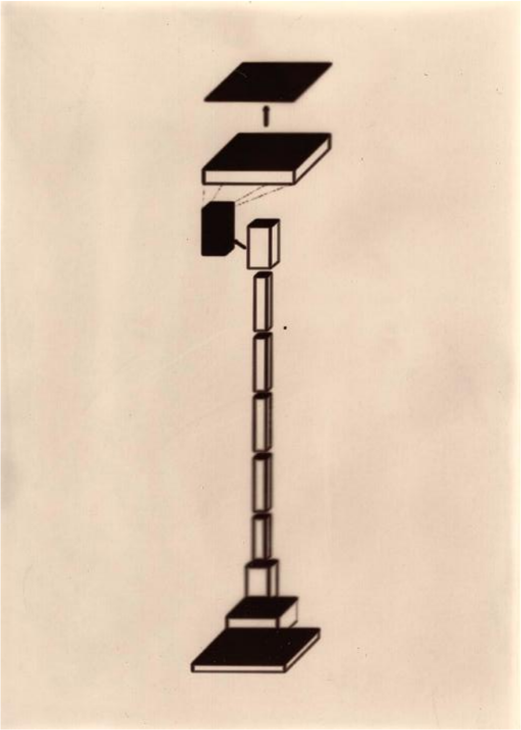


Moving Ethics Beyond Frameworks

A Qualitative Analysis of AI Ethics Operationalization Within Industry

ABSTRACT

As artificial intelligence (AI) has grown in popularity over the last decade, the field of AI ethics has grown in parallel. While many companies, governments, and civil society actors have released high-level AI ethics frameworks, these frameworks alone cannot operationalize AI ethics and cause real world change. While some scholars have provided theoretical frameworks outlining how to operationalize AI ethics, there remains little evidence-based work in the field. This dissertation provides qualitative evidence for the best practices in industry AI ethics operationalization, as well as recommendations for responsible technology professionals.

AUTHOR

Alayna Kennedy
University of Edinburgh
School of Social and Political Science

ACKNOWLEDGEMENTS

Many thanks to Drs. Eugénia Rodrigues, James Stewart, Jake Metcalf, and Manny Moss for their guidance during the STS programme and the thesis writing process. The team at Data & Society provided support, feedback, and a helpful framing for this project, so many thanks to them for allowing me to work with them on this placement. Finally, thanks to Ashley, Sammie, and Grace for their comments and support.

Table of Contents

List of Figures

1. Introduction

Artificial intelligence (AI) ethics is a popular term. In the last decade, the number of academic papers with the words "AI ethics" increased seven-fold (Borenstein, 2021), major tech companies like Google, Facebook, and Twitter have publicly launched responsible technology teams (Hao, 2021), and the market research firm Gartner has cited "Responsible AI" as one of the major competitive differentiators for technology companies (Goasduff, 2021). As part of this wave of attention to AI ethics, a proliferation of frameworks, principles, and policy documents have been released by companies, governments, and civil society (Nourbakhsh, 2021). According to the World Economic Forum estimates, there are now over 175 different AI ethics guideline documents globally (Firth-Butterfield and Madzou, 2020).

In addition to the publication of frameworks, many private-sector organizations have launched initiatives or teams within their organizations to implement responsible tech. Google's People + AI Research (PAIR) team states that their mission is to "do fundamental research, invent new technology, and create frameworks for design in order to drive a human-centered approach to artificial intelligence" (Google Research, 2022, p. 1). IBM's AI Ethics board claims to "support a culture of ethical, responsible, and trustworthy AI" and to serve as a "mechanism by which IBM holds our company…accountable" (IBM, 2022: 1). Even Spotify, the music streaming service, has a dedicated Algorithmic Responsibility team working to "ensure high quality data decisions and equitable algorithmic outcomes" (Spotify Research, 2022, p. 1). Other tech companies have similar teams espousing similar goals, from Microsoft's Fair, Accountability, Transparency, and Ethics (FATE) group to Facebook's research group focused on AI ethics and responsible technology, and (Hao, 2021).

However, the frameworks-first approach to responsible technology has been criticized as insufficient to combatting the harms enacted by AI technologies. Certain AI applications like facial recognition have been shown to be biased against people of color (Buolamwini, 2018), women (Ebell, 2021), and other historically marginalized groups (Acemoglu, 2021). In a meta-analysis of AI ethics guidelines, Hagendorff concludes that these frameworks fail to have a significant impact on human decision-making and organizational outputs in the field of AI (2020). Other researchers take Hagendorff's critique of frameworks even farther, claiming that ethical technology runs fundamentally counter to technology companies' business model, and that current industry AI ethics is merely a way for companies to receive better press or higher profits (Lauer, 2021; Morley, et al., 2021a; Wagner, 2018). As one critic writes, "while AI ethics frameworks may make for good marketing campaigns, they all too frequently fail to stop AI from causing the very harms they are meant to prevent" (Burt, 2020, p. 2).

Much of the literature on AI ethics frameworks, including its critiques, fail to base their findings on empirically grounded research about the implementation of ethical AI systems within the context of the companies who deploy them. Although the larger realm of responsible technology has a long history, corporate investment in AI ethics teams and tools has only been a significant source of investment for a few years (Markman, 2019). While this recent investment has led to the development of many high level frameworks, it is less clear how companies are implementing this contextual work into their business practices. (Georgieva, et al., 2022; Morley, et al., 2020).

This paper aims to understand what's going on behind the scenes of these industry ethics teams and to assess whether frameworks accomplish the state goals of "holding themselves accountable" and "ensuring equitable outcomes."

As part of Data & Society's AI on the Ground initiative, this paper uses findings from interviews with responsible technology professionals research to understand the impact of AI systems by speaking with people *on the ground* – those doing responsible tech work within industry. To understand the ways AI ethics is implemented beyond frameworks, I interviewed responsible tech professionals to ask about the artifacts they use and create in their work, how they navigate the business case for responsible tech, and how to implement frameworks into organizational practices (Data & Society, 2022).

Building on the findings from these interviews, this paper identifies five recommendations for implementing responsible tech frameworks within the context of an industry organization. Hopefully this paper can provide insight to industry professionals working on responsible tech, scholars, regulators, civil society groups, who are working on implementing real-world AI ethics into the messy and contextual nature of organizations.

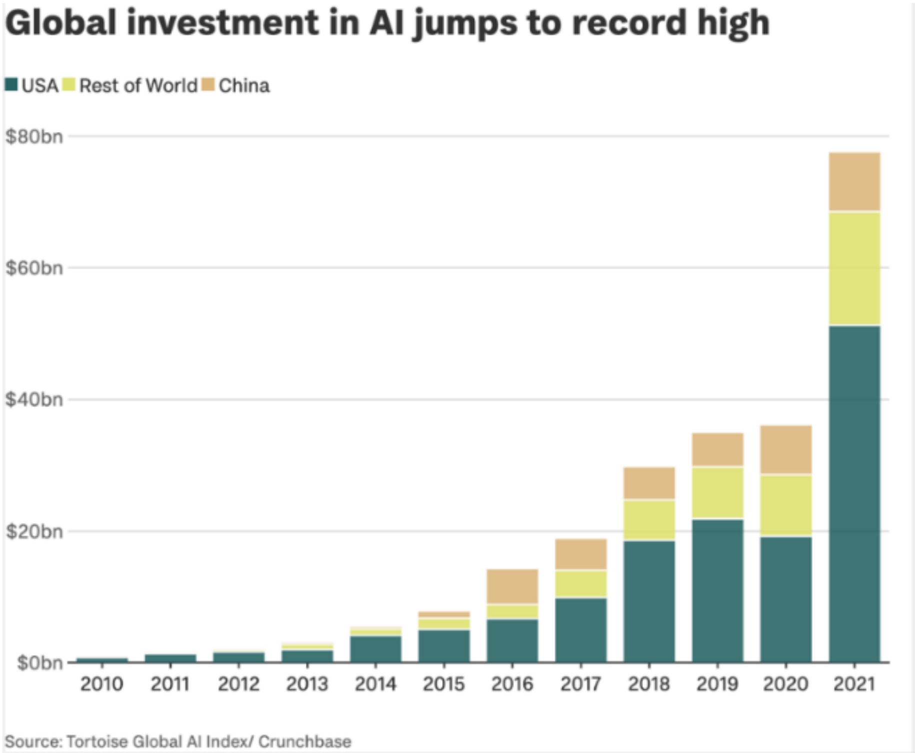
1.1 Structure of dissertation

This paper has four main sections: first, a historical analysis of AI development shows why AI within industry is an important current artifact. Second, I analyse the current AI ethics literature to show that much of the current analysis of ethical frameworks fails to focus on the operationalization of AI ethics. Third, I present my qualitative data collection methodology and my analysis of findings. Fourth, I discuss the findings from my interviews in more detail and presenting my recommendations for responsible tech professionals implementing frameworks. I conclude with a brief discussion of this study's limitations and ideas for future work.

2. Modern AI as an industry-based artifact

In the last decade, the term Artificial Intelligence (AI) has exploded in popularity, with technical breakthroughs in language models, image generators, and classification systems ushering in speculation about the effects of AI on our society, businesses, and governments. Around 2009, the term Big Data began gaining traction in corporate circles, with massive amounts of digital data promising better information and understanding, if only companies could process and analyse it. Shortly after, technical breakthroughs in the development of Deep Learning promised human-level AI performance (Alom, et al., 2018; Cukier and Mayer-Schoenberger, 2013). By the 2010s, the hype around big data had been replaced by AI, a sub-branch of machine learning (Tran, 2017).

In 2010, researchers at Google created a deep learning speech recognition tool that vastly surpassed the current cutting-edge industry approaches. The next year, IBM's Watson beat champion Ken Jennings at Jeopardy, and a year after that, Google's famous computer vision paper on identifying cats went viral (Dormehl, 2019). These prominent displays of AI's potential caught the imagination of the public and attracted the attention of businesses and investors, with investment in AI increasing sharply during the 2010s with the advances in AI technology. Gartner, a global IT research firm, surveyed 3,000 CIOs from 89 countries, and found that products claiming to use AI technologies grew 270% between 2014 and 2018 (Markman, 2019). Since 2010, 154,000 AI patents have been filed globally, with a "a strong uptick in the patent category covering AI, machine learning and neural networks" (Spencer, 2022, p. 2).



Source: Tortoise Global AI Index/ Crunchbase

Figure 1. Global investment in AI increases (Darrah, Mehta and Mousavizadeh, 2022)

The AI hype of the last decade builds upon a long and storied history of high expectations and public optimism around AI that dates to the term's inception in the 1950s (Goode, 2018). Throughout the last century, AI has captured the imaginations of scientists and publics alike with its preliminary results, but in the past, these results have failed to generalize or scale to real-world problems. Much like the automata of the past, these modern systems rely on human imagination to fill the gaps in their performance. Just like the mechanical Turk of the 18th century relied on mystery, so too does modern AI rely on our collective imagination of it. In this way, AI has been formed by the promises and performance of the social systems that created it, from the public imaginary of AI as a viable technology to the institutional investment in its development. In this next section, I show that AI is the product of a decades-long social imaginary, a "collectively held, institutionally stabilized, and publicly performed vision" (Jassanoff, 2015, p. 6). Furthermore, the sociotechnical artifact of AI is inherently bound up in the institution of private corporations.

2.1 Historical AI development – from myth to materiality

AI has become a buzzword in the last 10 years, and this paper focuses on the newest wave of sociotechnical developments that are calling themselves AI. However, the idea of automated machines has a long and storied past. Automata appear in myths and legends, from the ancient Greek machine-warrior Talos to the flask-bound homunculus, an alchemically made human intelligence imagined by Goethe (Borenstein, et al., 2021; Lungarella, et al., 2007; Walmsey, 2012). More recent examples include the famous Mechanical Turk of the late 18th century, a machine that seemed to be able to independently play chess and win against most opponents (Stephens, 2022).

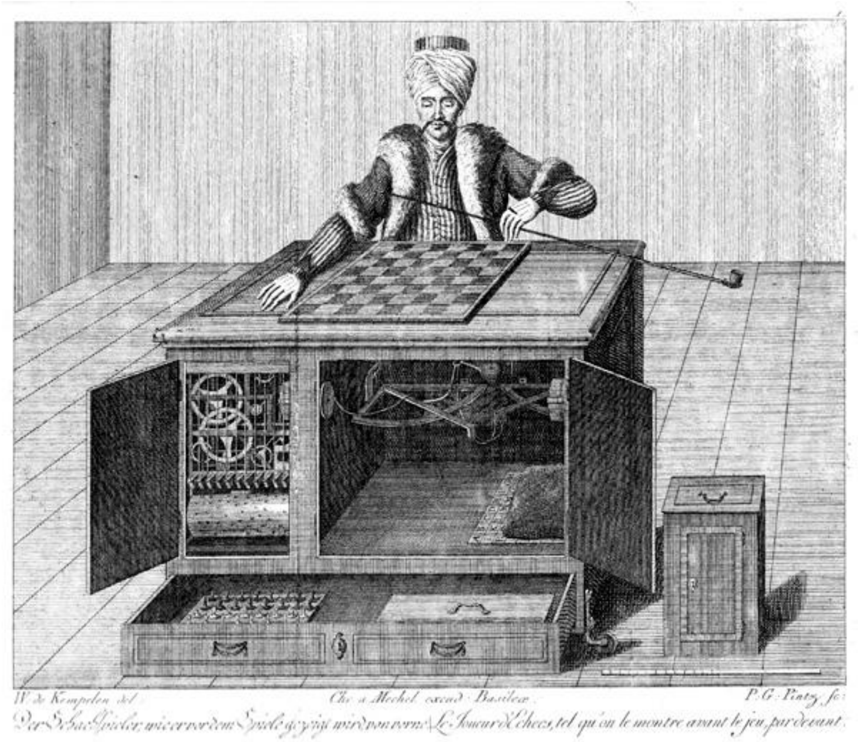

W. de Kempelen del.
Chr. a Mechel excud. Basileae.
P.G. Pintz fc.
Der Schachspieler, wie er vor dem Spiele gezeigt wird von vorne | Le Joueur d'Echecs, tel qu'on le montre avant le jeu, par devant.

Figure 2. A depiction of the Mechanical Turk at its debut in Vienna in 1770 (Stephens, 2022)

227 years after the Mechanical Turk made its debut in Vienna's Schönbrunn Palace, IBM's Deep Blue supercomputer beat world chess champion Gary Kasparov in a six-game match (Campbell, 2002). This time, there was no hoax, no trick to the machine, just lots of compute power, data, and the resources of one of the world's largest organizations. However, just like the Mechanical

Turk, Deep Blue caused uncertainty about the limits of automation and the potential of AI technology.

Deep Blue was the product of almost 50 years of advances in computer science since the origins of the field of artificial intelligence. While the precursors to AI technology (fields like informatics and cybernetics) developed during the World Wars, the term "Artificial Intelligence" came into being at the Dartmouth Workshop of 1956 (Lungarella, 2007). The workshop was populated with well-respected researchers: Norbert Wiener's field of cybernetics was prominent in the field of electrical engineering, while fellow attendee Claude Shannon had worked with Alan Turing on information theory describing digital signals in World War II (Cordeschi, 2007; Müller, 2021). In the early 1950s, their fellow computer scientists created programs that could play checkers autonomously and independently develop mathematical theorems, convincing many of the Dartmouth attendees that computer science needed to dedicate a specific field to the development of automated intelligence systems (Borenstein, et al., 2021). This conference coined the term "artificial intelligence," or "AI". The label was used throughout the 20th century in popular discourse surrounding machine intelligence. In popular and academic discourse, the term AI represented the promise of computers emulating ever more aspects of human cognitive performance.

Many of the attendees of the Dartmouth workshop were tremendously optimistic about the potential progress that AI could make. One attendee, Herbert Simon, predicted in 1965 that "Machines will be capable, within 20 years, of doing any work that a man can do." (Simon, 1965, p. 96). Indeed, in the two decades after the conference, symbolic AI systems produced results that astonished scientists and the general population. Computers began solving algebra word problems, communicating in natural language, and sensing the world through computer vision (Borenstein, et al., 2021; Müller, 2021). While limited, these results inspired wide-scale optimism about the development of AI technology. Early scholars' optimism about AI, coupled with high public expectations about the technology, led to huge investments in AI (Alom, et al., 2018). By the late 1960s, government agencies like the Defense Advanced Research Projects Agency (DARPA) had granted huge sums for AI research projects that focused on handcrafted knowledge, or rule-based systems capable of narrowly defined tasks (Jensen, Whyte, and

Cuomo, 2019). While a critical step forward for the field, these systems were, by DARPA's own admission, "fragile and limited" (DARPA, 2018, p. 1)

Systems like DARPA's showed that the early optimism about AI was unfounded, and disappointing results like DARPA's led to major funding cuts in AI research in the late 1960s (Toosi, et al., 2021). The removal funding led to the first AI "winter," a period marked by lack of financial and research interest in AI (Lim, 2018). There were several reasons for this first failure of AI to deliver expected results: many early AI systems attempted to replicate humans' thinking processes when creating algorithms, which resulted in computationally inefficient processes, and many researchers oversimplified their AI frameworks, leading to most early systems only succeeding on simplistic "toy problems" that had few real-world applications (Shevlin, et al., 2019; Toosi, et al., 2021).

Critiques of AI continued throughout the rest of the 20th century, with the AI winters of the 1970s and 1990s only broken by a decade of interest in expert systems in the 1980s. While these "expert systems" were quickly adopted by corporations, they too ultimately failed due to their small size, limited application, and lack of data to train their models on (McClay, 1995; Toosi, et al., 2021). Both AI winters occurred because AI failed to deliver value to institutional stakeholders – first DARPA, then corporations. When the optimism around AI failed to produce real results, economic funding dried up, and the technical project of AI development was discarded (Haenlein and Kaplan, 2019). After expert systems failed and AI entered its second winter, the field had no funding or credibility. AI was so scorned that researchers tended to avoid even using the term "AI," opting for terms like "informatics" or "analytics" instead (Floridi, 2020; Toosi, et al., 2021). Less than 30 years later, in 2016 the market for AI-related products, hardware, and software reached more than 8 billion dollars, and the New York Times reported that interest in AI had reached a "frenzy" (Lohr, 2016, p. 1). AI funding is now at an all-time high, with over $80 billion being spent on AI applications every year (Hao and Kruppa, 2022; Markman, 2019).

2.2 Resurgence of AI

Of course, the history of AI development described in the last section is an oversimplification of the complex ecosystem of technology and telecommunications development. The overall ecosystem of computing never stopped developing throughout the years of AI winter. However, AI's move from obscurity to market dominance in a few decades can illustrate a few key qualities of the modern AI hype. In the next section, I will show how the emergence of big data, deep learning, higher compute capability, and an economic model based on data-driven insights brought AI back to prominence. However, in addition to catapulting AI to prominence, these factors have also changed what most people mean by "AI."

2.2.1 Sociotechnical Imaginary of Modern AI

Scholars of science and technology studies (STS) have long espoused the belief that technology can be shaped by social forces; a theory known as the "social shaping of technology" (SST) (Edge, 1996). SST holds that the development of technology is not predetermined to follow a single path, but that innovation is a "garden of forking paths" (MacKenzie, 1999). Social choices and subtle contextual differences can lead to different technological outcomes in this garden. One of the important social choices that can direct technology's path along these forking paths is the collective imagination about technology.

The 'imaginary' as defined by Ricoeur is the "dialectic relationship between a horizon of expectations and a space made of experience" (Castoriadis and Ricoeur, 2016: 58). In this framing, the imaginary serves as a shaping force of culture and socio-economic forces – by imagining possible futures for the ways our technologies could work, the trajectory of innovation and investment can be changed (Flichy, 2002).

We've seen that AI has a long history of optimistic imaginaries, followed by institutional investment, and disappointing delivery. The modern iteration of AI development follows similar patterns – initial hype around the technology, promises made by corporate actors, and then a movement behind AI taking on a performative role to shape material reality.

2.2.2 Economic factors making insights and ads profitable

While AI was going through its second resurgence in the 1980s with the rise of expert systems, Western countries political economies began shifting toward neoliberal policies promoting free markets and decreasing government regulation (Harvey, 2007). Deregulation, privatization, and minimization of state interventions in markets were the cornerstone of many of these new policies, which became common in the UK and US throughout the 1980s. (Lim, 2011). Because regulation was seen as impeding innovation and hindering the United States and other Western countries from competing at a global level with countries like Japan and China, the computer industry managed to maintain their relative independence from regulation, within a policy environment that had come to recognise the value of low regulation. This regulatory environment led to the rise in profitability for technology and network-related companies in the 1990s, a phenomenon known as the dotcom bubble (Kavenna, 2019; Zuboff, 2015).

When the dotcom bubble burst in 2001, Google, Yahoo, and other internet-based companies began shifting their business model toward using data for profit – selling advertisements and behavioural insight products on the data generated by their customers/users interactions with their systems (Zuboff, 2015). Data-as-profit, the Big Data business model that proposed using previously unusable system data to create corporate value and profit, marked an important shift in the digital economy, when companies and marketers realized they were able to harvest data about individual's online behaviour (search queries, time spent on each website, clicks, etc.) to augment advertisement and drive online purchasing behaviour (Cheney-Lippold, 2011; Ball, 2019). This data could be used to "predict, modify, and actively control how individuals behave" (Ferreira, 2022: 274), making it massively valuable to ad-based companies like Google and Facebook (Zuboff, 2015). The promise, and in some case the reality, was that interaction data could be used categories consumer groups to open massive scale targeted marketing. In this way, the promise of AI profit helped shape the reality of AI investment, with the performance of companies shaping the material reality of the market (Glass and Rose-Redwood, 2014).

With the regulatory and economic environment making data collection profitable, internet companies began gathering unprecedented amounts of data about users. This led to a massive increase in the amount of global data, a phenomenon often referred to as “Big Data.”

2.2.3 Birth of Big Data

A mere 20 years ago, digital information accounted for only 25% of the globally stored information. Now, digital information is over 98% of the world’s stored information (JSTOR). In 2010, the entire amount of data created, consumed, and stored was just over 2 zettabytes (ZB), but by 2020, this number had skyrocketed to 59ZB (Statista, 2022; Vopson, 2021).

Digital information and digital data are closely related phenomenon, but crucially separate. As many STS scholars have noted, raw information becomes data through a process of social shaping – when human beings select, shape, and present key bits of information and call them “data” (Smagorinsky, 1995). In the rest of this section, I will refer to digital information as “data,” loosely defined as the information in a computer, or information used by computer scientists (Floridi, 2021).

The scale of data storage and creation is massive. However, big data is not characterized by size alone, but also by the process of “datafication” of the world – “the ability to render into data many aspects of the world that have never been quantified before” (Cukier and Mayer-Schoenberger, 2013, p. 3). For example, location becomes data when translated into latitude and longitude, words become data when processed by natural language processing programs, and even social relationships become datafied by systems like Facebook’s “friendship” and “like” structure (Barnes, 2013).

Much of this data creation happens on the Internet, with companies like Google recording clicks, location data, search history, and other user behaviour. In recent years, companies have begun collecting data from smart devices, mobile phones, wearable devices like smart watches, and virtual home assistants like Amazon’s Alexa (Enterprise Big Data Framework, 2019).

As discussed in the previous section, these companies are driven by the increasing economic profitability of data in a technology-driven market. According to Zuboff, user-generated data drives the manufacture of "prediction products," which are profitable for companies to sell, but which require massive amounts of user data (Zuboff, 2019). As more and more companies expand into internet-based products like internet of things connected devices, online stores, and app-based products, they continue to create massive amounts of data (Foster and McChesney, 2014). This drive toward datafication has generated massive amounts of data in the last few years alone. By some estimates, over 95% of the data that exists today was created in just the last 5 years (Vopson, 2021). While a significant amount of this data creation results from the digitization of previously existing sources like text and images, the majority of data creation is in the clicks, likes, and interactions of modern internet users (Vopson, 2021; Zuboff, 2019).

One of the limitations of historic AI was its lack of scalability, a problem that access to more data can solve. However, even with large datasets, traditional AI systems were still limited by their lack of depth and processing power. Along with the advent of Big Data, researchers in the 2010s solved the problem of AI depth with the advent of deep learning.

2.2.4 Technical breakthrough in "Deep Learning"

The term deep learning (DL) refers to a branch of machine learning that uses many layers of algorithms to process data. While the term traces its origins back to the 1950s, it became popularized in 2006 with Geoffrey Hinton's famous paper on Deep Belief Networks (Alom, et al., 2018). In 10 years after that paper's publication, researchers reported better and better performance using DL, including human-level speech recognition tools, image recognition, and text generators (Dormehl, 2019). Technical developments in the architecture of machine learning systems were part of the reason deep learning rose to prominence, but as technology writer Varun Bansal says, "you can't ignore the increase in computational powers and availability of large datasets" (Bansal, 2020, p. 2).

Deep learning was the technical tool needed to harness the raw material of data and transform it into real-world applications. While historic AI systems often failed due to lack of data and limited model size, deep learning, fed by big data, produced scalable results. However, this technology required lots of compute power to run, a resource that historically only a few large companies and major universities would have access to (Waldrop, 2019). Although models are become less computationally expensive at the same time as microprocessors are being redesigned to process larger models, most compute resources are still corporate controlled (Whittaker, 2021).

2.2.5 Growth of corporate compute power

While DL's technical advances in the last decade are not trivial, they are fundamentally based on ML techniques originally developed in the 1950s. Their astonishing results come from the combination of large amounts of data couple with increased compute power (Alom, et al., 2018). While compute power has been increasing rapidly for the last several decades, the amount of compute power needed for the largest and most sophisticated models is vast. For example, as of 2021, OpenAI's near-human language model, GPT-3 would need 1,024 GPUs to train the model in 34 days, a process that cost over $5M in electricity for the GPUs alone (Romero, 2021).

Industry actors, who hold the most capital, data, and compute power, and monetary funding, therefore have the best access to creating new AI models (Whittaker, 2021; Hagerty, 2019). As scholar Meredith Whittaker writes, "Modern AI is fundamentally dependent on corporate resources and business practices" (Whittaker, 2021, p. 1). The combination of economic profitability of data, technical advances in deep learning, and access to huge amounts of compute power, have made the historic foundations of AI into a new, profitable, and industry-based phenomenon.

2.3 Companies now use the term "AI" to refer to many combinations of big data, deep learning, and compute-driven products

More data, better DL architectures, and compute power have all driven the rise of AI systems. However, they have also changed the way that people talk about AI and chose to define the term. Because modern AI generally relies on data generated from large internet companies and large, expensive amounts of compute power, it is usually only deployed, and its profits only realized, by industry actors. Building off the profitability of data collection, AI promises industry investors a way to turn their massive amounts of data into insights, knowledge, and cost-reduction strategies (Lyon, 2019; Minkkinen, et al., 2022). Because of larger datasets, deep learning techniques, and huge amounts of computer power, the term "AI" is used much more broadly than it was in the 20th century. Many examples of broadly defined AI can be found in industry: an automated network of sensors, pattern-matching filters, and rules-programmed chatbots have all been referred to as "AI," even though they fail to meet the historical technical definition of the technology (Bogost, 2017).

While modern use of the term AI has been criticized for not adhering to the historic, technical definitions of the technology, the adoption of "AI" as a term to refer to multiple data-driven or computerized processes has spread widely enough to be taken seriously. Again, we return to the concept of promise and performativity in AI development: because of the situation of AI within a network of industry actors, corporate directors and marketing teams of technology firms have been able to set the agenda about AI that policy makers, politicians, and the general public care about. This definition and use of "AI" as a firmly industry-based sociotechnical artifact did not happen by chance, but rather through the combination of actions taken by these actors over years. Now, when people talk about AI, they are often talking about the combination of big data and computerized approaches to extracting knowledge from that data. They are talking about a new sociotechnical artifact that is entwined with the economic profitability of data collection and the industry-dominated computational ability. Therefore, this new wave of AI hype is a uniquely industry-based phenomenon. In the next chapter, I will discuss the AI ethics literature that has arisen in tandem with AI's popularity, and how most of the current literature fails to address the industry-based nature of this technology and the ways it intersects with other industry processes like change management, innovation, and business ethics.

3. AI ethics literature has moved from frameworks toward operationalization but lacks qualitative evidence

In the first chapter, I showed that the rise of the data-driven ad economy, the birth of big data, developments in deep learning, and the consolidation of compute power in large institutions has transformed AI into a new sociotechnical artifact inherently bound up in industry. The adjacent field of AI ethics has emerged in tandem with this new, industry-centric form of AI. Just as the new form of AI draws on, yet differs from, its historical roots, so too does AI ethics have a long historical trajectory that has recently shifted.

Historical conceptions of AI ethics focused on philosophical issues about computer consciousness and existential issues of human survival in the case of emergent machine superintelligence. While there are scholars who still work on these issues, recent years have seen a rise in AI ethics frameworks, tooling kits, and operationalization methods (Wong, Madaio, and Merrill, 2022).

3.1 Historical conception of AI ethics

While there has been a distinct rise in AI ethics work over the last decade, the core issues of the responsible technology field are not new. Dating back to the origins of the term AI in the 1950s, scholarly optimism led to speculation among journalists, filmmakers, and the public about the potential of automated computer system. In pop culture alone, dozens of films, TV shows, and other pieces of popular media about the potential impacts of AI systems were released (Borenstein, et al., 2021; Douetteau, 2020).

Early versions of AI ethics emerged in academia in the 1980s and 1990s within the realm of science and technology studies. “Computer ethics” was a term coined by Deborah Johnson in 1978 in her paper about using legislation to control to use of software, and the first textbook on computer ethics was published in 1985, followed by a “steady flow of monographs, textbooks, and anthologies in the 15 years that followed” (Müller, 2022, p. 10) By the 1990s, journals that specialized in computer ethics were being established, including the Journal of Information

Ethics in 1992, Science and Engineering Ethics in 1995, Ethics and Information Technology in 1999, and Philosophy & Technology in 2010 (Müller, 2021). Despite popular interest in the philosophical implications of intelligent machines, as recently as 2010, scholars considered digital ethics as “marginal or specialist…with very few presentations at mainstream conferences, publications in mainstream journals, or posts at mainstream departments” (Müller, 2021, p. 11).

3.2 The rise of AI and the parallel rise of AI ethics

Just as AI moved from relative obscurity in the 1990s to narrative prominence 30 years later, so too have issues of AI ethics risen in popularity in the last decade. Last chapter, we discussed the multiple factors that led to a rise in the use of AI within industry – deep learning development, big data, etc. Each of these waves of AI development have been accompanied by ethical concerns and mitigation strategies (Kennedy, 2020).

Much AI ethics work originally came from academic criticism of industry, like the seminal “Gender Shades” paper from MIT researcher Joy Buolamwini (2018) that showed facial recognition technology is biased against people of color, and women of color in particular. Buolamwini’s work was part of a wave of other papers about the ethics of AI, responsible technology, and the principles of AI. Papers that focus on these topics have become exponentially more popular in the last several years, paralleling the rise in AI investment and development.

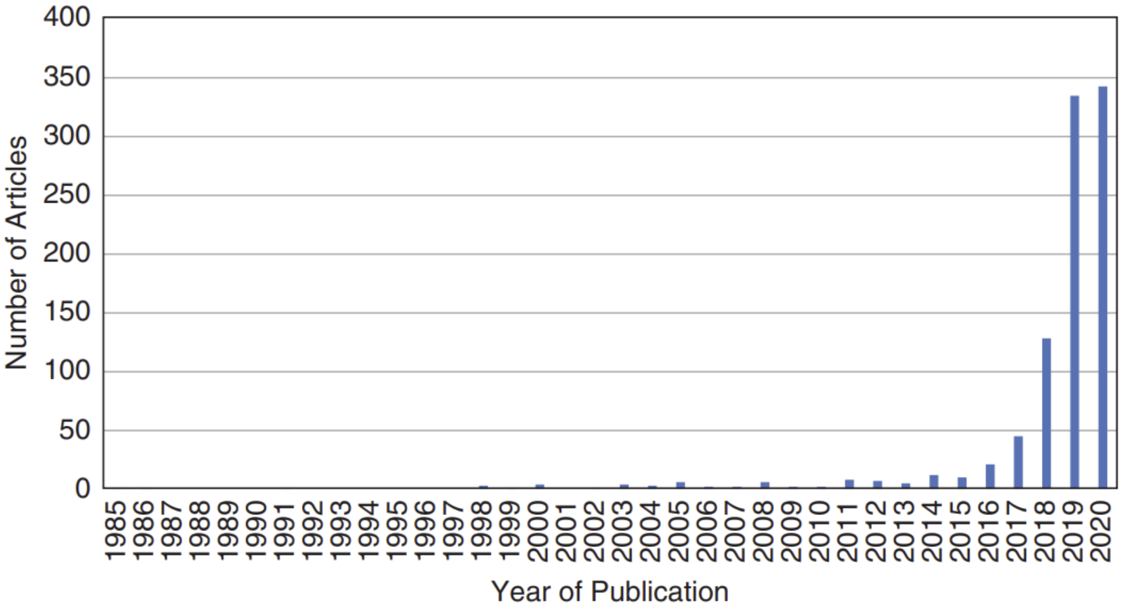


Figure 3. The counts of Google scholar citations about AI ethics (Borenstein, et al., 2021)

Just as AI has become a buzzword in industry use cases, academics and civil society researchers have begun to talk about the need for "AI ethics" to mitigate these harms (Ebell, 2021; Hagerty, 2019). A great deal of the research efforts in the field of AI ethics have focused on mitigating bias and unfair machine learning (ML) in the technical stages of development, such as detecting historically biased data and removing disparate impact from a model's output (Mitchell, et al., 2019).

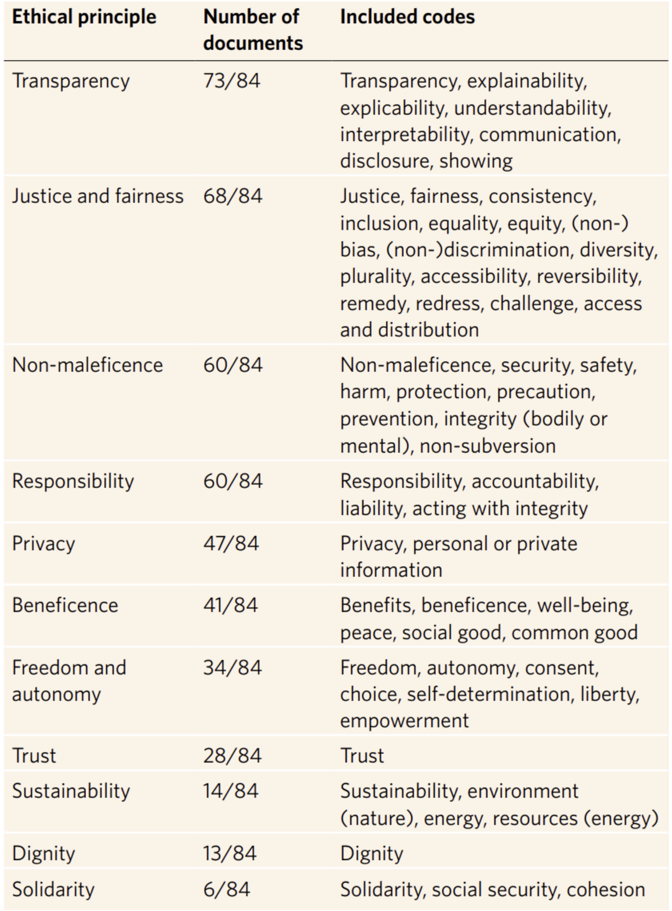
Although the field of AI ethics was borne out of academia's criticism about industry use of AI, industry actors have adopted many of the artifacts produced by academic and civil society researchers. While the field of AI ethics was borne from academics like Buolamwini who were concerned about racial and gender disparity, the idea of responsible AI quickly spread to industry and government. Many companies have released their own work on technical bias and fairness mitigation techniques (Buolamwini and Gebru, 2018; Lauer, 2021).

Industry actors have begun investing in "Responsible AI" as a differentiator for technology companies (Markman, 2019). As one study from the Economist Intelligence Unit concludes there is a substantial business case for AI ethics, from increasing employee retention to preparing for upcoming regulation and alleviating corporate customers concerns about the reliability of AI (Haider and Islan, 202). As the analyst firm Gartner wrote in a recent report about the AI market landscape, "moving forward, organizations must develop and operate AI systems with fairness and transparency and take care of safety, privacy, and society at large" (Markman, 2019).

Because of these motivating factors, companies have been investing in responsible technology teams and research. Most prominently, companies, governments, and civil society organizations have begun to release high-level AI ethics frameworks. Between 2016 and 2019, 74 sets of ethical principles or guidelines for AI were published, focusing on high-level guidance like "creating transparent AI" (Hilligoss and Fjeld, 2021). The World Economic Forum estimates that over 175 global AI ethics and responsible technology frameworks have been released to date (Firth-Butterfield and Madzou, 2020).

3.3 As AI ethics moves from academia to industry, the dominant artifact is the "framework"

Frameworks are the most prominent emergent artifact in the field of AI ethics (Ashok, et al., 2022). While many different frameworks have been released, they generally coalesce around similar principles. A meta-analysis of principles and guidelines released by private companies, research institutions, and public sector organizations show general convergence around shared principles (Jobin, et al., 2019). These general principles are Transparency, Fairness, Non-Maleficence, Responsibility, and Privacy (Jobin, et al., 2019). However, while there is some tenuous consensus on the high-level principles, there remains a gap in these principles and practical techniques for implementation of AI ethics within an organizational context (Giziński, Kuźba, and Biecek, 2020; Sanderson, et al., 2021).



Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

Figure 4. Ethical principles identified in global AI guidelines (Jobin, et al., 2019)

As this meta-analysis suggests, while a general consensus around *what* principles should be a part of AI ethics frameworks has emerged, these high-level guidelines fail to provide substantive guidelines for *how* to operationalize those principles. Some critics of frameworks claim that they fail to enact any real changes in the field of AI and ML, like Hagendorff (2019), who reviewed 22 major AI ethics guidelines and concluded that they "most often" fail to have an actual impact on the development and deployment of AI technologies (Hagendorff, 2019, p. 2). Other critics have noted that AI ethics development lacks methods to translate these principles into practice, and that no external mechanisms of accountability for private companies developing AI yet exist (Mittelstadt, 2019a).

3.4 Lack of operationalization of AI ethics frameworks

The fact that most high-level guidelines on AI ethics tend to converge around these major principles is a good sign: it suggests the emergence of industry standards and best practices (Mökander, et al., 2021). However, despite the proliferation of AI ethics principles, in the last few years there were still many public instances of AI systems causing harm. From examples of unfair treatments in the healthcare industry (Villarreal 2020), to biased A-level prediction tools (Coughlan 2020), to discriminatory use of AI by law enforcement (Pastaltzidis, et al., 2022), it has become clear that AI principles alone cannot prevent AI-related harms (Mittelstadt, 2019b; Morley, et al., 2021b).

Frameworks alone have limited practical applicability to the real-world design and deployment of AI systems. However, implementing principles in practice requires organizational trade-offs, consensus among different stakeholders, and clear metrics and tools to measure and score the success of AI ethics initiatives, all challenging tasks that have caused technology companies to struggle with operationalizing AI ethics *(*Mökander, et al., 2021; Olah, et al., 2018).

Initial efforts to move ethics beyond frameworks have tended to focus on technical solutions like fairness toolkits and protocols for keepings humans in the loop of technical decision making (Mökander, et al., 2021; Hagendorff, 2021). While these tools serve a purpose in the large system of AI development, many of the current toolkits are immature (Morley, et al., 2020) and ineffective in addressing the wider social harms of AI. An ML model trained on a dataset tested for statistical bias by a toolkit will not automatically perform ethically without oversight, but requires periodic oversight and assessment, since ethics is not a "bolt-on," but a continuous process (DeBrusk, 2018; Kennedy, Coates, and Lindquist, 2020).

In a survey of these tools, scholars from Oxford found that "almost all existing translational tools and methods are either too flexible (and thus vulnerable to ethics washing) or too strict (unresponsive to context)" (Morley, et al. 2021a, p. 2). However, the same study suggests that the limitations of current tools and frameworks can be overcome by a theoretical concept called "Ethics as a Service" (Morley, et al., 2021a).

3.5 Moving beyond frameworks to “Ethics as a Service”

AI developed over the last decade in response to the promise of corporate profit, and industry actors currently have an outsized role in shaping the development of both AI and AI ethics (Whittaker, 2019). While frameworks have emerged as a prominent artifact, critics note that they often fail to have real impact, potentially because an implementation of systemic ethical principles presents a threat to the profitability of AI (Hagendorff, 2019; Lauer, 2021). Frameworks fit easily into an existing business context without changing business structures. However, a new wave of AI ethics work is emerging that focuses on operationalizing AI ethics in practice. Like frameworks, this operationalization work is firmly rooted in an industry context and focuses on using traditional business development techniques like change management, business ethics, and corporate education (Georgieva, et al., 2022).

Scholars from the Oxford Internet Institute define this new focus on operationalization as “Ethics as a Service,” based on the platform-as-a-service or infrastructure-as-a-service models common in industry (Morley, et al., 2021b). This method of AI ethics is fundamentally rooted in existing industry practices and involves distributing the responsibility of operationalizing AI ethics across internal business units and external stakeholders. This process involves distributing organizational responsibility across multi-agent systems of companies, individuals, and external stakeholders like governments (Floridi, 2016). In the world of AI ethical governance, this could involve “several components including, but not necessarily limited to: an independent multi-disciplinary ethics board; a collaboratively developed ethical code; and AI practitioners themselves” (Morley, et al., 2021b, p. 259).

However, while scholars theorize that framing ethics as a service will overcome some of the limitations of the current framework-based approach, whether it will work in practice “is yet to be seen” (Morley, et al., 2021b, p. 250).

Most of this current work lacks input from responsible technology professionals, failing to examine the current embedded practices for implementing AI ethics (Canca, 2020; Zhu, et al., 2022). While there have been focused case studies (Mökander, et al., 2021) and workshop discussions on AI ethics (Ehsan, et al., 2021), the lack of on-the-ground evidence of best practices represents a major gap in the current literature. In this study, I aim to provide qualitative evidence for the effectiveness of ethics as a service by interviewing responsible technology professionals and using their perspectives and knowledge as a basis for normative recommendations.

4. Methodology of qualitative data collection and analysis

We have seen AI and AI ethics are bound up in industry promises, performativity, and corporate actor networks. My primary question for this study was to examine the ways that AI ethics is operationalized within an industry context: beyond just publishing a high-level framework, what are technology companies *actually* doing to implement "ethical AI?" I interviewed eight professionals working on AI ethics in industry to gain their perspectives and understanding of the current state of AI ethics within industry. While I was constrained by the time (only 2 months to organize and conduct interviews) and access (not everyone wanted to speak to a master's student), I believe that getting this qualitative data from professionals on-the-ground presents valuable insight into the current state of AI ethics in industry. In this section, I outline my methodology in conducting the interviews and my approach to collection and analysing my qualitative data.

4.1 Semi-structured interview methodology

There are many theoretical approaches to asking people for information, from highly structured surveys to unstructured feedback from focus groups (Adams, 2015). In between these two extremes lie the methodology of *ethnographic interviews* or *semi-structured interviews* (SSI). These conversations, conducted with one interviewee at a time, use a combination of closed- and open-ended prompts, follow up questions, and dialogue to collect qualitative knowledge from the interviewee (Rabionet, 2016). SSIs provide a breadth of contextual information about specific questions when conducted by an interviewer familiar with the subject matter (Adams, 2015; Creswell and Poth, 2016).

4.2 Establishing ethical guidelines

As per University of Edinburgh requirements, I submitted an ethics proposal for the interview process. Importantly, since the dissertation is in collaboration with Data & Society, I had to ensure that all interviewees knew their transcript data and the findings from the research might be used in future Data & Society reports. In addition, I provided every interviewee the option to remain anonymous, so I had to ensure to completely de-identify the transcripts of the two interviewees who requested anonymity.

4.3 Incorporating research questions into an interview structure

The flexible nature of SSIs means that a structured series of questions cannot be created for the interviews. Instead, an interview guide, or outline of planned topics should be laid out in a tentative order according to their priority (Adams, 2015; Allen, 2017)). In crafting the interview guide, I needed to work with the team from Data & Society, since my interviews would be used in their larger AI on the ground initiative, I needed to align my interviews with the information they were looking to incorporate into their larger project.

Below are the high-level thematic changes to the study influence by Data & Society's work:

AI on the Ground research objectives		**Influence on interview structure and analysis**
Looking for the smallest details about to tools, outputs, and specific artifacts responsible tech professionals use	→	My highest priority area of questioning centred around artifacts – frameworks, tools, structures, and other outputs
Ethnographically oriented research approach focuses on the ways AI governance is integrated within social contexts	→	My questions focused as much on the social and organizational processes as the artifacts, trying to ask interviewees about the social context of their work
Focus on articulating emerging issues and developing guidelines, best practices, and recommendations for new approaches	→	Not only did I want my interviewees to be descriptive, to tell me about what's currently happening, but I also had to ask them to be normatively prescriptive, to tell me what they think *should* happen in the future.

Figure 5. Data & Society's research objectives' influence on methodology

4.4 Process of conducting and recording interview data

Each interview lasted 1 hour and was recorded on Zoom. At the start of the interview, I addressed the ethical consent of each participant, confirming orally if they wanted to remain anonymous or consented to having their information shared in the final report.

Preparation for each of the interviews was a vital part of the process, since SSIs require a competent and dedicated interviewer with enough experience and knowledge to guide unstructured conversations in a way that still answers the core research questions (Adams, 2015).

Voice recordings of the interviews were stored on my personal laptop, where I used the Otter.AI transcription system to automatically transcribe them into text format. I reviewed each transcript for accuracy before uploading them to NVIVO where I began the process of analysis.

4.5 Qualitative data analysis of interviews

After collecting and processing the interview data, I began the process of analysis. In analysing the interviews, I applied a thematic analysis grounded in social constructivist and inductive knowledge-gathering theory. In this section, I discuss my theoretical approach to the data analysis, my findings from coding the interviews, and the themes that ultimately emerged from my analysis.

4.5.1 Data familiarization

Because I performed the interviews, I had a deep sense of familiarity with the data before I even began coding. However, I also performed a cursory initial assessment with NVivo's "auto code" feature to generate preliminary "start" codes to ground me in the data. Below is a word cloud of the start codes generated by NVivo.

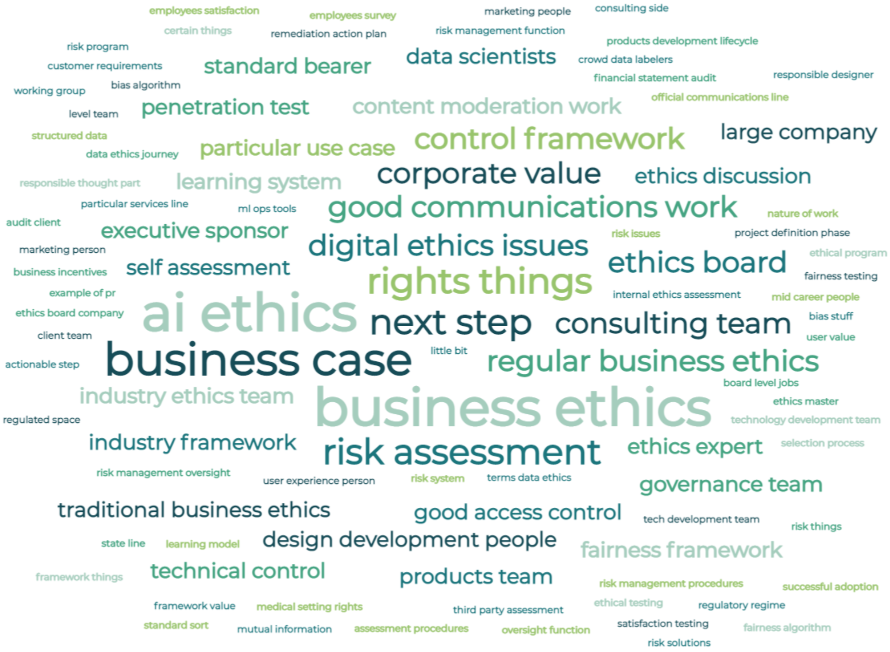


Figure 6. Initial codes from NVivo

The NVivo codes focused on the key topics of conversation, namely AI ethics and business ethics. However, my research question was more focused on the actions companies and responsible tech professionals were taking beyond just ethics and frameworks. While these codes gave me a good grounding for the more focused coding analysis I performed, I had to keep my more pointed research question in mind.

4.5.2 First and second rounds of coding

The term "open coding" was originally coined by grounded theory researchers in the 1960s, a practice of free form coding of data which Saldaña redefines as "eclectic coding" (Richards, 2020; Saldaña, 2021). Both terms refer to the process of coding data in qualitative studies using multiple methods in the first cycle of coding, with the goal to clarify codes in later cycles of coding. Importantly, this process is iterative (Given, 2008) – as you work through material, new categories emerge, and previous codes become clearer. After a first cycle approach using the eclectic coding techniques I mentioned earlier, I transitioned to further analysis by combing codes, beginning to identify emergent themes, and eliminating codes that were not central to my research questions. After trimming my code list, I examined the code landscape that I had created with another word cloud, this time using the codes I had generated in my first and second rounds of coding.

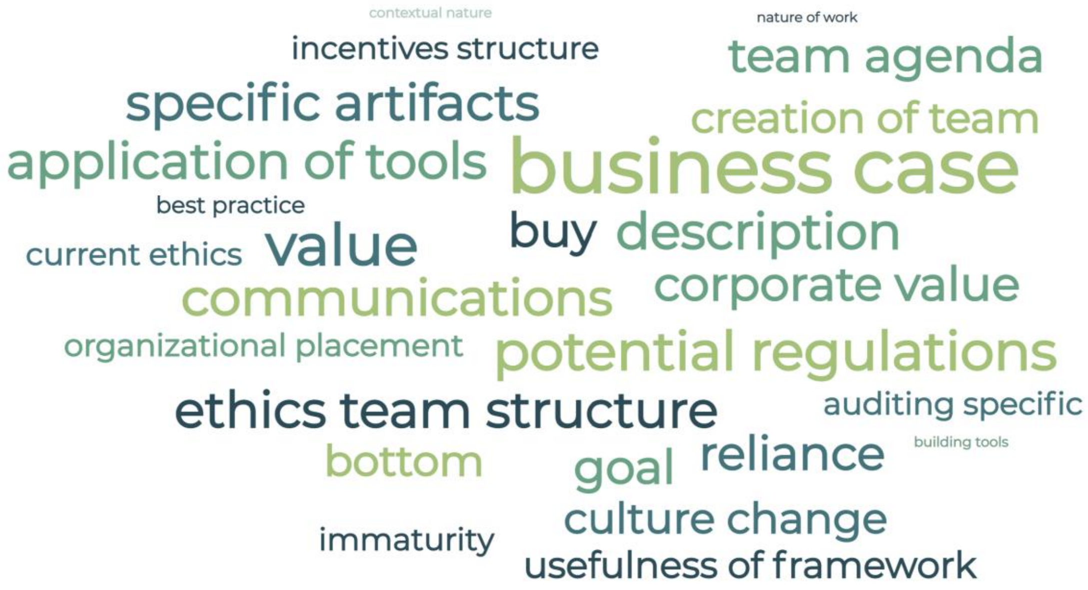


Figure 7. Focused codes generated after analysis

4.5.3 Generating themes from qualitative data

In the process of focusing my codes down, I began to see themes emerge about the process of operationalization of AI ethics, the common challenges faced, and the best practices or recommendations of the interviewees. Using the word cloud of codes, counts from NVivo, and pen and paper, I began to organize and arrange these themes into three categories: the process of operationalization, challenges, and recommendations for other responsible technology professionals.

I then returned to the original coded data to re-examine the ways my three emerging themes were reflected in the original transcripts (Braun and Clarke, 2006). The goal of this phase of analysis was to make sure I had kept the data in mind in creating themes (Guest, 2011). Throughout the process of refining my themes, I kept in mind the larger picture of how they all fit together. After several iterations of checking themes against data and reworking my theme diagram, I finalized my themes.

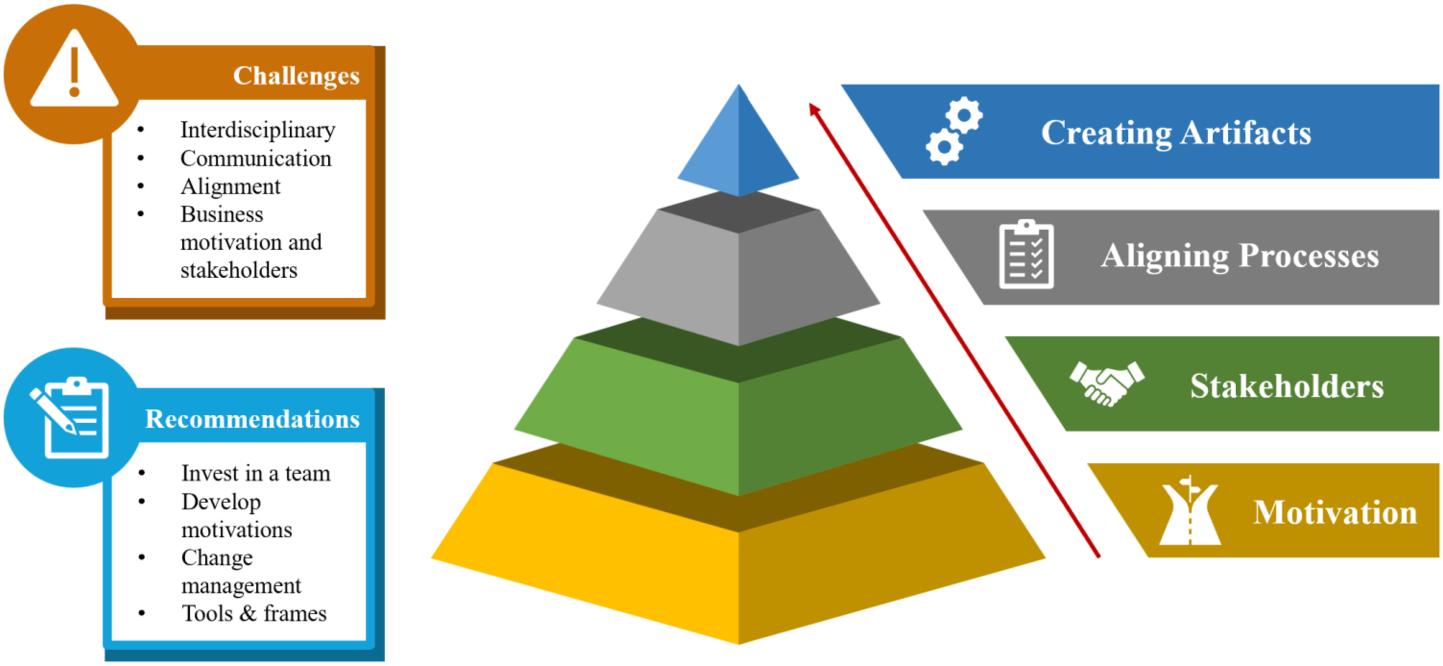


Figure 8. Overarching themes that emerged from qualitative analysis of interviews

4.6 Reporting findings of qualitative research

The final, and perhaps most important step, of qualitative research is reporting the findings. In the next chapter, I detail the findings from the interviews, including challenges and recommendations for responsible technology professionals.

5. Findings: operationalizing AI ethics involves business motivations, challenges, and creation of specific artifacts

5.1 Multiple factors incentive and motivate companies to work on AI ethics

Despite the current popularity of AI ethics, the field is still relatively immature. Most AI ethics boards, principles, and teams have been created within the last several years, so there are still challenges to implementing AI ethics work on the ground. In interviews with responsible tech professionals, I found that having a strong business case was an important shared quality for most responsible tech teams, that artifacts from these teams tend to focus either on governance or technical tooling, and that challenges to implementation often involve miscommunication or misalignment in culture and team goals.

5.2 Business case – the motivation to work on responsible tech

One of the critiques of industry AI ethics is that their profit-driven structure inherently precludes them from any real impact. However, there are many reasons a company would want to adopt responsible technology guidelines for its own benefit, even within a profit-driven structure. Interviewees in this study expressed multiple different motivating factors driving AI ethics implementation within organizations. Each of the business cases for AI ethics differed according to the organization's structure – for example, audit-based firms like EY and KPMG took a risk management approach to justify their work, while client-based firms like IBM and Intel were more likely to develop artifacts that they could use with clients. Companies start from their business case, and knowing their motivations inform the responsible technology artifacts they ultimately create.

Ultimately, all companies looking to operationalize AI ethics principles needs buy in from both executives and the company at large. To ensure that AI ethics work has buy in, teams need to align early with a business case grounded in the current structure of the organization.

It’s important to understand what drives this work, because it will not continue throughout the process of operationalization if it does not have both personal and institutional buy in.

The main motivations or business cases for AI ethics I found are:

1 Early pushes by AI ethics evangelists

Employees at large tech organizations recognize the importance of AI ethics – from the bottom up. At IBM, the trustworthy AI centre of excellence is “finding more of a groundswell, saying, ‘we’re really interested in this space, we want hands on experience, point me to a project that I can work on, because I want to learn everything I possibly can about this space, because I care.’” Large organizations like Microsoft, Intel, and Google have groups made of volunteers who work on responsible tech research without being compensated, solely because they care and want to drive the work.

However, as researcher Jessie Smith says, volunteer teams take a long time to develop team structure since responsible tech work is “probably not going to be a part of like a KPI for years after the motivation piece starts.” Most of the professionals interviewed agreed that early motivational pushes by employees needed to be matched by enthusiasm or buy-in from executives, like in the case of KPMG’s responsible AI team in the Netherlands, created by partner Sunder Klaus three years ago. Because of his position within the company, he was able to create the team. As Marc van der Meel from KPMG recalls, “long story short, it basically came out of his head, that [responsible tech] would be a future successful proposition. I think at that moment, it was still very much ahead of the curve…there wasn’t a lot of market demand yet.” Employees and executives can push for this work within their companies even before the market demands it. As one study from the Economist found, employees at large companies want to be doing this work, and to retain top talent, many companies have begun forming responsible tech teams under pressure from their own employees and executives (Haider and Islan, 2020).

2 Fiscal losses from bad AI PR

While early adopters and ethics champions may have begun the process of forming AI ethics teams, there are still many sceptics. As Ian Eisenberg from CredoAI says, "it's our impression right now, that people are not going to do that much work for ethical AI…most companies are not going to for number of reasons: they just don't have the capacity, they're immature, they don't know how to do it, and it's not actually clear that that would lead to any improvement in their bottom line."

To garner buy-in from sceptics, responsible tech teams need to show that their work saves time and money. One oft-cited example of a financial business case for AI ethics is avoiding the fallout from a public display of unfair AI. For example, when Volkswagen's software system was discovered to "cheat" on emissions testing, they lost $4.3B. Olivia Gameblin, an ethical AI consultant, says she uses that example to gain buy-in from sceptics. "Their stock price is still 35% lower today than it would have been if they didn't have that scandal…. everyone complains about how expensive it is to take a model out of production, retrain it, fix it, but you're going to put models into production that you're going to have to recall because you did not make these [ethical AI] decisions in the first place. If you'd already made these decisions, you wouldn't have to recall those models. So again, saving you time and money."

Fundamentally, technology systems require users, and keeping users satisfied requires a broad ecosystem of trust and positive coverage. Responsible tech professionals all seem to agree that harmful AI practices will eventually be revealed to the public. As Smith says, "even if you're you have no transparency practices, and you never tell anybody how things work, eventually people will discover that they've been harmed….and that's a PR nightmare for people who don't do AI ethics or who don't care about it. That's the business case."

3 Provide services or tools to clients

In addition to avoiding the losses caused by bad press, companies also have an incentive to capitalize on the wave of AI ethics popularity to actively sell work and develop their businesses. Companies that provide client services are especially motivated to wrap AI ethics offerings into their current practices. Firms like KPMG and EY that focus on providing audit and governance services have developed AI ethics team that work closely with those clients to provide AI ethics frameworks as a service. Ansgar Koene, the AI Ethics lead at EY, noted that "in order to be a consultant, auditor in this space, it's all about being the trust provider." Because EY's clients already expect audits and compliance services from them, their AI ethics work takes the form of audit and compliance artifacts.

At companies like IBM and Microsoft, however, their clients often expect technical solutions to AI ethics, in keeping with these companies' other suite of software solutions. Many large tech companies have created technical toolkits to address bias and fairness issues within AI – Microsoft's fair learn and IBM's Fairness360 are just two examples.

While frameworks have been criticized for not providing enough actual organizational change, jumping straight into tools and solutioning misses important organizational context and alignment. As Boinodiris says of her IBM clients, "we have clients out there who want the easy checkbox tooling…and we tell them, we try to tell them, we're not just using tools for our own organizational AI governance, because we know it's not enough."

4 Prepare for upcoming regulation

Since the General Data Protection Act (GDPR) went into effect in 2018, governments around the world have developed more comprehensive laws to govern the use of AI. The United States' proposed Algorithmic Accountability Act would require companies to conduct impact assessments of high-risk systems using AI, and the European Commission's proposed AI act would implement a legal obligation on developers of AI to meet standards of governance, design, transparency, and data security (MacCarthy, 2019).

In the interviews, I found a general consensus among interviewees about the AI industry moving into a space with more government regulation. While there are no "objective criteria" or industry standard currently, if the AI act and other regulation were to pass, they would start companies to use standardized governance tools. Even now, although the AI act has not been implemented, KPMG’s clients have begun to take the regulation into consideration to “future proof” their work.

5.3 Operationalizing AI ethics -implementing organizational artifacts

Making change in a company’s practices, projects, and deliverables requires action beyond theory and principles. For many organizations, the embedding of ethics into AI development and implementation involves change to governance, policy, and procedures. Many companies, like IBM and Intel, have created brand new teams, roles, and responsibilities that specifically address the operationalization of AI ethics principles. In conjunction with team creation, companies have developed AI ethics artifacts, like governance frameworks and process flows.

Just like the motivations for responsible tech work tie back to the organization’s original business case, so to do the artifacts. In our interviews, I found two main categories of AI ethics artifact: audit and governance process flows and technical toolkits. The deployment of these artifacts was most successful when they aligned with the tools and processes already being used by the company.

1 Workshops and organizational alignment processes

One AI governance lead discussed the usefulness of their workshop methodologies, citing a simple checklist application that asks internal stakeholders a series of questions about their processes. They described it as a form of “triage,” a quick way to align with the people working on AI and discover the vulnerabilities of their work quickly. Another discussed their work building “accelerator” programs to help guide people to make better decisions about developing and deploying AI.

Boinodiris describes their compares this early stage of alignment as the first few chapters of a book, saying “think about organizational AI governance as a book with chapters. This is a World Economic Forum case study for IBM - when we started developing AI, we didn't have principles, we didn't have an ethics board. We made mistakes…. we learned. And if you think about the book, we say to clients, chapter one has to be establish principles based on your values. Chapter two, start to focus on AI literacy with respect to ethics, chapter three, get your ethics board established so it's got power. Chapter, like seven, is to start to use tooling."

Responsible technology professionals agreed about the importance of setting up the right channels of communication and responsibility and drawing clear lines between AI ethics work and the work already done by existing teams. Artifacts in this initial stage can simply be a process workflow, an Agile team plan, or an ML Ops integration. Ethics consultants like Gambelin say that creating workflows for stakeholders, like data science teams, is a crucial part of their process.

Even at Credo AI, which focuses explicitly on technical metrics and toolkits, the first step in their process is a governance alignment phase, where they "try to figure out the governance approaches and map that to a particular use case, the alignment step of the process." These artifacts often take the form of more traditional consulting outcomes, like in the case of KPMG's work with government clients, where they "help set up an ethical compass, which consists out of twelve norms or values, which they as an organization want to adhere to…help them with setting up an ethical committee, so people within the organization that are qualified to judge an AI system…and we also created a training program for them for their employees to become more familiar with the topic of data ethics."

2 Audit frameworks deliver trust as a service

While alignment artifacts generally fit into existing business practices and processes, AI audit and governance frameworks are an artifact emerging entirely from the new AI ethics space. In companies that describe themselves as selling "trust as a service", like KPMG and EY, their teams have developed structured ways of assessing and auditing AI systems against their high-level principles of fairness, trust, and transparency.

One team lead discussed their impact assessment, which generates a heat map of the major issues AI teams and projects need to focus on. Projects are scored against principles relevant to the work, like accessibility or bias. For example, a resume screening service like Amazon's that was biased against women would be considered actively harmful, and would get a negative score, whereas a system that had been tested for bias could get a positive score. Importantly for this framework, projects cannot receive a zero on the scale, because, as the team lead said, "I don't think there's such thing as neutral technology." When a project gets a negative score, it receives intervention to try and move it from a negative, harmful impact score to a positive, where there are no foreseeable negative outcomes.

Other companies use forms that are less formalized and designed more as tools to spark critical reflection among stakeholders. For example, one such form discussed in the interviews asks different questions to developers depending on their current phase of AI development. This format asks the project managers about what they're doing and asks for them to reflect on ways that it could be harmful to end users or counter to high-level responsible AI principles. As one of the developers of that form said, "we calibrated what the different areas were, and then we ask, who does this, what are the [responsible technology] issues, and explain what you're doing about it."

Audit based frameworks allow clients and technology producers to measure AI outputs, assess their risk, and provide ways for developers to reflect on the principles of the technology during the development process. Other examples of this kind of artifact include Google's model cards for model reporting (Mitchell, et al., 2019), which many interviewees cited as a gold standard for AI assessments.

3 Technical toolkits allow for model management

Technical toolkits for addressing AI bias and transparency issues have been highly visible artifacts, with many technology companies finding technical ways to fix the harm created by their AI systems (Durmus, 2021). For example, Credo AI formalizes the AI assessment process with their tool called Lens, which collects evidence about AI models and records performance-based metrics. After aligning on compliance controls, Lens allows companies to record and store evidence about those controls. For example, recording a metric about model fairness, which the Lens tool can record, upload, and automate to make that assessment process faster and more scalable.

Toolkits like this are important, and part of the broader and holistic solution to issues of responsible technology. However, even companies that use and deploy tools, like IBM, recognize that tools alone cannot address the whole of AI ethics. Boinodiris says, “the tools are useful. They just can't be used alone, without the design thinking, the right intent and ethics boards and principles, and all this other stuff, AI literacy, training practitioners, so much you have to do to get it right.” Other interviewees echoed this sentiment, including Credo AI’s Eisenberg, who said “automated assessments and checks are scalable, and therefore, very helpful, but they don't cover all things that people want from an ethical assessment, and they potentially undermine the more difficult work that's needed to think through the impacts or harms that are caused. It can get you into, ‘Oh, we set up our monitoring we're good.”

5.4 Challenges to operationalization

AI is a term used for many different applications of technology, and its very breadth presents a huge challenge to developing organizational AI ethics teams. Below are four of the main challenges to AI ethics implementation that came up in our interviews:

1 Balancing organizational tensions

Many of the aspects of responsible tech require a balance of opposing forces. The driving force for creating AI ethics teams needs to be a combination of top-down executive buy-in and a bottom-up culture of ethics. Frameworks need to be specific enough to provide guidance, while vague enough to allow for contextual re-interpretation. Tying ethics to product teams allows for faster implementation but can cause an over-reliance on profit. These organizational tensions need to be balanced by responsible technology professionals, and a successful balancing act requires knowledge of the company’s business case, existing toolkits, and team structure.

2 Aligning interdisciplinary stakeholders

Responsible technology work involves many different existing aspects of business. Our interviewees cited important collaborators in legal, compliance, research, education, marketing, HR, PR, governance, and policy teams that already existed in their respective organizations. AI ethics teams are new and need to find their place within this corporate ecosystem.

3 Immaturity of AI ethics space – lack of formalized structure and standards

While there seems to be an emerging AI legal regime, currently there are no formalized, industry-accepted standards for AI governance and assessment. Interviewees made a number of comparisons to other industry standards that AI ethics work could imitate, like environment, social, and governance (ESG) standards, which have been incorporated into legal compliance regimes, SOC2, a voluntary compliance standard for cybersecurity and data privacy, or Control Objectives for Information and Related Technologies (COBIT), a framework created by to bridge crucial gap between technical issues, business risks and control requirements. Despite these comparisons to standard comparisons, nothing equivalent has yet emerged in the AI ethics space. As AI ethics consultant Gameblin said, "there's nothing to certify against, so it's more of question of who wins the certification race to become the standard certification."

4 Reliance on both frameworks and tools deployment

A 2020 paper analysing 27 AI ethics toolkits found that approaching AI ethics from a tool-based perspective frames the work as purely technical and avoids contending with AI's political and organizational effects (Hagendorff, 2020). Responsible technology professionals also need to navigate organizational power dynamics and align with stakeholders, which tools alone cannot help with.

> Frameworks are important to inform the organizational alignment of AI ethics work but need to be supplemented by tools that can scale. However, tools alone fail to account for the contextual nature of each implementation of AI technology.

5.5 Recommendations for responsible tech professionals

Frameworks are an important first step in a longer process of change management and operationalization of AI ethics withing organizations. While frameworks fail when they exist alone, jumping right to solutions and implantation of tools without the necessary organizational alignment provided by frameworks also leads to a failure of responsible tech systems.

My interviewees discussed the current state of AI ethics work in industry, but they also had normative recommendations. Many of them discussed what worked, what didn't work, and provided tentative recommendations for AI ethics teams in the future looking to operationalize their work in an industry context. In analysing the data from the interviews, I saw five general recommendations emerge from all my interviewees.

1 Invest in a dedicated AI ethics team

Any company trying to do AI ethics work needs a dedicated team. Companies should start with their own values, whether in the form of an existing business case or set of business ethics values that can be leveraged to generate buy in from their organization. Articulating frameworks that link to core company values is an excellent place to start the process of organizational responsible tech work, and to implement these frameworks. In the IBM case study on AI ethics, the process of implementing an ethics team is described as chapters. Chapter one – define principles. Chapter two – focus on AI literacy and cultural-buy in. Chapter three – establish an AI ethics board and dedicated team. Companies looking to implement AI ethics should follow w similar step-by-step process of integrating a responsible tech team into their organization.

A dedicated AI ethics team will need to fit within existing business structures where they can be most effective. For some organizations, like KPMG, this is within their assurance and audit teams. Others, like IBM, have their AI ethics team focused on client outputs. Importantly, companies must recognize that the creation of a new, dedicated AI ethics team will require an investment in training and educating responsible tech professionals. This initial investment in people and processes will pay off – according to Ansgar Koene of EY, "once the AI act comes into operation, there will be a severe shortage of people who can actually support compliance with the with the act.

2 Leverage cultural and business motivations to generate organizational-wide buy-in

Although values and a dedicated ethics team are a promising start to AI ethics implementation, their work needs organizational buy-in to progress. Individual actors throughout the company can be targeted with education, training, and consistent messaging about the importance of AI principles.

To quote researcher Jessie Smith, "a key step is motivation. Somebody has to be motivated to convince everyone else that this is important. People will be business minded by default, so there has to be somebody who comes in and says that ethics is important." Just the process of developing values and a team requires some amount of motivation, and actively building upon this foundation is crucial to continuous AI ethics work (Raji, et al., 2020).

Artifacts like workshops and trainings become important to cultural buy in, like IBM's focus on educational training on AI ethics. Boinodiris says, "it's three hours per day, where it's hyper personalized, people are making a playbook, they're very getting very excited…. The primary for purpose is get people to care. Get people to care and recognize how this has meaning for them in where they sit within the organization. The second thing this does, is it gets people to raise their hand and say, 'I want to be part of the next phase of design thinking workshops, where we start to establish principles and governance, including being a part of an ethics board.' It's a way to start mining for stakeholders across the bigger org."

Cultural buy-in can garner individual support, and AI ethics need to leverage the business case to gain buy-in from the company at large. As one governance professional said, “companies that are more mature in this space are thinking beyond regulations. They're also thinking, ‘what kind of harms could come out if I do the wrong thing?’ There's reputation damage. There's potential like operational risk outcomes.” Leveraging the real business case for AI ethics allows teams to generate more motivation for this work.

3 Align stakeholders by addressing their context and lowering barriers to entry

Importantly, AI ethics teams do not need to reinvent the entire organizational structure. They should clearly define what their work focuses on, and what business processes can be owned by other teams. Once companies have developed their values and formed their AI ethics teams, they should identify important stakeholders and separate out the work.

Crucially, AI ethics teams should lower the barrier to entry for their stakeholders as much as they can. If they want volunteers to work with their group, they should adjust their monthly quotas or bottom-line requirements, freeing up individuals to prioritize AI ethics work. Additionally, they should automate what metrics they can, and scale up appropriate processes.

4 Use existing practices of change management

Much AI ethics work has little to do with AI. Setting up clear channels of communication, outlining responsibility for new AI ethics teams, and setting an agenda with key stakeholders are all parts of organizational change management practices. Companies should leverage existing business channels when possible, refocusing existing resources on emerging tech ethics teams. AI ethics tools and governance artifacts exist within the larger ecosystem of the business case motivation, existing organizational structure, and team incentives – all of these factors can be addressed by conventional change management practices.

5 Embrace the holistic nature of responsible tech work – both frameworks and tools are unhelpful alone

Tools and frameworks are both unhelpful alone. Companies are rightly criticized for having a high-level ethical framework about being transparent, unbiased, and fair, but not actually doing any harm mitigation work. However, companies also cannot jump right into applying technical tooling to a complex, interdisciplinary, multi-stakeholder issue. The organizational process of alignment required to even write frameworks is incredibly useful and grounding when those same companies begin to implement AI fairness tools and practices. Neither works without the other – there are no quick and easy answers to generating responsible technology.

6. Discussion of findings and suggested future work

6.1 Limitations of Study

Of course, as with all research, this study has limitations. Because of the relatively small number of interviews, the findings of the study only represent a part of the much larger AI ethics ecosystem. The interviewees were generally part of larger organizations and were all located in Europe or the United States; therefore, the study is unrepresentative of the AI ethics development in smaller companies and non-western countries. Although I have attempted to augment these biases with existing literature and my own experiences as a researcher, future work should attempt to cultivate a more diverse and representative sample of practitioners.

6.2 Conclusion

As AI has become a more powerful force in our world, it has become clear that governance and ethical design will be necessary to avoid AI harms. While companies cannot do everything on the AI ethics agenda, many companies have the business motivation to substantially reduce harm and implement responsible technology. This paper attempted to show the current state of AI ethics work within these industry actors and provide recommendations for policymakers, industry leaders, and tech workers to implement organizational and structural AI ethics work.

Taking an ethnographic approach, I provided qualitative support for the current best practices and future recommendations for industry applications of AI ethics. Industry actors have already begun to form an AI ethics ecosystem and internal apparatus for governance. Any government policy attempting to ground this ecosystem with further laws and regulations should take these current best practices into account. Ultimately, my findings confirmed that AI ethics is ultimately contextual, dependent on an organization's business case, stakeholders, challenges, and outputs. Frameworks themselves are artifacts of this organizational context, reflecting the values and norms of organizations. The more concrete artifacts of AI ethics, like governance tools and technical toolkits, emerge from the same organizational context of frameworks, and rely on stakeholder alignment, communication, and buy-in from organizational leaders.

Appendices

Appendix A: Semi-structured interview questions

Semi Structured Interview plan:

1. Who & What of "responsible tech" and tools
   1. Who is responsible for advancing this team's agenda?
   2. What's the main driving force of the work organizationally? For example, at IBM we had a combination of bottom-up push from developers with some top-down support from executives. What's it like on your team?
      1. What business unit does the team sit under?
         1. Who "sponsors" this work? Is there an executive who advocated for it?
   3. **What is the team's main agenda?** Education, product development, research, etc.
      1. How is this shaped by the organizational location of the team?
         1. What do you want the team's agenda to be? How is that different from its current state?
         2. How do you see the team's agenda developing?
   4. **What tools and outputs does the team focus on?**
   5. **How are the outputs of the team structured? As products, thought leadership, etc.**
      1. Again, how is this shaped by the organizational location of the team? Focus specifically on artifacts! **Any small details on deliverables. Follow up on this question for details!**
   6. Who are the main stakeholders in the team's work? CXOs or lower-level employees, customers or clients?

Daily work and responsibilities

a. **Tell me about your day-to-day responsibilities, stakeholders, tasks, outputs, etc.** Warm up question - emphasis this should be a BRIEF response

Placement within larger business case

a. **What IS the business case that drives your teams' work? How does it fit into a larger organizational mission?**

a. **How do you feel the business case of your company has shaped your team's work and agenda?**

b. How does having a business case constrain the work of responsible tech?

c. How do you feel the business case is expanding or changing? Is there anything you are doing to change it? Anything others can do?

d. Goal of this section is to specifically pry into the economic drivers of the new wave of AI - is it insight focused? Product based? What motivates the companies and the business case?

e. Also, do professionals agree on the economic factors?

Moving beyond frameworks - **What are the frameworks you engage with?**

. Where do frameworks emerge from industry itself and not just high-level ethics boards?

a. What is the nature of the frameworks? Regulatory, educational, aspirational?

b. How does the process of certain values appearing and being negotiated into practice work within your organization?

c. How have frameworks shaped the outputs of your team?

Policy Recommendations

a. Normatively, how would you recommend moving ethics beyond frameworks? What has worked for your team and what do you want to see that isn't being done yet?

How does industry interact with the other actors / social worlds - government, academic, think tanks, etc?

. Who are the most important stakeholders? How is their input accounted for within your company?

Appendix B: Project Diary

PROJECT DIARY

Placement Based Dissertation

Picture Caption: View from University of Edinburgh's campus, where much of my early thesis work was done

Beginnings
May 9th - 22nd

Main lesson:

Just get started

In the first two weeks of the dissertation, my biggest lesson was just to start somewhere. By picking one area to start working on, my longer-term plan eventually fell into place.

Getting started

The challenges of finding somewhere to begin

Ever since I learned it was an option, I badly wanted to secure a placement for my dissertation. Coming from a corporate job at the IBM AI ethics board, I knew that I wanted to return to industry after the completion of my master's degree, and the placement would be a valuable way to make connections and pad my resume. After some last-minute changes, I was lucky enough to partner with the research group Data & Society on my thesis on moving AI ethics beyond frameworks.

Even before the placement began, I chatted with Jake Metcalf, my supervisor from Data & Society, about the way my work would fit into their larger initiatives.

RESEARCH TRACK

AI ON THE GROUND

We use social science to develop robust analyses of AI systems; effectively assess the impact of AI systems; and inform future design, use, and governance.

Picture Caption: Data & Society's webpage on their AI on the Ground initiative

My project aligns with their team's AI on the Ground initiative, ethnographic research project analysing the ways current industry actors are using AI ethics. This years-long project's goal is to understand AI's risks and potential harms to people, communities, and institutions through empirically grounded research rather than speculation. Data and Society's website says, "We argue that the impact of AI systems and policies can only be fully understood by observing, listening, and speaking with people *on the ground*–from government and business leaders to scientists and engineers, to community activists and vulnerable groups."

To support this research, I needed to schedule interviews, structure the outline of the dissertation, develop a methodology for the interview process itself, and align the entire scope of the project with Data & Society's current projects. I felt overwhelmed.

First, I focused on scheduling interviews with responsible tech professionals. The goal of the thesis is to provide some qualitative evidence on industry professionals in responsible tech fields, so actually gathering that qualitative data seemed like a good place to start.

In the first two weeks, I reached out to about a dozen individuals, with only a few responses back. The team at Data & Society were very helpful in connecting me with people to interview. One of my personal goals is to extend my interviews across a wide and diverse spectrum of experiences, not just my own network. The Data & Society team helped with that, as well as some of my professors and academic contacts within the University.

In preparing for interviews, I developed a methodology and questionnaire for my semi-structured interviews. Last semester, I took a class on analysing qualitative data, and I was able to use the textbooks and materials from that class to develop a methodology based on the literature.

My goal for the interviews was to provide enough structure in the questions to answer a few main questions, including:

- How do you feel the business case of your company has shaped your team's work and agenda?
- How does the process of certain values appearing and being negotiated into practice work within your organization?
- Normatively, how would you recommend moving ethics beyond frameworks? What has worked for your team and what do you want to see that isn't being done yet?

I also clarified the scope of the secondary research I'll be pursuing. I will be reading blog posts, ethics frameworks, corporate websites, etc. - the purpose of this secondary research will be to assess the outward facing perception of what a responsible tech advocate's daily work is.

The main challenge of this beginning stage of the dissertation was the ambiguity of getting started. When I began, I only had my research plan and my own theories about AI ethics in industry. However, I needed structure, and a long-term plan. Luckily, I was able to have several calls with Manny Moss from Data & Society about their AI on the Ground Initiative.

Manny was able to offer me guidance on how to structure the research questions for my interviews. At the end of the first two weeks of the placement, I had a better idea of what Data & Society's current projects are within their AI on the Ground Initiative, a finalized methodology for my interviews, and 3 interviews scheduled. In the next phase of the project, I planned to start writing my introduction and schedule more interviews.

During this initial phase of work on my placement-based dissertation, I had to learn how to manage my own time and expectations. While the team from Data & Society were helpful in setting expectations and providing feedback on my ideas, they were hands-off in the logistical work of setting deadlines and recruiting participants. This was difficult for me both practically and intellectually. Practically, I needed to set myself deadlines, manage my own time, and send out emails to all the participants myself. More difficult, however, were the intellectual hurdles I had to overcome to start the thesis. How do I move from my research proposal document to a full-fledged thesis? Where do I even start? Without someone to lay it out for me, I found myself reading online articles with desperate-sounding titles like “How to write a master’s dissertation: a step-by-step guide.” After some reading, thinking, and advice from fellow grad students, I realized that the best way to start was to just start! So, I began with a thorough reading of every article I could find that seemed aligned with my thesis topic. When I started a deep dive into the literature, ideas and connections began popping up in my brain. Soon, I had the rough outlines for a literature review, with more ideas in the back of my draft document. Although the prospect of writing over 10,000 more words is daunting, it started to feel less overwhelming.

Picture Caption: The Data & Society conference team at a picnic on the river Isar in Munich

Connections
May 23rd – June 5th

Main lesson:

The importance of collaboration

Spending a weekend with the team from Data & Society was an incredible opportunity to collaborate on my ideas, theories, and structure – and it was a lot of fun!

Meeting the team

At a conference in Munich, I met my collaborators

One of the reasons I was excited to do a placement-based dissertation was the networking opportunities afforded by working with an organization. In the 3rd week of my placement, I was invited to attend just such an event.

On May 27th, I travelled to Munich for a conference with Data & Society on sociotechnical AI governance systems. I was able to meet the team that I had been working with on my placement and listen to an entire day's worth of talks and discussion about responsible tech, safety, and sociotechnical systems. Jake Metcalf, my placement supervisor, opened the conference by discussing the work that needs to be done to define the term "sociotechnical" – involving hands-on ethnography work like my thesis. In following discussions, I was able to discuss my projects with academics and professionals, who gave me feedback on the framing of my project.

I met some fellow researchers from the Ada Lovelace institute and the Technical University of Munich who were interested in my project and wanted me to send them the link to my completed thesis research. Meeting them felt like a valuable addition to my network of academics and professionals working on responsible tech. Overall, it was an incredibly generative and helpful experience for a young researcher – exactly the type of opportunity that I had hoped to have during the placement!

I also began working on the academic text of my thesis in earnest. With the help of some of the discussions with the Data & Society team in Munich, I finished outlining my entire project.

This was a helpful academic experience, since I have learned that the process of writing the thesis is a lot less about cranking out words to put on the page, and a lot more about being thoughtful with the types of analyses and citations used.

In addition to finalizing my dissertation outline during this two-week period, I also finished a first draft of my first chapter and started outlining the second chapter. At the end of week 4, I had completed 4 interviews as part of my ethnographic research process and have started reaching out to more professionals to interview as well.

Overall, meeting the team in Munich and finalizing the outline for the project helped me realize the scope and goals of the thesis in more detail. The biggest lesson from this period of the dissertation work has been the value of collaboration. As I mentioned in my last project diary, my supervisor from Data & Society, while helpful, has been mostly hands-off throughout the placement, so many of the decisions about the project and structure have fallen to me. However, getting the chance to connect with the team and with other academic collaborators at the Munich conference afforded an incredible opportunity to share ideas, restructure the project, and get feedback. After I left Munich, I had a working structure for the thesis, which was incredibly valuable as I continued to research and write the thesis.

In my last project entry, I wrote about how I struggled to find a place to begin. The collaboration in Munich also showed me the value of having a structured and intellectually disciplined approach to research. Having this feedback allowed me to tighten up the structure of the dissertation itself, which helped immensely in writing and researching it. Once I knew the general flow of all my arguments, it was much easier for me to sit down and write. In the table below, I have included the final structure and flow of my dissertation.

Dissertation structure for “Moving Ethics Beyond Frameworks”

Chapter	Title	Description	Purpose
1	21st century AI as a new artifact	How artificial intelligence has become a sociotechnical artifact entwined with industry	Providing an academic backdrop for my work – why focusing on industry AI ethics is important
2	The gap in AI ethics literature	Current AI Ethics critiques centre industry, but fail to provide an ethnographic view of the inner workings of companies	Showing that my interviews address a real gap in the current literature
3	Methodology of this study	The goal and methodology of this study is to fill that ethnographic gap	Discussing my thinking and methodological choices
4	How to move ethics beyond frameworks	Findings from interviews	Sharing what I’ve learned
5	Policy recommendations for responsible tech professionals	Policy Recommendations	Promoting an action-oriented output

Picture Caption: Schönbrunn palace, which I visited this summer and which I write about in my thesis as the location of the first show of the Mechanical Turk in 1770

Logistics
June 6th – 20th

Main lesson:

Logistical work is real work

Academic work isn't just reading papers and writing more words – it involves difficult days of long LinkedIn searches. Logistical work is valid as a type of work!

Getting the work done

Dissertation work takes on many forms

With a finalized outline, I spent much of this period focusing on interviewing participants for my study. My target participant is a responsible technology professional who works for a large industry actor. Since I'm examining the ways that AI ethics frameworks are enacted within organizations, I've been looking for individuals who work at companies that have released AI ethics frameworks – IBM, Google, KPMG, etc.

I've spent quite a bit of time sending out emails – since the response rate for cold emails like this is low, I've sent out probably 30-40 emails in this two-week window. Also, since I'm not looking for just any participants, I've done a lot of research trying to find appropriate interviewees. That translates into a lot of time spent on LinkedIn!

Although I had a plan and process laid out before my interviews, by the nature of semi-structured interviews, each conversation grew and adapted in its own way. I learned not to let my interviewees talk for too long, to cut in and get to the root of what they were saying. I wanted to make the most of my time with them to figure out what was *really* going on within these companies, not just the surface-level PR lines they feed to everyone. My abilities as an interviewer got sharper and more thoughtful with each interview I did. I also started to see some connections between the findings from each interview – something that one person from a consulting company said something different than someone from an auditing company, but they agreed on some of the underlying challenges of the responsible AI space. Throughout the interview period of the thesis, I felt more and more confident in my own abilities as a qualitative researcher.

While my focus for these past two weeks has been the qualitative data collection portion of the thesis, I've continued to work on the written parts of the thesis. I've got a rough draft of my first two chapters finished and started working on the methodology section. Although I don't like leaving the first two chapters unfinished, I wanted to make sure to record the information on the methodology of the study while the information was still fresh in my mind – in other words, while I was still actually doing the data collection!

Again, having taken a class last semester about qualitative data analysis was extremely helpful in the framing of my third chapter on methodology. I've been able to reference several textbooks on qualitative analysis to provide much-needed structure to my data collection and analysis approach. As I learned in that class, the process of data collection and storage are already important for the process of analysis. Even though I haven't started the actual data analysis yet, I have written about the ways that I structured the interviews, the theories behind my choices, the ways I processed and stored the data, and more details on my process of qualitative data collection and preliminary analysis.

The biggest lesson I had to learn in this period is that dealing with the logistics of the dissertation IS work! Academic research requires a lot of background logistical work too. LinkedIn searches, transcription, etc. was all just as important!

Picture Caption: My dining room table, where many hours have been spent on dissertation work

Weaving
June 21st – July 4th

Main lesson:

Letting go of my overly ambitious plans

Approaching the end of the placement, I had to realize that having a smaller scope of work than I originally planned for is okay and part of the research process.

Pulling threads together

Accepting the reality limited scope for master's work

My first two thesis chapters are essentially a long STS essay about the history of AI, and the gap in current literature around AI ethics. Essentially, the first 5,000 words are a long persuasive piece convincing the reader that the topic of my thesis is important, grounded in past research, and addresses a real gap in the current literature. The next chapter jumps to methodology. Again, my class last semester on qualitative data analysis was crucial in the formation of this chapter. My fourth chapter, and perhaps most substantial one, involves the analysis of the data I've collected during my interviews. Although I have started preliminary data analysis, I've been holding off on really diving into that chapter before I finish all my interviews. Therefore, for the last two weeks, I've been bouncing back and forth between academic writing, qualitative analysis, and the logistical challenge of trying to get more people to interview with me.

This last task has been among the most challenging – I severely underestimated how difficult it would be to piece together a large enough sample of interviewees. I currently have 5 interviews completed, which is only half of my original goal of 10 interviews for the study. Although I've probably emailed 100 people, I haven't been able to land the number of interviews I wanted.

Although I still have some time before I need to close the interview window officially (my target date is July 8th), my difficulties in obtaining interviews have brought me to another lesson – accepting the smaller scope of a master's project. My academic thesis advisor warned me back in May that master's theses are supposed to challenge new researchers, teach us new skills, and add some value to the academic literature, but that the scope of a master's project is necessarily smaller given the time and resource constraints. Of course, I didn't really internalize that lesson, feeling ambitious as I was at the beginning of the summer. Now, I realize that just having a smaller scope of study and a few less interviews doesn't make my work worthless or bad.

As the project evolves and starts to fill out into its final shape, I can see that my original aspirations for the dissertation output were a bit ambitious.

The final big lesson that I learned in this last segment of my placement was not to be disappointed when my final scope falls somewhat short of my earliest ambitions! The project is coming together well, I feel good about my academic research, and the interviews I do have provide valuable insights. Even if the final result isn't ground-breaking, it's still something to be proud of.

Bibliography

Acemoglu, D., 2021. *Harms of AI* (No. w29247). National Bureau of Economic Research.

Adams, W., 2015. Conducting Semi-Structured Interviews. In: K. Newcomer, H. Hatry and J. Wholey, ed., *Handbook of Practical Program Evaluation*, 4th ed. [online] Hoboken, New Jersey: John Wiley & Sons, pp.492-506. Available at: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119171386> [Accessed 1 June 2022].

Allen, M. ed., 2017. *The SAGE Encyclopaedia of Communication Research Methods*. SAGE publications.

Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Van Esesn, B.C., Awwal, A.A.S. and Asari, V.K., 2018. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*.

Anton, E., Behne, A. and Teuteberg, F., 2020. The Humans Behind Artificial Intelligence–An Operationalisation of AI Competencies.

Ashok, M., Madan, R., Joha, A. and Sivarajah, U., 2022. Ethical framework for Artificial Intelligence and Digital technologies. *International Journal of Information Management*, *62*, p.102433. Available at: <https://www.sciencedirect.com/science/article/pii/S0268401221001262?casa_token=v12xM7Vy3MYAAAAA:AQJKnwJatcdpdaxO9BAJ7U92ER2ikkmgslq22rSJgK8QPRK_1tAC5Me7d1fqy1MExar8gSW55w>

Ball, K., 2019. Review of Zuboff's The age of surveillance capitalism. *surveillance & society*, *17*(1/2), pp.252-256.

Bansal, V., 2020. *The Deep History of Deep Learning*. [online] Medium. Available at: <https://towardsdatascience.com/the-deep-history-of-deep-learning-3bebeb810fb2> [Accessed 1 July 2022].

Barnes, T.J., 2013. Big data, little history. *Dialogues in Human Geography*, *3*(3), pp.297-302.

Bogost, I., 2017. *'Artificial Intelligence' Has Become Meaningless*. [online] The Atlantic. Available at: <https://www.theatlantic.com/technology/archive/2017/03/what-is-artificial-intelligence/518547/> [Accessed 18 June 2022].

Borenstein, J., Grodzinsky, F., Howard, A., Miller, K. and Wolf, M., 2021. AI Ethics: A Long History and a Recent Burst of Attention. *Computer*, 54(1), pp.96-102.

Braun, V. and Clarke, V., 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), pp.77-101.

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

Burt, A., 2020. *Ethical Frameworks for AI Aren't Enough*. [online] Harvard Business Review. Available at: <https://hbr.org/2020/11/ethical-frameworks-for-ai-arent-enough> [Accessed 29 July 2022].

Campbell, M., Hoane Jr, A.J. and Hsu, F.H., 2002. Deep blue. *Artificial intelligence*, *134*(1-2), pp.57-83.

Canca, C., 2020. Operationalizing AI ethics principles. *Communications of the ACM*, *63*(12), pp.18-21.

Cheney-Lippold, J., 2011. A new algorithmic identity: Soft biopolitics and the modulation of control. *Theory, Culture & Society*, *28*(6), pp.164-181.

Christoforaki, M. and Beyan, O., 2022. AI Ethics—A Bird's Eye View. *Applied Sciences*, *12*(9), p.4130.

Cordeschi, R., 2007. AI turns fifty: revisiting its origins. *Applied Artificial Intelligence*, *21*(4-5), pp.259-279.

Coughlan, S., 2020. *A-levels and GCSEs: Boris Johnson blames 'mutant algorithm' for exam fiasco*. [online] BBC News. Available at: <https://www.bbc.co.uk/news/education-53923279> [Accessed 31 July 2022].

Creswell, J.W. and Poth, C.N., 2016. *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Cukier, K. and Mayer-Schoenberger, V., 2013. The rise of big data: How it's changing the way we think about the world. *Foreign Aff.*, *92*, p.28.

DARPA.mil. 2018. *DARPA Announces $2 Billion Campaign to Develop Next Wave of AI Technologies*. [online] Available at: <https://www.darpa.mil/news-events/2018-09-07> [Accessed 29 July 2022].

Darrah, K., Mehta, B. and Mousavizadeh, A., 2022. *AI Boom Time*. [online] AI Global Index. Available at: <https://www.tortoisemedia.com/2021/12/02/ai-boom-time/> [Accessed 29 July 2022].

Data & Society. 2022. *AI on the Ground*. [online] Available at: <https://datasociety.net/research/ai-on-the-ground/> [Accessed 29 July 2022].

DeBrusk, C., 2018. *The Risk of Machine-Learning Bias (and How to Prevent It)*. [online] MIT Sloan Management Review. Available at: <https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/> [Accessed 31 July 2022].

Dormehl, L., 2019. *Revisiting the rise of A.I.: How far has artificial intelligence come since 2010? | Digital Trends*. [online] Digital Trends. Available at: <https://www.digitaltrends.com/cool-tech/biggest-ai-advances-of-the-2010s/> [Accessed 6 June 2022].

Douetteau, F., 2020. *The Tyranny of Appearances in AI*. [online] Blog.dataiku.com. Available at: <https://blog.dataiku.com/the-tyranny-of-appearances-in-ai> [Accessed 16 June 2022].

Durmus, M., 2021. *A Brief Overview of Some Ethical-AI Toolkits*. [online] Medium. Available at: <https://medium.com/nerd-for-tech/an-brief-overview-of-some-ethical-ai-toolkits-712afe9f3b3a> [Accessed 29 July 2022].

Ebell, C., Baeza-Yates, R., Benjamins, R., Cai, H., Coeckelbergh, M., Duarte, T., Hickok, M., Jacquet, A., Kim, A., Krijger, J. and MacIntyre, J., 2021. Towards intellectual freedom in an AI Ethics Global Community. *AI and Ethics*, *1*(2), pp.131-138.

Ehsan, U., Wintersberger, P., Liao, Q.V., Mara, M., Streit, M., Wachter, S., Riener, A. and Riedl, M.O., 2021, May. Operationalizing human-centered perspectives in explainable AI. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-6).

Enterprise Big Data Framework. 2019. *Where does 'Big Data' come from?* [online] Available at: <https://www.bigdataframework.org/short-history-of-big-data/> [Accessed 15 July 2022].

Ferreira, A. 2022. The age of surveillance capitalism: the fight for a human future at the new frontier of power, by Shoshana Zuboff. *Journal of Urban Affairs, 44(2),* pp. 274-276. DOI: 10.1080/07352166.2021.1939586

Firth-Butterfield, K. and Madzou, L., 2020. *Rethinking risk and compliance for the Age of AI*. [online] World Economic Forum. Available at: <https://www.weforum.org/agenda/2020/09/rethinking-risk-management-and-compliance-age-of-ai-artificial-intelligence/> [Accessed 29 July 2022].

Floridi, L., 2016. Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), p.20160112.

Floridi, L., 2020. AI and its new winter: From myths to realities. *Philosophy & Technology*, *33*(1), pp.1-3.

Foster, J.B. and McChesney, R.W., 2014. Surveillance capitalism: Monopoly-finance capital, the military-industrial complex, and the digital age. *Monthly Review*, *66*(3), p.1.

Georgieva, I., Lazo, C., Timan, T. and van Veenstra, A.F., 2022. From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. *AI and Ethics*, pp.1-15.

Given, L., 2008. *The Sage encyclopedia of qualitative research methods*. Los Angeles, Calif.: Sage Publications.

Giziński, S., Kuźba, M. and Biecek, P., Meta-analysis of academic discourse about interpretability, transparency, and fairness. Available at: <https://docs.mlinpl.org/virtual-event/2020/posters/39-Pipeline_for_automated_metaanalysis_and_tracking_of_academic_discourse_about_models_interpretability_transparency_and_fairness.pdf>

Glass, M.R. and Rose-Redwood, R. eds., 2014. *Performativity, politics, and the production of social space* (Vol. 15). New York: Routledge.

Goasduff, L., 2021. *The 4 Trends That Prevail on the Gartner Hype Cycle for AI, 2021*. [online] Gartner. Available at: <https://www.gartner.com/en/articles/the-4-trends-that-prevail-on-the-gartner-hype-cycle-for-ai-2021> [Accessed 2 June 2022].

Goode, L., 2018. Life, but not as we know it: A.I. and the popular imagination. *Culture Unbound*, 10(2), pp.185-207.

Google Research. 2022. *PAIR – Google Research*. [online] Available at: <https://research.google/teams/brain/pair/> [Accessed 29 July 2022].

Guest, G., MacQueen, K.M. and Namey, E.E., 2011. *Applied thematic analysis*. SAGE publications.

Haenlein, M. and Kaplan, A., 2019. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, *61*(4), pp.5-14.

Hagendorff, T., 2021. Blind spots in AI ethics. *AI and Ethics*.

Hagendorff, T., 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), pp.99-120. Available at: <https://link.springer.com/content/pdf/10.1007/s11023-020-09517-8.pdf>

Hagerty, A. and Rubinov, I., 2019. Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. *arXiv preprint arXiv:1907.07892.*

Haider, A. and Islan, W., 2020. *Staying ahead of the curve – The business case for responsible AI.* [online] The Economist Intelligence Unit. Available at: <https://pages.eiu.com/rs/753-RIQ-438/images/EIUStayingAheadOfTheCurve.pdf> [Accessed 29 July 2022].

Hao, K., 2021. *How Facebook got addicted to spreading misinformation*. [online] MIT Technology Review. Available at: <https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/> [Accessed 4 May 2022].

Hao, K. and Kruppa, M., 2022. *Tech Giants Pour Billions Into AI, but Hype Doesn't Always Match Reality*. [online] Available at: <https://www.wsj.com/articles/tech-giants-pour-billions-into-ai-but-hype-doesnt-always-match-reality-11656508394> [Accessed 29 July 2022].

Harvey, D., 2007. *A brief history of neoliberalism*. Oxford University Press, USA.

Hilligoss, H. and Fjeld, J., 2021. *Introducing the Principled Artificial Intelligence Project*. [online] Berkman Klein Center. Available at: <https://cyber.harvard.edu/story/2019-06/introducing-principled-artificial-intelligence-project> [Accessed 8 December 2021].

IBM. 2022. AI Ethics. [online] Available at: <https://www.ibm.com/artificial-intelligence/ethics> [Accessed 29 July 2022].

Jasanoff, S., 2015. Future imperfect: Science, technology, and the imaginations of modernity. *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*, pp.1-33.

Jensen, B., Whyte, C. and Cuomo, S., 2019. Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence. *International Studies Review*, [online] 22(3), pp.526-550. Available at: <https://academic.oup.com/isr/article-abstract/22/3/526/5522301?redirectedFrom=PDF&casa_token=4ft9XkVcNscAAAAA:5D1TNTVx9J5MThH74HsBY8Ggoi6yBtSDjb1QUYhtUH3db9MyeKNegRDoKvnhZUcac4HILlajKzl_xg>.

Jobin, A., Ienca, M. and Vayena, E., 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), pp.389-399. Available at: <https://www.nature.com/articles/s42256-019-0088-2.pdf>

Kavenna, J., 2019. *Shoshana Zuboff: 'Surveillance capitalism is an assault on human autonomy'*. [online] the Guardian. Available at: <https://www.theguardian.com/books/2019/oct/04/shoshana-zuboff-surveillance-capitalism-assault-human-automomy-digital-privacy> [Accessed 5 July 2022].

Kennedy, A., 2020. *Why We Need to Audit Government AI | Montreal AI Ethics Institute*. [online] Montreal AI Ethics Institute. Available at: <https://montrealethics.ai/why-we-need-to-audit-government-ai/> [Accessed 8 June 2022].

Kennedy, A., Coates, D. and Lindquist, K., 2020. Auditing Government AI: How to assess ethical vulnerability in machine learning. In Workshop on Navigating the Broader Impacts of AI Research Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020).

Lauer, D., 2021. You cannot have AI ethics without ethics. *AI and Ethics*, *1*(1), pp.21-25.

Lim, W.K., 2011. Understanding risk governance: Introducing sociological neoinstitutionalism and foucauldian governmentality for further theorizing. *International Journal of Disaster Risk Science*, *2*(3), pp.11-20.

Lim, M., 2018. *History of AI Winters*. [online] Actuaries Digital. Available at: <https://www.actuaries.digital/2018/09/05/history-of-ai-winters/> [Accessed 13 June 2022].

Lohr, S., 2016. *IBM Is Counting on Its Bet on Watson and Paying Big Money for It (Published 2016)*. [online] Nytimes.com. Available at: <https://www.nytimes.com/2016/10/17/technology/ibm-is-counting-on-its-bet-on-watson-and-paying-big-money-for-it.html?emc=edit_th_20161017&nl=todaysheadlines&nlid=62816440> [Accessed 29 July 2022].

Lungarella, M., Iida, F., Bongard, J. and Pfeifer, R., 2007. AI in the 21st Century - With Historical Reflections. In: *50 Years of Artificial Intelligence: Essays Dedicated to the 50th Anniversary of Artificial Intelligence*, 1st ed. [online] Berlin: Springer-Verlag, pp.1-18. Available at: <https://link.springer.com/content/pdf/10.1007/978-3-540-77296-5.pdf> [Accessed 13 July 2022].

Lyon, D., 2019. Surveillance capitalism, surveillance culture and data politics. In: D. Bigo, E. Isin and E. Ruppert, ed., *Data Politics: Worlds, Subjects, Rights*, 1st ed. New York: Routledge, pp.64-79.

MacCarthy, M., 2019. *An Examination of the Algorithmic Accountability Act of 2019*. Content Moderation Online and Freedom of Expression. [online] Transatlantic High Level Working Group. Available at: <https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/Algorithmic_Accountability_TWG_MacCarthy_Oct_2019.pdf> [Accessed 29 July 2022].

Markman, J., 2019. *Artificial Intelligence Beats the Hype With Stunning Growth*. [online] Forbes. Available at: <https://www.forbes.com/sites/jonmarkman/2019/02/26/artificial-intelligence-beats-the-hype-with-stunning-growth/?sh=1fa4fa3f1f15> [Accessed 10 July 2022].

McClay, W.J., 1995. Surviving the AI Winter. In *ILPS* (pp. 33-47).

Moss, E., Watkins, E., Singh, R., Elish, M. and Metcalf, J., 2022. *Assembling Accountability*. AI on the Ground Initiative. [online] Data & Society. Available at: <https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability-Policy-Brief_pdf.pdf> [Accessed 18 July 2022].

Minkkinen, M., Niukkanen, A. and Mäntymäki, M., 2022. What about investors? ESG analyses as tools for ethics-based AI auditing. *AI &amp; SOCIETY*, [online] Available at: <https://link.springer.com/article/10.1007/s00146-022-01415-0>.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T., 2019, January. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Mittelstadt, B., 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, [online] 1(11), pp.501-507. Available at: <https://www.nature.com/articles/s42256-019-0114-4>.

Mittelstadt, B., 2019. AI Ethics–Too principled to fail. *arXiv preprint arXiv:1906.06668*.

Mökander, J., Morley, J., Taddeo, M. and Floridi, L., 2021. Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. *Science and Engineering Ethics*, *27*(4), pp.1-30.

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J. and Floridi, L., 2021. Ethics as a service: a pragmatic operationalisation of AI ethics. *Minds and Machines*, *31*(2), pp.239-256.

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M. and Floridi, L., 2021. Operationalising AI ethics: barriers, enablers and next steps. *AI & SOCIETY*, pp.1-13.

Morley, J., Floridi, L., Kinsey, L. and Elhalal, A., 2020. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, *26*(4), pp.2141-2168.

Müller, V., 2021. History of digital ethics. In: C. Veliz, ed., *Oxford Handbook of Digital Ethics*, 1st ed. Oxford: Oxford University Press, pp.1-18.

Nourbakhsh, I.R., 2021. AI ethics: a call to faculty. *Communications of the ACM*, *64*(9), pp.43-45.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K. and Mordvintsev, A., 2018. The building blocks of interpretability. *Distill*, *3*(3), p.e10.

Pastaltzidis, I., Dimitriou, N., Quezada-Tavarez, K., Aidinlis, S., Marquenie, T., Gurzawska, A. and Tzovaras, D., 2022, June. Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2302-2314).

Rabionet, S., 2016. How I Learned to Design and Conduct Semi-structured Interviews: An Ongoing and Continuous Journey. *The Qualitative Report*, [online] 16(2), pp.563-566. Available at: <https://nsuworks.nova.edu/tqr/vol16/iss2/13/>.Bazeley, P., 2020. *Qualitative data analysis*. 2nd ed. London: SAGE eBooks.

Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P., 2020, January. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44).

Richards, L., 2020. *Handling qualitative data: A practical guide*. London: SAGE eBooks.

Romero, A., 2021. *Meet M6—10 Trillion Parameters at 1% GPT-3's Energy Cost*. [online] Medium. Available at: <https://towardsdatascience.com/meet-m6-10-trillion-parameters-at-1-gpt-3s-energy-cost-997092cbe5e8> [Accessed 29 July 2022].

Saldaña, J., 2021. *The coding manual for qualitative researchers*. Washington D.C.: SAGE eBooks.

Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., Newnham, G., Hajkowicz, S., Robinson, C. and Hansen, D., 2021. AI ethics principles in practice: Perspectives of designers and developers. *arXiv preprint arXiv:2112.07467*.

Shevlin, H., Vold, K., Crosby, M. and Halina, M., 2019. The limits of machine intelligence. *EMBO reports*, [online] 20(10), pp.1-5. Available at: <https://www.embopress.org/doi/epdf/10.15252/embr.201949177>.

Simon, H.A., 1965. *The shape of automation for men and management* (Vol. 13). New York: Harper & Row.

Smagorinsky, P., 1995. The social construction of data: Methodological problems of investigating learning in the zone of proximal development. *Review of educational research*, *65*(3), pp.191-212.

Spencer, M., 2022. *Artificial Intelligence Hype Is Real*. [online] Forbes. Available at: <https://www.forbes.com/sites/cognitiveworld/2019/02/25/artificial-intelligence-hype-is-real/?sh=2aa7a14625fa> [Accessed 29 July 2022].

Spotify Research. 2022. *Algorithmic Responsibility*. [online] Available at: <https://research.atspotify.com/algorithmic-responsibility/> [Accessed 29 July 2022].

Statista. 2022. *Total data volume worldwide 2010-2025 | Statista*. [online] Available at: <https://www.statista.com/statistics/871513/worldwide-data-created/> [Accessed 29 July 2022].

Stephens, E., 2022. The mechanical Turk: a short history of 'artificial artificial intelligence'. *Cultural Studies*, [online] pp.1-23. Available at: <https://www.tandfonline.com/doi/full/10.1080/09502386.2022.2042580?cookieSet=1>.

Toosi, A., Bottino, A.G., Saboury, B., Siegel, E. and Rahmim, A., 2021. A brief history of AI: how to prevent another winter (a critical review). *PET clinics*, *16*(4), pp.449-469.

Tran, D., 2017. *Why the AI Hype Train is Already off the Rails and Why I'm Over AI Already*. [online] Medium. Available at: <https://medium.com/built-to-adapt/why-the-ai-hype-train-is-already-off-the-rails-and-why-im-over-ai-already-e7314e972ef4> [Accessed 29 July 2022].

Villerreal, A., 2020. *US healthcare workers protest chaos in hospitals' vaccine rollout*. [online] The Guardian. Available at: <https://www.theguardian.com/world/2020/dec/21/us-healthcare-workers-protest-chaos-hospitals-vaccines-vaccinations> [Accessed 31 July 2022].

Vopson, M., 2021. *The world's data explained: how much we're producing and where it's all stored*. [online] The Conversation. Available at: <https://theconversation.com/the-worlds-data-explained-how-much-were-producing-and-where-its-all-stored-159964#:~:text=In%202018%2C%20the%20total%20amount,mind%2Dboggling%20175ZB%20by%202025.> [Accessed 29 July 2022].

Wagner, B., 2018. Ethics as an escape from regulation. From "ethics-washing" to ethics-shopping?. In *Being Profiled* (pp. 84-89). Amsterdam University Press.

Waldrop, M.M., 2019. What are the limits of deep learning?. *Proceedings of the National Academy of Sciences*, *116*(4), pp.1074-1077.

Walmsey, J., 2012. *Mind and Machine*. 1st ed. New York: Palgrave Macmillan, pp.5-17.

Whittaker, M., 2021. The steep cost of capture. *Interactions*, *28*(6), pp.50-55.

Wong, R., Madaio, M., and Merrill, N. *Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics.* arXiv:2202.08792v1, 17 Feb. 2022.

Zhu, L., Xu, X., Lu, Q., Governatori, G. and Whittle, J., 2022. AI and Ethics—Operationalizing Responsible AI. In *Humanity Driven AI* (pp. 15-33). Springer, Cham.

Zuboff, S., 2015. Big other: surveillance capitalism and the prospects of an information civilization. *Journal of information technology*, *30*(1), pp.75-89.

Zuboff, S., 2019. *The age of surveillance capitalism: The fight for a human future at the new frontier of power: Barack Obama's books of 2019*. Profile books.