in how ethics is applied in the AI workflow and ethics tools can play a role in ethics education. In addition, a renewed perspective towards ethical oversight and the role of research ethics boards in an era of big data also seems like it could go a long way towards helping address how AI researchers think about and apply ethics.

Conclusion

Ethics tools can play a role in helping AI researchers address ethical issues in the context of their work by offering a means to broaden ethical thinking in ways that go beyond technical fixes. There are many places where AI systems intersect with ethical issues involving the data used to construct AI systems and the choices made in constructing the mathematical models used in these systems. Ultimately, AI is made by people and the AI research community is a fairly elite and homogenous group, funded by a handful of corporate and government interests. Many AI systems currently being used in real-world applications have caused people harm. AI systems developed for healthcare applications are especially concerning because they operate in a high-stakes environment where they may inform decisions having significant impacts on humans, including life and death decisions.

My study involved interviews with AI researchers working on healthcare projects to gain a better understanding of how they think about ethics and where they see ethics intersecting with their work. While there was agreement about the importance of ethics, there was a gap in how this translated into actual practices. Most ethical considerations were narrow and technical in nature, missing many of the bigger picture issues identified in the literature. These interview findings were then used to select three ethics tools that were explored in a focus group. The objective of the focus group was to discuss how these tools might be useful in opening up the range of ethical issues considered within the work of developing AI systems. The focus group findings suggested that ethics tools could be useful to help guide ethical thinking and broaden perspectives. While this study was limited in scope, making the results not fully generalizable, it still provides some useful insights. Two noteworthy issues that were raised in this study are the potential limitations in the role of research ethics boards in an era of big data and a gap in ethics education within computer science curriculum.

There are many opportunities to build on this study and to explore issues that due to time and resource constraints were not able to be fully investigated. Here are some of the comments and areas for future work:

- How will the use of AI in medicine impact or shape medical practices? This question was raised by one participant but was not explored beyond this minor reference.
- There is a gap between the belief that funding in academia has no influence over research and the literature that suggests how funding can set agendas at the institutional or government level. Why is this connection being missed and what impact does that have for AI research in healthcare?
- There could be further exploration around who is responsible for AI and the role of third parties to govern or oversee ethics vs the role of individual researchers to assess ethics in a day to day context.
- There was little discussion about what might happen if a technology works as intended and how that might lead to ethical concerns. While AI researchers seemed to see accuracy as an ethical duty, accuracy of the technology does not necessarily guarantee ethical outcomes.
- There is a tension and ambiguity that needs to be more fully explored which relates to an appreciation for the non-binary thinking that goes into ethical considerations.

Any of these issues would make for interesting follow up research. However, I believe an opportunistic next step would be to explore how university curricula might be enhanced in order to include ethics training for computer science students. I think a quantitative study focusing on Canadian universities to explore what training is currently delivered and where gaps exist would be an interesting next step. The ethics toolkit that I will be developing as a companion piece to this paper can also play a role in supporting education for AI researchers.

There are signs that the AI community is starting to understand the importance of applying ethics in their work. Important conferences, such as NeurIPS, are now requiring AI researchers to include a social impact statement with their papers (Johnson, 2020). Ruha Benjamin, a sociologist who specializes in the intersection of race and technology, was a keynote speaker at the 2020 International Conference on Learning Representations (ICLR), a deep learning conference. These developments seem particularly relevant in pointing towards addressing the more radical "second wave" critique of AI referenced in the literature. However, it is still early days and as the AI community navigates ethics there are varying perspectives about the scope of the issues and how best to approach them. One recent heated example of AI community infighting centred around racial bias in AI and whether it was just a question of biased data sets or a bigger structural problem (Synced, 2020). The battle took place on Twitter between two well-known AI researchers, Timnit Gebru and Yann LeCun and ended with LeCun exiting Twitter permanently. Gebru is part of a new generation of AI researchers who have been critical of limited perspectives within the field. She is an advocate for the use of ethics tools and was the lead designer of *Datasheets for Datasets,* the first tool discussed in the focus group. I expect we will see more debate within the community as well as calls to include expertise from other domains and even the general public, as we seek ways to implement ethics in order to build trustworthy AI. I believe ethics tools can be used to support AI researchers to develop ethically aligned AI.

Appendices

Appendix A: Sample information letter (excerpt)

INFORMATION LETTER and CONSENT FORM

Study Title: Ethically Aligned AI: Applying ethics in AI for healthcare

Research Investigator:	**Supervisor:**
Katrina Ingram	Dr. Geoffrey Rockwell
University of Alberta	University of Alberta

Background

I’m a masters student working on my final capstone project. My research study is on applied ethics within the development of artificial intelligence systems for healthcare. The purpose of the study is to document ethical practices currently being followed in the design of AI systems for healthcare applications, identify gaps, develop ethics tools, and to make recommendations for ethical guidelines.

I am recruiting participants from the following groups:

- AI researchers (PhD students and professors in computer science) who are researching and/or developing AI related to healthcare.

I’m recruiting 6-8 participants for this study through word of mouth connections in my network. I am asking for your participation in this study

Purpose

The purpose of the study is to document ethical practices currently being followed in the design of AI systems for healthcare applications, identify gaps, develop ethics tools, and to make recommendations for ethical guidelines.

Appendix B: Sample interview questions (one on one interviews)

General

What would you say are the major areas of ethical concern for AI researchers in general?

Who, in your opinion, should be responsible for ensuring AI meets ethical standards? Please explain your choice.

Data

Let's discuss your data sets. I'm curious to know what data sets you are using and how your data sets were gathered – can you tell me about that process.

- If you are using secondary data sets, where did you obtain this data and what was the methodology used to gather it?
- Do you know how your training datasets were labelled and who labelled them? How have you defined "ground truth" in your project?
- Are you using social media data (i.e. Twitter, FB, Instagram)? What steps, if any, have you taken to obtain consent? What steps, if any, have you taken to protect privacy?
- Are you using health records or other information in your data set that are protected under provincial or federal privacy laws? How are you ensuring that you abide by these legal requirements in managing and analyzing your data?

Models

How do you know if your AI model is actually doing what you hoped it would? How do you ensure your model works?

How do you ensure your model does not produce harmful or unintended consequences? For whom does your AI model fail? For example, what if is your AI algorithm is pretty accurate on population X but not on population Y? Have you tested for its accuracy on different populations such as different mixes of gender, race, age, socio-economic background etc.?

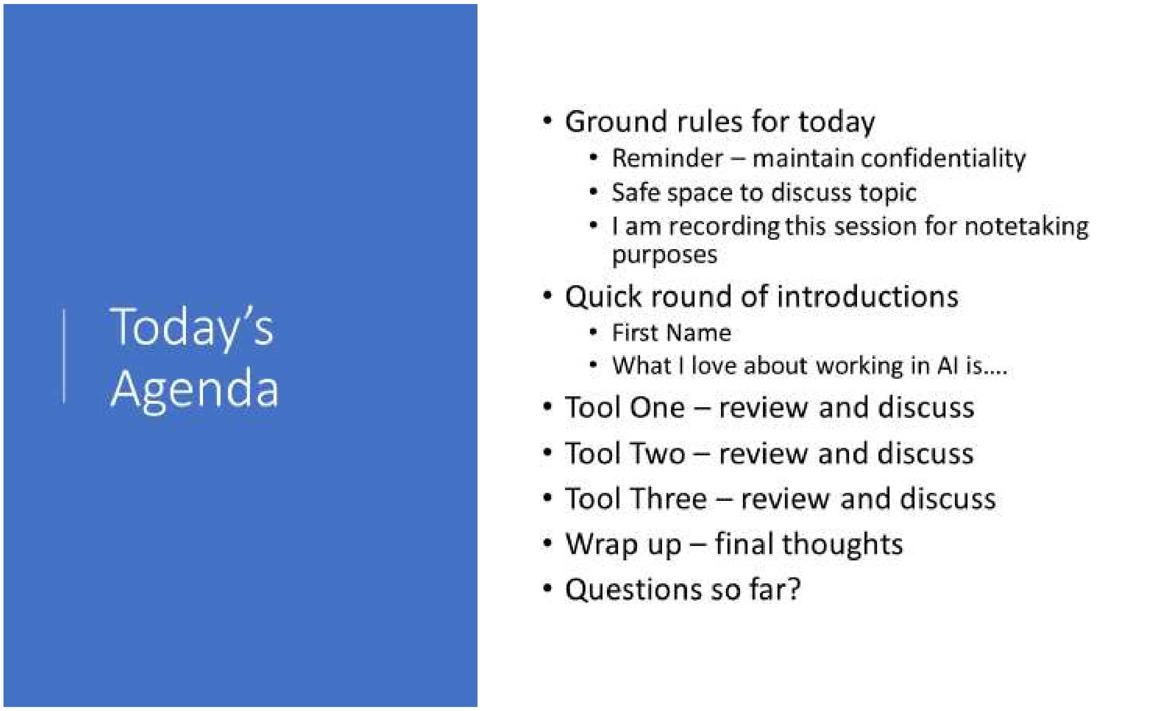
To what degree is your AI model explainable? If you are using deep learning or other techniques that are not explainable, how do you weigh the trade-offs between the use of this technique vs explainability of outcomes?

Ethical Codes and Tools

Several ethical codes for AI development have been published in the last two years by government, industry, not for profits, industry associations and academia. Which codes are you aware of and do you personally adhere to any particular ethical code(s) or principles in the course of your work?

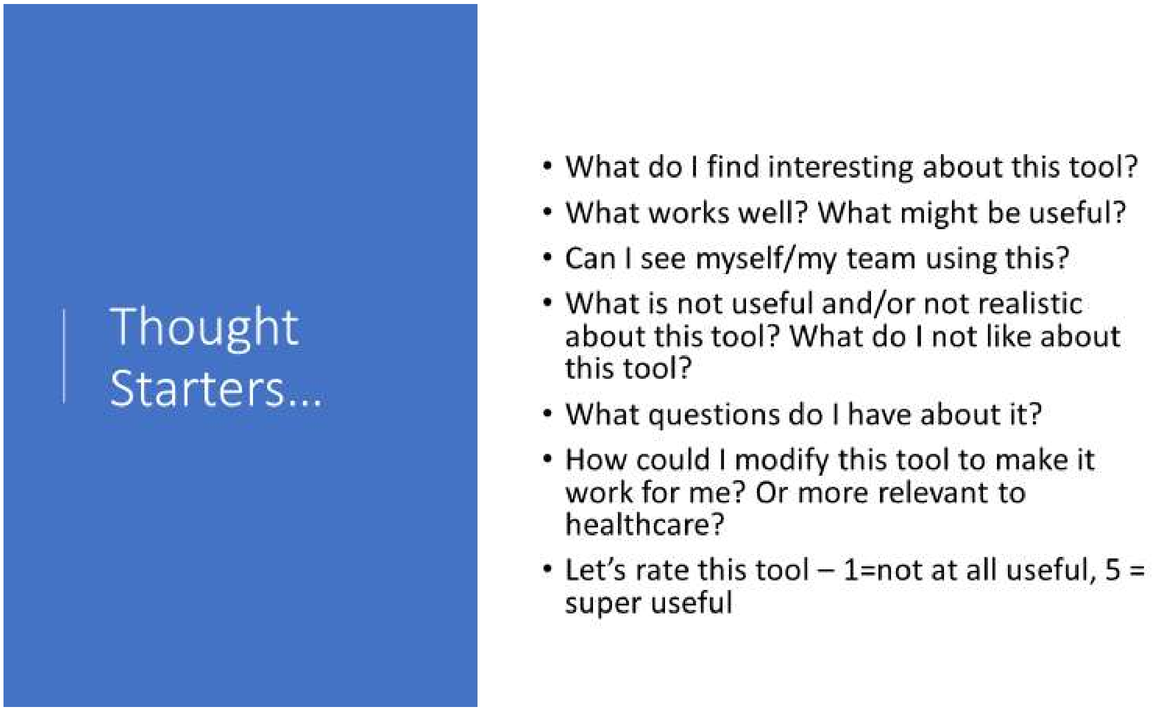
Given your focus on healthcare, how has bioethics informed your research project if at all?

Appendix C: Sample focus group slides and materials



Today's Agenda

- Ground rules for today
  - Reminder – maintain confidentiality
  - Safe space to discuss topic
  - I am recording this session for notetaking purposes
- Quick round of introductions
  - First Name
  - What I love about working in AI is....
- Tool One – review and discuss
- Tool Two – review and discuss
- Tool Three – review and discuss
- Wrap up – final thoughts
- Questions so far?



Thought Starters...

- What do I find interesting about this tool?
- What works well? What might be useful?
- Can I see myself/my team using this?
- What is not useful and/or not realistic about this tool? What do I not like about this tool?
- What questions do I have about it?
- How could I modify this tool to make it work for me? Or more relevant to healthcare?
- Let's rate this tool – 1=not at all useful, 5 = super useful

Tool One: Datasheets for Datasets

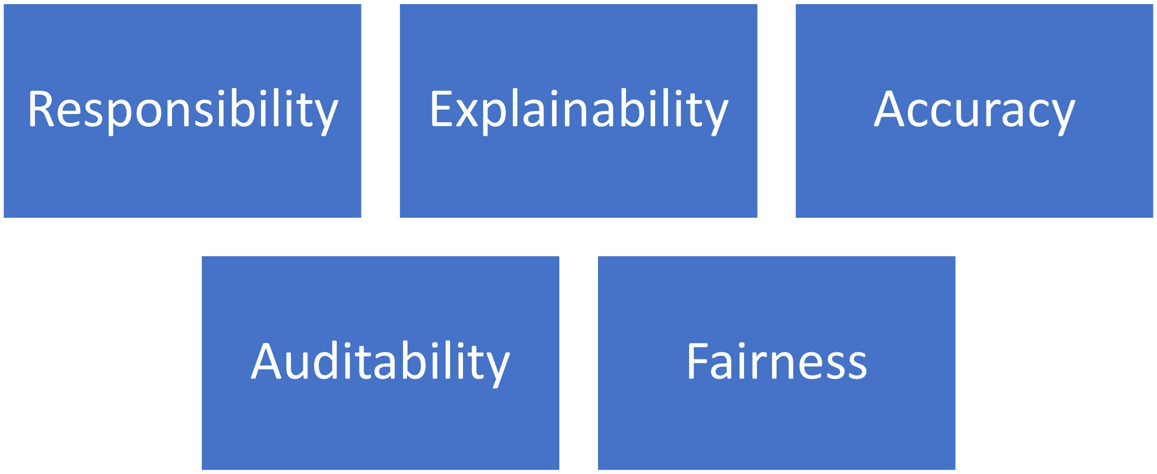
Datasheets for Datasets was written as a paper. I have taken the concepts from the paper and reformatted it into a tool with some of the suggested questions. https://arxiv.org/abs/1803.09010

Area	Questions	Answer
Motivation	For what purpose was the dataset created?	
	Who funded the creation of the dataset?	
Composition	What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances?	
	Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	
	Are there any errors, sources of noise, or redundancies in the dataset?	
Collection Process	If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?	
	Were any ethical review processes conducted (e.g., by an institutional review board)?	
	Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?	
Pre-processing	Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?	
Uses	Is there a repository that links to any or all papers or systems that use the dataset?	
	What (other) tasks could the dataset be used for? Are there tasks for which the dataset should not be used?	
Distribution	Will the dataset be distributed to 3rd parties outside of the entity who created it?	
	Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?	
Maintenance	Will the dataset be updated?	
	If the data relates to people is there a limit on retention? **Are there healthcare regulations that guide this limit?**	

Tool Two: Principles for Accountable Algorithms and Social Impact Statement

I have taken the concepts from the FATML website and reformatted them for discussion purposes. For more info - https://www.fatml.org/resources/principles-for-accountable-algorithms

The FIVE Principles



Responsibility

Explainability

Accuracy

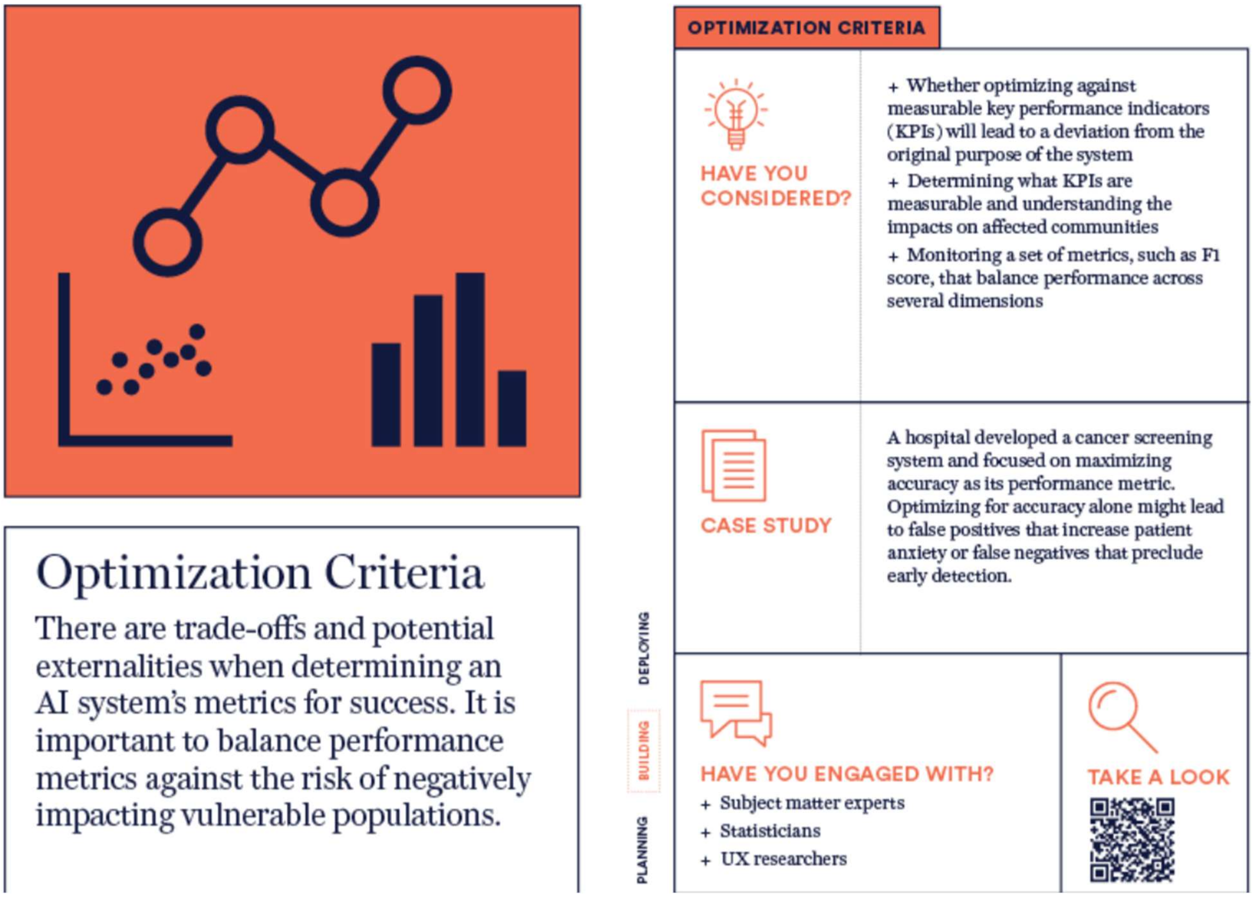
Auditability

Fairness

Sample questions:

- Who is responsible if users are harmed by this product?
- Who will have the power to decide on necessary changes to the algorithmic system during design stage, pre-launch, and post-launch?
- Who are your end-users and stakeholders?
- How much of your system / algorithm can you explain to your users and stakeholders?
- Can you provide for public auditing (i.e. probing, understanding, reviewing of system behavior) or is there sensitive information that would necessitate auditing by a designated 3rd party?
- What sources of error do you have and how will you mitigate their effect?
- How confident are the decisions output by your algorithmic system?
- What are realistic worst-case scenarios in terms of how errors might impact society, individuals, and stakeholders?

Tool Three: AI Blindspot



Optimization Criteria

There are trade-offs and potential externalities when determining an AI system's metrics for success. It is important to balance performance metrics against the risk of negatively impacting vulnerable populations.

OPTIMIZATION CRITERIA

HAVE YOU CONSIDERED?

+ Whether optimizing against measurable key performance indicators (KPIs) will lead to a deviation from the original purpose of the system
+ Determining what KPIs are measurable and understanding the impacts on affected communities
+ Monitoring a set of metrics, such as F1 score, that balance performance across several dimensions

CASE STUDY

A hospital developed a cancer screening system and focused on maximizing accuracy as its performance metric. Optimizing for accuracy alone might lead to false positives that increase patient anxiety or false negatives that preclude early detection.

HAVE YOU ENGAGED WITH?

+ Subject matter experts
+ Statisticians
+ UX researchers

TAKE A LOOK

.**Card contains:**

- Questions for consideration related to the card's topic
- A case study example
- A list of people who might prove useful to engage with on this topic.
- Links/QR code to additional resources (papers etc.)

How to use the cards:

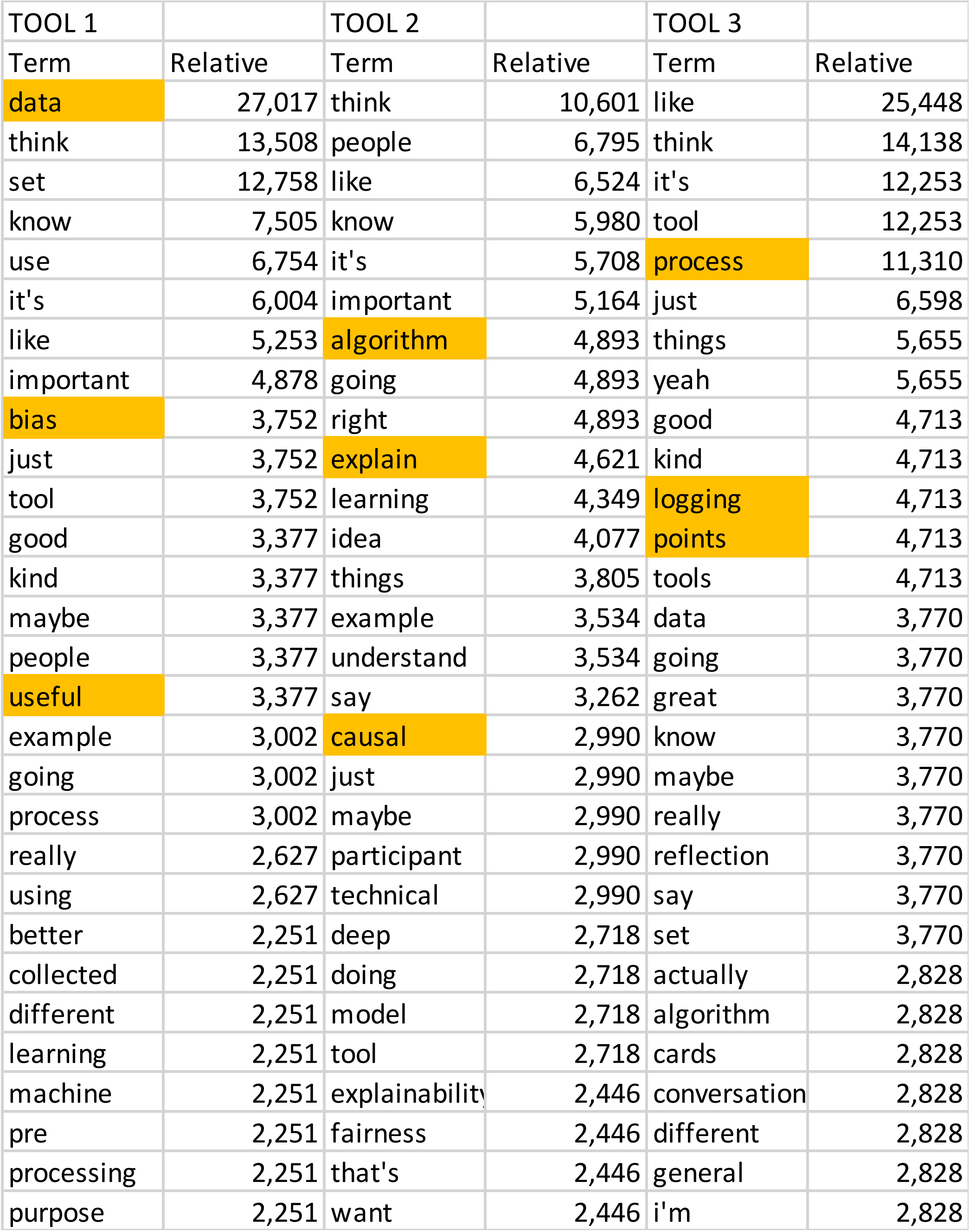
- The cards can be used beginning to end during the AI design cycle or individually at a particular part of the workflow.
- Since the cards are easy to understand, they have also been used with non-technical audiences (policy makers, stakeholders etc.) to help better explain, educate and drive conversation around the AI workflow.

Retrieved from - https://aiblindspot.media.mit.edu/

Appendix D: Summary of Key Findings and Demographic Data

Category	Variable	Summary of Key Findings
Demographics	Gender	Two of five participants were female (40%) In general only 16-20% of AI researchers are female.
	Major	All participants except one were computer science (CS) majors.
Research Perspective	Description of research	The CS participants were focused on problem solving and used words such as “finding, analyzing, coming up with, validating, and predicting” to describe their work. The non-CS participant used “help, advise and learn from” to describe their research.
Ethics in General	Ethics Board Review	Most participants did not obtain ethics board approval. If this was done, it was indirectly by another team member or professor. Most used secondary data/public data - not covered by ethics review boards.
	Research Motivation	Most were influenced by their supervisor's research interests and shared a sense of wanting to "do good" and to use AI to solve a problem.
	Who is responsible for ethics in AI	Everyone agreed that there is a system with many players and that everyone has a role, from individual developers to industry to the end user to ethics boards.
	Barriers to applying ethics	Issues vary but in general they involve non-tech stakeholders, whether that is the fear of misuse or the fear of the system delivering bad outcomes or non-tech people in the process who don't understand (i.e. on ethics board or in charge of data). One person mentioned time pressures and not having enough time to really learn more about ethics.
	Stakeholder engagement	Some had indirect stakeholder engagement while others were more involved directly with clinicians. Only one person had no stakeholder engagement.
Data	Data Used	All researchers are using secondary data sets and thus face the issues inherent in appropriating data for another use. One participant was generating synthetic data based on real world data.
	Data Security	Everyone is using secure storage and following university or company protocols.
	Type of Data	There were a range of data types: images, text, social media, questionnaires, synthetic.
Models	Use of Deep Learning	Deep Learning (DL) is a powerful technique, but it creates problems for explainability. Everyone except for one participant (the non-CS major) is using DL as their primary technique as it’s more accurate and able to deal with bigger datasets.
	Explainability	The CS majors reported that explainability is either not important, focused on the wrong question, that the math is explainable (even if outcomes are not) or that while it’s important it’s not as important as accuracy. The non-CS major said explainability was very important
Ethical codes, tools and training	Ethics training and understanding of bio ethics	Only one participant (non-CS major) received formal ethics training. All agreed that training would be valuable. Only one participant understood concepts of bioethics (non-CS major).
	Ethical Codes	Most have little to no knowledge of any specific ethical codes for AI. One person mentioned the GDPR. Most think of ethical codes in terms of technical AI capabilities to protect or anonymize data. Some mentioned having personal principles that they apply.
	Ethical Tools	No one is using ethical tools or had any knowledge of any specific ethical tools. One person (non-CS major) mentioned a data management plan. Most participants have informal discussions with other team members, such as clinicians, where the topic of ethics is touched on.

Appendix E: Ethics Tools: Terms Frequency – Inverse Document Frequency



TOOL 1		TOOL 2		TOOL 3	
Term	Relative	Term	Relative	Term	Relative
data	27,017	think	10,601	like	25,448
think	13,508	people	6,795	think	14,138
set	12,758	like	6,524	it's	12,253
know	7,505	know	5,980	tool	12,253
use	6,754	it's	5,708	process	11,310
it's	6,004	important	5,164	just	6,598
like	5,253	algorithm	4,893	things	5,655
important	4,878	going	4,893	yeah	5,655
bias	3,752	right	4,893	good	4,713
just	3,752	explain	4,621	kind	4,713
tool	3,752	learning	4,349	logging	4,713
good	3,377	idea	4,077	points	4,713
kind	3,377	things	3,805	tools	4,713
maybe	3,377	example	3,534	data	3,770
people	3,377	understand	3,534	going	3,770
useful	3,377	say	3,262	great	3,770
example	3,002	causal	2,990	know	3,770
going	3,002	just	2,990	maybe	3,770
process	3,002	maybe	2,990	really	3,770
really	2,627	participant	2,990	reflection	3,770
using	2,627	technical	2,990	say	3,770
better	2,251	deep	2,718	set	3,770
collected	2,251	doing	2,718	actually	2,828
different	2,251	model	2,718	algorithm	2,828
learning	2,251	tool	2,718	cards	2,828
machine	2,251	explainabilit	2,446	conversation	2,828
pre	2,251	fairness	2,446	different	2,828
processing	2,251	that's	2,446	general	2,828
purpose	2,251	want	2,446	i'm	2,828

References

Ahmed, W., Bath, P. & Demartini, G. (2017). *Chapter 4 Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges.* In Woodfield, K., (ed.) The Ethics of Online Research. Advances in Research Ethics and Integrity (2). Emerald, pp. 79-107. ISBN 978-1-78714-486-6

AI for Society. (2018, December) *Inclusion in AI Development and Deployment*. Retrieved from - https://www.ic.gc.ca/eic/site/133.nsf/vwapj/1_Discussion_Paper_-_AI_for_Society_EN.pdf/$FILE/1_Discussion_Paper_-_AI_for_Society_EN.pdf

Alshenqeeti, H. (2014) Interviewing as a data collection method: a critical review. *English Linguistics Research*. *3*(1), 39-45

Anaconda. (2020). 2020 State of Data Science Report. Retrieved from - https://know.anaconda.com/rs/387-XNW-688/images/Anaconda-SODS-Report-2020-Final.pdf

Ananny, M. (2016). Toward an Ethics of Algorithms. *Science, Technology, & Human Values, 41*(1), 93-117. doi:10.1177/0162243915606523

Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., & Lawless, M. (2019). Using Zoom Videoconferencing for Qualitative Data Collection: Perceptions and Experiences of Researchers and Participants. *International Journal of Qualitative Methods*, *18*, 1609406919874596. https://doi.org/10.1177/1609406919874596

Bajarin, T. (2020, June 18). Why it matters that IBM has abandoned its facial recognition technology. Forbes. Retrieved from - https://www.forbes.com/sites/timbajarin/2020/06/18/why-it-matters-that-ibm-has-abandoned-its-facial-recognition-technology/#f66ec3fafaf3

Bellefontaine, M. (2020, April 21). Alberta Privacy Commissioner investigating Babylon by Telus Health. CBC. Retrieved from - https://www.cbc.ca/news/canada/edmonton/alberta-privacy-commissioner-investigating-babylon-by-telus-health-1.5539900

Berendt, B., Büchler, M., & Rockwell, G. (2015). Is it Research or is it Spying? Thinking-Through Ethics in Big Data AI and Other Knowledge Sciences. *KI - Künstliche Intelligenz*, *29*(2), 223–232. https://doi.org/10.1007/s13218-015-0355-2

Bijker, W. E. (2015). Technology, Social Construction of. *International Encyclopedia of the Social & Behavioural Sciences. 24* (2), 135-140.

Boddington, P. (2017). *Towards a Code of Ethics for Artificial Intelligence*. Springer. UK

Bogen, M. (2019, May 6). All the Ways Hiring Algorithms Can Introduce Bias. *Harvard Business Review*. https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias

Boyce, C., & Neale, P. (2006). Interviews: A guide for designing and conducting in-depth interviews for evaluation input. *Pathfinder Org.* Retrieved from - http://www2.pathfinder.org/site/DocServer/m_e_tool_series_indepth_interviews.pdf

Budds, D. (2017, October 17). Exclusive: Ideo's Plan To Stage An AI Revolution. Retrieved December 9, 2019, from Fast Company website: https://www.fastcompany.com/90147010/exclusive-ideos-plan-to-stage-an-ai-revolution

Burdick, A. (2017, September 15). The A.I. "Gaydar" Study and the Real Dangers of Big Data. *The New Yorker*. Retrieved from - https://www.newyorker.com/news/daily-comment/the-ai-gaydar-study-and-the-real-dangers-of-big-data

Burke, A. (2020, May 11) Airline temperature checks raise privacy concerns, experts say. CBC. Retrieved from - https://www.cbc.ca/news/politics/air-canada-temperature-checks-covid-19-privacy-concerns-1.5562939

Burt, C. (2020, June 20). Facial recognition to play key role in travel reopening as biometrics industry weighs social responsibility. BiometricUpdate.com. Retrieved from - https://www.biometricupdate.com/202006/facial-recognition-to-play-key-role-in-travel-reopening-as-biometrics-industry-weighs-social-responsibility

[C4EJournal] (2019, October 1). Marzyeh Ghassemi, Can Machines Learn from Our Mistakes? [Video File] Retrieved from - https://c4ejournal.net/2019/10/03/marzyeh-ghassemi-can-machines-learn-from-our-mistakes-2019-c4ej-37/

Calderon, A., Taber, D., Qu, H. & Wen, J., 2019. AI Blindspot. Retrieved from - https://aiblindspot.media.mit.edu/

Cassell, J. (2001, November). *Genderizing HCI.* Retrieved from https://pdfs.semanticscholar.org/4810/c28fe3523b52bd39b1eeb3b6225ab2145fa7.pdf

Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 376(2133). Retrieved from - https://doi.org/10.1098/rsta.2018.0080

Dastin, J. (October 9, 2018*) Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. Retrieved from - https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Denscombe, M. (2010). *The Good Research Guide: For small scale social research projects.* Buckingham, UK. Open University Press.

Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, H.V., Unsworth, K., Sahuguet, A., Venkatasubramanian, S., Wilson, C., Yu, C. & Zevenbergen, B. (n.d.) Principles for accountable algorithms and a social impact statement for algorithms. FatML.org. Retrieved form - https://www.fatml.org/resources/principles-for-accountable-algorithms

Domingos, P. (2015) *The Master Algorithm: How the quest for the ultimate learning machine will remake our world*. New York, NY: Basic Books.

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. American Journal of Theoretical and Applied Statistics, 5(1), 1-4

Fabbri, A., Lai, A., Grundy, Q., Bero, L. A. (2018). The influence of industry sponsorship on the research agenda: A scoping review. *American Journal of Public Health. 108* (11), e9-e16.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, *363*(6433), 1287–1289. doi: 10.1126/science.aaw4399

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Frey, L.R. and SunWolf. (2009). "Across applied divides: Great debates of applied communication scholarship" in *Routledge Handbook of Applied Communication Research*, pp.26-54.

Garrett, N., Beard, N., & Fiesler, C. (2020). More Than "If Time Allows": The Role of Ethics in AI Education. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 272–278. https://doi.org/10.1145/3375627.3375868

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2018). Datasheets for Datasets. In *arXiv [cs.DB]*. arXiv. http://arxiv.org/abs/1803.09010

Gerein, K. (2019, November 1). Keith Gerein: AHS's $1.4-billion gambit to transform the health system begins in Edmonton. *Edmonton Journal.* Retrieved from https://edmontonjournal.com/news/politics/keith-gerein-ahss-1-4-billion-gambit-to-transform-the-health-system-begins-in-edmonton

Goldsmith, J., & Burton, E. (2017). Why Teaching Ethics to AI Practitioners Is Important. *Thirty-First AAAI Conference on Artificial Intelligence*. Retrieved from https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPaper/14271/13992

Gomes de Andrade, N., Pawson, D., Muriello, D. *et al.* Ethics and Artificial Intelligence: Suicide Prevention on Facebook. *Philos. Technol.* 31, 669–684 (2018) doi:10.1007/s13347-018-0336-0

Hagendorff, T. (2019). *The Ethics of AI Ethics An Evaluation of Guidelines*. Retrieved from https://arxiv.org/abs/1903.03425

Hao, K. (2019, February 13). Police across the US are training crime-predicting AIs on falsified data. *MIT Technology Review*. Retrieved from - https://www.technologyreview.com/2019/02/13/137444/predictive-policing-algorithms-ai-crime-dirty-data/

Harari, Y.N. (2015). The Data Religion. In *Homo Deus A Brief History of Tomorrow* (pp 428-462). Toronto, ON. Penguin Random House Canada.

Heffetz, O., & Ligett, K. (2014). Privacy and Data-Based Research. *The Journal of Economic Perspectives: A Journal of the American Economic Association*, 28(2), 75–98. https://doi.org/10.1257/jep.28.2.75

Henle, T., Matthews, G. J., & Harel, O. (2019). Data Confidentiality. In A. Levy, S. Goring, C. Gatsonis, B. Sobolev, E. van Ginneken, & R. Busse (Eds.), *Health Services Evaluation* (pp. 717–731). https://doi.org/10.1007/978-1-4939-8715-3_28

Hern, A. (2017, July 3). Royal Free breached UK data law in 1.6m patient deal with Google's DeepMind. *The Guardian*. Retrieved from http://www.theguardian.com/technology/2017/jul/03/google-deepmind-16m-patient-royal-free-deal-data-protection-act

Hsieh, H. F., Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research. 15*(9), 1277-1288.

Johnson, D. R., Scheitle, C. P., & Ecklund, E. H. (2019). Beyond the In-Person Interview? How Interview Quality Varies Across In-person, Telephone, and Skype Interviews. *Social Science Computer Review*. Retrieved from - https://doi.org/10.1177/0894439319893612

Johnson, K. (2020, February 24). *NeurIPS requires AI researchers to account for societal impact and financial conflicts of interest*. VentureBeat; VentureBeat. https://venturebeat.com/2020/02/24/neurips-requires-ai-researchers-to-account-for-societal-impact-and-financial-conflicts-of-interest/

Kelion, L (2020, April 20). Coronavirus: Why are there doubts over contact-tracing apps? BBC News. Retrieved from - https://www.bbc.com/news/technology-52353720

Kitchin, Rob. (2017). *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Sage, UK.

Knight, W. (2019, October 8). An AI pioneer wants his algorithms to understand the why. *Wired*. Retrieved from - https://www.wired.com/story/ai-pioneer-algorithms-understand-why/

Kugler, Matthew B., From Identification to Identity Theft: Public Perceptions of Biometric Privacy Harms (2018, November 24). *U.C. Irvine Law Review*. (Forthcoming). Available at SSRN: https://ssrn.com/abstract=3289850 or http://dx.doi.org/10.2139/ssrn.3289850

Lincoln, Y. S., & Guba, E. G. (2006). *Naturalistic inquiry*. Newbury Park: Sage Publ.

Levin, J. (2019). Three Ways AI Will Impact The Lending Industry. *Forbes Magazine*. https://www.forbes.com/sites/forbesrealestatecouncil/2019/10/30/three-ways-ai-will-impact-the-lending-industry/

Lohr, S. (2015) *Data-ism: The Revolution Transforming Decision Making, Consumer Behaviour And Almost Everything Else*. New York, NY: Harper Collins.

MacDougall, H, Langley, G.R. (n.d.) Medical Ethics: Past, Present and Future. Royal College. Retrieved from - http://www.royalcollege.ca/rcsite/bioethics/primers/medical-ethics-past-present-future-e

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*, 45(3), 156–160. https://doi.org/10.1136/medethics-2018-105118

Maani, N., & Galea, S. (2020). COVID-19 and Underinvestment in the Public Health Infrastructure of the United States. *The Milbank Quarterly*, *98*(2), 250–259. https://doi.org/10.1111/1468-0009.12463

Mantha, Y. and Kiser, G. (2019) *Global AI Talent Report 2019*. Retrieved from https://jfgagne.ai/talent-2019/

Marcus, G. and Davis, E. (2019). *Rebooting AI Building Artificial Intelligence We Can Trust*. New York, NY: Penguin Random House.

Mason., J, Morrison, A., & Visintini, S. (2018, September). *An Overview of Clinical Applications of Artificial Intelligence*. Canadian Agency for Drugs and Technologies in Canada. Retrieved from https://www.cadth.ca/sites/default/files/pdf/eh0070_overview_clinical_applications_of_AI.pdf

Mayan, M. J. (2009). Essentials of qualitative inquiry. Walnut Creek, CA: Left Coast Press

Merrigan, G., Huston, C. L. and Johnston, R. (2012). *Communications research methods.* Don Mills, ON: Oxford University Press.

Mertz, E. (2020, May 1). Alberta launches ABTraceTogether app to improve contact tracing, fight COVID-19 spread. Global News. Retrieved from - https://globalnews.ca/news/6894997/covid-19-alberta-health-contact-tracing-app/

Miltner, K. (2019). Girls who coded gender in twentieth century U.K. and U.S. computing. [Review of the books *Programmed Inequality: How Britain Discarded Women Technologists and Lost its Edge* by Marie Hicks, *Recoding Gender: Women's changing participation in computing* by Janet Abbatte and *The Computer Boys Take Over: Programmers and the Politics of Technical Expertise* by Nathan Ensmenger]. *Science, Technology & Human Values. 44*(1), 161-176.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Moor, J. H. (1990). The Ethics of Privacy Protection. *Library Trends*, 39 (1 and 2), 69–82.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/1905.06876

Nebeker, C., Torous, J., & Bartlett Ellis, R. J. (2019). Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Medicine*, *17*(1), 137. https://doi.org/10.1186/s12916-019-1377-7

Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5), e262–e273. https://doi.org/10.1016/S1470-2045(19)30149-4

Noble, S. U. (2018). *Algorithms of Oppression.* New York, NY: New York University Press.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Crown Publishers.

Onwuegbuzie, A. J., Dickinson, W. B., Leech, N. L., & Zoran, A. G. (2009). A Qualitative Framework for Collecting and Analyzing Data in Focus Group Research. *International Journal of Qualitative Methods*, *8*(3), 1–21. https://doi.org/10.1177/160940690900800301

Pandey, S.C., and Patniak, S. (2014, March) Establishing reliability and validity in qualitative inquiry: A critical examination. *Jharkhand Journal of Development and Management Studies*. XISS, Ranchi, Vol. 12, No.1, pp. 5743-5753

PERVADE. (n.d.) Pervasive Data Ethics for Computational Research. Retrieved from - https://pervade.umd.edu/about/general/

Puckett, B., Eckhard, M.O. (2020, March 19). The Latest Covid-19 Conundrum: Can employers institute temperature checks at workplaces? OlgeTree Deakins. Retrieved from - https://ogletree.com/insights/the-latest-covid-19-conundrum-can-employers-institute-temperature-checks-at-workplaces/

Richards, N. M. and Hartzog, W. (2019, April 11) The Pathologies of Digital Consent. *Washington University Law Review*. Retrieved from https://ssrn.com/abstract=3370433

Ricks, B and Surman, M. (2020) Creating Trustworthy AI: A Mozilla whitepaper on challenges and opportunities in the AI era. Mozilla Foundation. Retrieved from - https://drive.google.com/file/d/1LD8pBC-cu7bkvU-9v-DZEyCmpWED7W7Z/view

Rigby, M. J. (2019). Ethical Dimensions of Using Artificial Intelligence in Health Care. *AMA Journal of Ethics*, *21*(2), 121–124. https://doi.org/10.1001/amajethics.2019.121.

Sandelowski, M. (2000). Whatever happened to qualitative description? *Research in Nursing & Health. 23*, 334-340.

Schneier, B. (2020, April 3) Privacy and security implications of Zoom. SchneieronSecurity.com. Retrieved from - https://www.schneier.com/blog/archives/2020/04/security_and_pr_1.html

Semuels, A. (2018, January 23). The Internet Is Enabling a New Kind of Poorly Paid Hell. *The Atlantic*. Retrieved from - https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/

Serebrin, J. (March 30, 2019) *E is for ethics in AI – and Montreal's playing a leading role*. Montreal Gazette. Retrieved from - https://montrealgazette.com/news/local-news/can-montreal-become-a-centre-not-just-for-artificial-intelligence-but-ethical-ai

Shilton, K. (2015). Anticipatory ethics for a future Internet: analyzing values during the design of an Internet infrastructure. *Science and Engineering Ethics*, *21*(1), 1–18. https://doi.org/10.1007/s11948-013-9510-z

Sinclair, S. and Rockwell, G. (2016). *Voyant Tools*. Web. http://voyant-tools.org/.

Slaughter, S., Archerd, C. J., Campbell, T. I. D. (2004). Boundaries and Quandaries: How Professors Negotiate Market Relations. *The Review of Higher Education. 28*(1), 129-165.

Smith, C. S. (2019, November 10). Building a World Where Data Privacy Exists Online. *The New York Times*. Retrieved from https://www.nytimes.com/2019/11/19/technology/artificial-intelligence-dawn-song.html

Solove, D. J. (2013). *Nothing to Hide: The False Trade-off between Privacy and Security*. New Haven, CT: Yale University Press.

Steacy, L. (2020, April 19) T &T Supermarket starting to check temperatures of shoppers, workers. City News. Retrieved from - https://www.citynews1130.com/2020/04/19/t-and-t-temperature-check-covid/

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. Retrieved from http://arxiv.org/abs/1906.02243

Synced. (2020, July 1). *Yann LeCun Quits Twitter Amid Acrimonious Exchanges on AI Bias*. Synced. Retrieved from - https://syncedreview.com/2020/06/30/yann-lecun-quits-twitter-amid-acrimonious-exchanges-on-ai-bias/

Topol, E. (2019) *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York, NY: Hachette Book Group.

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, *114*(2), 246.

West, S.M. Whittaker, M and Crawford, K. (2019). Discriminating systems: Gender, race and power in AI. AI Now Institute. Retrieved from https://ainowinstittue.org/discriminatingsystems.html

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S.M., Richardson, R., Schultz, J. & Schwartz, S. (2018). *AI Now Report 2018.* AI Now Institute, NYU. Retrieved from - https://ainowinstitute.org/AI_Now_2018_Report.pdf

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 195–200. https://doi.org/10.1145/3306618.3314289

Wing, J. M. (2020). Trustworthy AI. In *arXiv [cs.AI]*. Retrieved from - http://arxiv.org/abs/2002.06276

Winner, L. (1993). Upon Opening the Black Box and Finding It Empty: Social Constructivism and the Philosophy of Technology. *Science, Technology & Human Values*, *18*(3), 362–378. https://doi.org/10.1177/016224399301800306

Wittkower, D.E. (2017, October 18-21). Disaffordances and dysaffordances in code. Paper presented at AoIR 2017: The 18th Annual Conference of the Association of Internet Researchers. Tartu, Estonia: AoIR. Retrieved from - http://spir.aoir.org

Zang, J., Dummit, K., Graves, J., Lisker, P., & Sweeney, A. L. (2015). Who Knows What About Me? A Survey of Behind the Scenes Personal Data Sharing to Third Parties by Mobile Apps. *Technology Science*. Retrieved from https://techscience.org/a/2015103001.pdf

Zastrow, M. (2020, May 19). Coronavirus contact-tracing apps: can they slow the spread of COVID-19? *Nature*. Retrieved from - https://www.nature.com/articles/d41586-020-01514-2

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy & Technology*. https://doi.org/10.1007/s13347-018-0330-6

Zimmerman, A., Di Rosa and Kim, H. (2020, January 9) Technology can't fix algorithmic injustice. *Boston Review*. Retrieved from - https://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fix-algorithmic?fbclid=IwAR1_S4wnpOWMDamWmOFaxNEY1m5WoPgc83RqcSROg9j8XMW718AVc88DYMw

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.