THE UNIVERSITY
of EDINBURGH

Transmuting Values in Artificial Intelligence:

Investigating the Motivations and Contextual Constraints Shaping the Ethics of Artificial Intelligence Practitioners

SJ Bennett

Thesis submitted for the degree of Doctor of Philosophy

The University of Edinburgh

2023

Abstract

Advances in Artificial Intelligence (AI) research and development have seen AI applied in various high-stakes domains such as healthcare and welfare. Furthermore, portrayals of AI are often characterised by narratives of perpetual progress and sleek optimisation, obscuring the intricate interactions of materiality and socio-political decision-making inherently embedded within wider systems of design and development. The resulting ethical and social concerns have prompted proposal of numerous frameworks, tools and guidelines for the ethical design and development of AI. However, translating these proposals into practice has proven challenging, and there is a paucity of research into the practical contexts shaping the ethico-onto-epistemology of AI practice. In this thesis I illustrate these contexts via the accounts of 24 AI practitioners, complemented by ethnographic observations from an industry research lab, examining the values which motivate practitioners, the constraints which shape their practice, and their approaches to ethics.

Weaving through these discussions of practice, values, and the nature of responsibility, I examine how ambiguities pervade practice and shape the realities of ethical reflection and engagement at all stages of development. My findings uncover practitioner motivations linked with interconnected intellectual and moral values, how these related to intellectual conduct and culture within the field, and how practitioner heuristics for ethical decision-making are often relational and character-based in nature. This realization of values in practice is tempered by numerous constraints including hardware limitations, epistemic cultures, and ethical knowledge. Drawing upon the *Ethics of Ambiguity,* I discuss how the uncertainty, ambiguity and unequal access to resources shaping AI practice necessitate a process-focused ethics which pivots away from solutions, towards critical contextual reflexivity and awareness of how contexts impact realisation of values. To this end, I demonstrate how *The Ethics of Ambiguity* can offer a path forward for ethical AI practice. This vision of AI practice embraces ambiguities rather than attempting to segment and sideline them, focusing on how practitioner decisions (and their eventual outputs) impact others' freedoms while acknowledging the multiplicity of values across socio- and geo-political contexts.

Lay Summary

Artificial Intelligence (AI) models are being applied in domains impacting many aspects of life, such as healthcare, sentencing, recruitment and even dating. In designing and developing AI models, AI practitioners are making day-to-day decisions which have serious ethical implications. Numerous ethical tools and frameworks have been developed to try and guide ethical development of AI, however, there is still a large gap between policy and practice, and a sparsity of studies which take into consideration how contexts of AI practice, and practitioner motivations, play into AI ethics.

Aiming to better understand AI ethics in practice, I used qualitative research methods to investigate the values which motivate practitioners, the contexts and barriers shaping their work and outputs, their understanding of responsibility with regards to ethics of AI, and their existing approaches to ethics.

I found that AI practitioners are motivated by intellectual values such as efficiency and a keen pursuit of knowledge, as well as moral values such as equality of treatment and opportunity. The way these values and AI models are built are expressed is highly influenced by contexts such as culture, access to resources and the inherent ambiguities of AI practice. Reflecting upon these findings, I demonstrate how Simone de Beauvoir's ethical theory as laid out in *the Ethics of Ambiguity* can provide a useful framework for thinking about and guiding ethics in AI practice.

Acknowledgements

I wish to thank the wonderful people who have supported me through the PhD process and made it possible for me to get this far. Firstly, thanks to my supervisors Ewa Luger, Chris Speed and Richard Banks for your sage advice and openness to the wildly divergent path I have taken, but also the support in my many non-PhD related ventures, which in the end have also contributed massively to my thinking regarding my research.

I am deeply grateful to my family and friends, who have kept me going throughout this process and inspired me so many times. A special shout-out to Sonia Garcia de Alba for your tireless support, I owe you several thesis-reading sessions! I was incredibly lucky with my Edinburgh College of Art PhD cohort Patricia Wu Wu, Asad Khan and Pushpi Bagchi - thanks for the many thought-provoking conversations which really challenged my thinking and influenced my direction with this research. I owe a lot to my friends at AI Ethics & Society - Benedetta Catanzariti, Vassilis Galanos, Aditi Surana and Lara Dal Molin - who had a foundational impact on my research as well as many great nights chatting over drinks. Thanks also to my talented friends Jessica Alberg and Carolina Hayes for keeping me sane during the COVID-19-and-writing-up combination with our Zoom drawing sessions.

I am especially grateful to my amazing partner Imo, my parents and my sister for their endless patience and kindness during the process of writing up – and most of all lovely Polo, my biggest cheerleader.

Table of Contents

CHAPTER 1: Introduction

The research set out in this thesis began in a meeting at a Big Tech research lab, during the early stages of designing and developing an AI-driven assistive device. Jointly introduced to both the lab and my doctoral research, I keenly jotted down notes and roughly sketched ideas before presenting an interlude of slides which invoked mantras of "ethics", "consent" and "trust", punctuating an afternoon of technical discussion and logistical planning. Over the course of several such discussions, it became clear that the salient ethical implications and decisions were bound up in the everyday practices of the Artificial Intelligence practitioners in the team[1]. Whilst the domain knowledge, software engineering and design skills of other team members were crucial in materialising the proposed device, these were contingent on the output of computer vision models. Thus, to a large extent, the AI researchers set the pace of the project: it was their expertise informing how to approach the aims set up in project management, in combination with the willingness of senior lab figures to fund this work. In this way, the embedded nature of my PhD research provided first-hand experience of the conflict between abstracted ethical principles/frameworks and the dynamic, complex realities of practice.

Weaving through discussions of practice, values, and the nature of responsibility, in this thesis I examine how ambiguities pervade practice and shape the realities of ethical reflection and engagement at all stages of development. Initially I had aimed to map the ethical implications of the project. However, I quickly realised the limitations of the concepts I was trying to apply, observing the same iterative and uncertain patterns of work which had dominated my brief previous experiences with AI. Combined with the highly specialised knowledge of the AI practitioners, who in my view made the important ethical decisions in their own day-to-day work, this uncertainty complicated ethics work to a significant extent, whether this work be developing models of consent, or mapping ethical implications of a project. These experiences served as the starting point for my interest in, and interrogations of, ethical conduct within the field. Following from this, in

[1] AI Practitioner is used as an umbrella term for individuals with significant training or experience in the field.

this thesis, I investigate the motivating values and ethical perspectives shared by AI practitioners in qualitative interviews and observations, sketching out AI practices as experimental, highly iterative, and difficult to predict, compounded by contextual pressures such as access to resources and cultural norms. I then reflect upon how the ambiguities formed in these contexts stretch beyond their immediate conditions to form an uncertain human-AI relationship, drawing upon *the Ethics of Ambiguity* (de Beauvoir, 1962) to consider the implications of this for developing a meta-ethics of AI.

In the years since the start of this research in 2017, themes of ethical and social implications of AI have gone from barely discussed outside of specific academic and technical circles, to a burgeoning domain with regular news stories, and numerous frameworks and principle-sets. Part of this is terminology change, given overlaps between different communities working on the ethics of AI and computer science. It also reflects the meteoric rise in funding and application of AI, which is even beginning to permeate everyday life. In light of this, in the next section I discuss some of the contemporary shifts in AI development and applications, to provide some contexts for the field at large.

The Confluences Shaping Artificial Intelligence

The AI industry is projected to value 422.37 billion dollars by 2028, with China indicated to be the world leader in investment whilst Silicon Valley in the United States (US) has been the forerunner up to now, hosting industry giants like Alphabet (includes Google), Cisco, Intel and [Facebook's] Meta (Bloomberg 2022). AI adoption in business increased as much as 270% between 2015 and 2019, accompanied by a shortage in AI practitioners (Costello 2019). Meanwhile, at the time of writing this thesis, even a cursory glance at the news and social media reveals that Artificial Intelligence is burgeoning across numerous domains, positioned as providing novel insights into real-world problems (Zhou 2018), impacting on medical decision-making (Fisher *et al* 2019), hiring practices (Chalfin *et al* 2016), entertainment (Amatraian 2013), education (Luan and Tsai 2021) and even art (Fiebrink 2019). These far-reaching and influential applications of AI have implicit social

and ethical dimensions, which I discuss throughout this thesis, which not only affect existing systems but also shape potential implementations in new domains.

At the centre of many of these outputs lies mention of efficiency, objectivity and simplicity, values which are also evident in the AI literature (Birhane *et al* 2022). However, human constructs and societies are inherently complex, multi-dimensional, and subjective. In attempting to collapse this complexity into formats amenable to AI, the result is that people and their circumstances are represented by reductive, partial data-points (Birhane 2021). Omitted information tells as much of a story as the explicitly included information - datasets/models are "partial, situated and contextual" (Leurs 2017, p. 150). Thus, even projects such as the one mentioned in this introduction, which aim to do good and aid marginalised groups, can risk resulting in further marginalisation, by "reproducing existing power/knowledge frameworks that marginalize underrepresented groups" (Parson 2019, p. 16). This contributes to epistemic injustice, unfair composition of forms of knowledge available for use, due to inequities which shape the process through which such knowledge is obtained and distributed (Fricker, 2007). I revisit the concept of epistemic injustice in the methods chapter, as it serves a useful lens for interrogating the ethical impacts and responsibilities of AI practice.

Many AI projects consist of large, distributed teams consisting of various roles. To complicate this further, AI models are shaped by context: by time constraints, cultural influences, even the tools used to generate them (Jo and Gebru 2020), whilst the datasets the models are trained on are products of the systems and social contexts they were formed within (Miceli *et al* 2022). Therefore, building and using datasets/models involves making decisions with inherent moral and social implications. Viewing data and models as objective - and considering them in isolation from shaping factors - reproduces harms and discrimination in the resulting outputs (Mhlambi and Tiribelli 2023). Conversely, perceptions of AI are constructed and shared according to a multiplicity of values and aims, for example framing the epistemic contributions of AI to appear more favourable. Practitioners in pattern recognition intentionally framed the field as data-led rather than human-defined, enabling a move away from perceptions of "mechanical identification of significance and the reproduction of human judgment"

(Mendon-Plasek 2021, p. 32). This allowed reorienting of the problem of incomplete data from a flaw into a feature, a liability into an "epistemic virtue of greater contextual sensitivity" (Mendon-Plasek 2021, p. 34). Thus, even the framing of the methods themselves belie certain values.

Indeed, in the words of Artificial Intelligence pioneer Marvin Minsky, AI is a *suitcase word* (Winston 2016, p. 282), carrying a unique connotation to each person who uses it[2]. There is a rich history of AI practitioners reflecting on the implications and values of their work, even bemoaning a lack of paralleled reflection on the part of their colleagues (McDermott 1976). These reflections may be broader, analysing contrasts between the reality of their own practices and the perceptions of those on the outside, or more nuanced, for example how AI practice differs from the work of a software developer. Early critiques of AI practice had undertones of concern about the impact of AI hype cycles (Elish and Boyd 2018) - that is, patterns of successes followed by failures, which form part of the external constraints impacting AI practice. The early 70s had seen a rapid decline in funding of AI, or 'winter', in the wake of the Lighthill Report (Agar 2020), with memory of this forming part of the culture of the revived 'summer' of the field. In *Alchemy and Artificial Intelligence,* Dreyfus argued that the grandiose claims made by AI researchers were creating false expectations and obfuscating understanding of AI, illustrating this using the example of alchemy which "has shown that any research which has had an early success can always be justified and continued by those who prefer adventure to patience" (Dreyfus 1965, p. 85). Although approaches to AI have shifted, this metaphor, and the concerns which it alludes to, remains apt today, in often vague yet evocative descriptions of both product and practice.

Intertwined Practices, Values and Ethics

In designing technological artefacts which impact human life, AI practitioners are practicing ethics, contributing to how these lives are lived (Vallor 2016), whether these decisions are consciously considered, or conducted without express reflection on moral

[2] Given the practice-based focus of this research, I use 'AI' to refer to limited and domain-bound AI rather than Artificial General Intelligence (AGI)

implications. Like AI, ethics is a suitcase word. It might refer to consideration of philosophical arguments for the status of AI as a moral agent (Himma 2009), seek to understand placement of responsibility of systems in automated warfare (Horowitz 2016), aim to elucidate how treatment of AI impacts attitudes to certain demographics of society (van Grunsven and van Wynsberghe 2019), enshrine certain principles such as autonomy within human-AI interaction (Sankaran *et al* 2020), seek to develop moral AI agents (Shaw *et al* 2018), or present professional ethics frameworks intended to guide design and development of ethical AI (Floridi *et al* 2018). All these interpretations tackle important and timely aspects of AI ethics. They also raise implicit questions about the role that AI practitioners play in creating the AI models which these ethical approaches address (except, perhaps, for the treatment of the hypothetical AI moral agent), and indeed which considerations are reasonable for a practitioner to be aware of and consider in their practice. Fundamentally, ethics asks that we identify and understand the moral and social impacts of decisions and helps us to better navigate the complexities of these.

The increase in AI ethics and governance approaches over the past few years provides a necessary and important reaction to the incursion of AI into many aspects of life and determination of aspects of futures. However, often the focus is placed upon values and outputs, often side-lining the materiality of AI practice, and in doing so missing key ethical considerations. Furthermore, there is emerging concern in the literature about the principle-practice gap, that well-meaning ethics frameworks, principles and tools are not being applied in practice, whether that be due to unsuitability or unwillingness (Schiff *et al* 2021; Morley *et al* 2021). Also, there is a paradox of guidance being most impactful at the earliest stages of a project, whilst difficult to properly provide early on due to the uncertainties of how the project will unfold in practice (Rittel and Webber 1973). Meanwhile, the continually increasing complexity of algorithms and richness of data is leading to a "gap between the design and operation of algorithms and our understanding of their ethical implications" (Mittelstadt *et al* 2016, p. 2), which may result in "severe consequences affecting individuals, groups and whole segments of a society." (Mittelstadt *et al*, 2016, p. 2).

These factors combine to produce a fundamental ambiguity inherent to ethical decision-making in the context of framing and developing AI. An irony of ambiguity is that it has multiple definitions pertaining to it, as well as a myriad of related terms. Referring to situations, phrases and other objects which can signify multiple meanings, it bears a close relation to uncertainty and risk (Sennet 2011). Ambiguity refers to a lack of clarity in meaning, interpretation and even states of being, all subject to inherent nuances, or the "inescapability of interpretation" (Best 2012, p. 88). Uncertainty also has implications for conceptualising AI, with uncertainty regarding output and impact not only "bad evidence for inferring which values are embedded in the technology" (Miller 2021, p. 63) but also, I would argue, a shaky pretext upon which to design ethical interventions.

Taken together, these considerations regarding the framing of AI, the values and ideologies surrounding its development, the ambiguities around ethical considerations, and the principle-practice gap present complexities for anyone engaged in the growing industry. Through investigating the day-to-day practices and values of Artificial Intelligence practitioners, in this thesis I consider how these might inform conceptualisations of AI ethics and the design of ethical interventions in AI practice. Despite increasing recognition regarding the multiplicity of reasons why principles and guidelines for AI ethics are difficult to translate into practice, such as tensions in terms of incentives, and complexities of AI impacts (Schiff *et al* 2021), there is still limited insight into their socio-material contexts.

Aims and Contribution of Thesis

This thesis attends to the role and constitution of socio-material practice in negotiating ethics and values in AI, mapping key facets shaping the infrastructuring of AI and building upon a growing body of work in this space (Feldman & Orlikowski 2011; Hoffman 2017; Moss 2022). I employ concepts from philosopher Simone de Beauvoir's *The Ethics of Ambiguity* as a conceptual tool to scaffold my discussion of findings, reflecting upon their implications for the "ethico-onto-epistemology" (Barad 2007) of AI (further discussed in Chapter 2). Given this, values and ethics are conceptualised in this

thesis as distinct concepts, with values part of the descriptive understanding of a context, whilst ethics is construed as a normative way of identifying, reflecting upon, and making decisions, perhaps drawing upon on certain values during its practice. Accordingly, my overarching research questions are as follows,

> *'What kinds of constituent factors and values shape contemporary AI practice? What are the implications of these for developing a practical ethics of AI?'*

In doing so, the research focuses upon a specific role within teams which are often multidisciplinary and distributed, a role which holds power within decision-making, conceptualising and designing AI models. I investigate the role of the AI practitioner, and how the conditions of AI practice impact on their understandings of ethical responsibility and how to engage in it. The perspectives and decisions of practitioners, like anyone, are tempered by their values and contexts.

To address this research focus, I used a qualitative, practice-focused ethnographic approach. With origins in the fields of Science and Technology Studies (STS) and Human-Computer Interaction (HCI), practice-based studies of technologies frame algorithms as complex entanglements of humans, technologies, and institutions (Christin 2020). Most of this work has focused upon the study of AI outputs as comprised of situated entanglements between technology and users. However, as this thesis demonstrates, the practices of producing AI are also fundamentally relational. In order to convey knowledge which would allow for adequate definition of distinct ethical considerations such as informed consent, it is crucial to understand the motivations driving the project, and the contexts forming it, as well as the specific workings of the outputs. To expand upon this, "rather than considering context to be information", this thesis considers context as "a relational property that holds between objects or activities" (Dourish 2004, p .5), shaping them. Relational conceptions of ethics claim that "addressing moral problems involves first, an understanding of identities, relationships and contexts" (Robinson 1999, p. 31). Drawing these facets together, the thesis argues that relational ethics which address context and relationality, can best account for the contingencies and ambiguities of AI practice.

My findings have been informed by attending to the everyday practices of individuals working with AI, the motivational values influencing their choices, and their perspectives on responsibility and ethics in AI work. As such, in addition to presenting novel findings regarding the values and other facets co-constituting AI practice, I hope this thesis can provide insight for those engaged in the work of developing ethics in AI and be of use to AI practitioners in navigating practice.

The Research Context

As mentioned at the start of this chapter, the direction of this research was influenced by my experience embedded in a Big Tech research team which was developing an AI assistive device. This consisted of three sub-teams; AI, software engineering and human-computer interaction, with project-funded AI PhD projects included as part in these, in addition to the one represented in this thesis. Each PhD project included an external supervisor who consulted on the project, and a supervisor internal to the lab. However, this thesis is not a mapping of the workings of this specific institution or project, nor is it my intention to be constrained to these, therefore I have anonymised descriptions and employ them as illustrative stories rather than detailed analyses.

I became involved with the research team at such an early stage that the ethical implications of the project were incredibly broad, whilst the team-members with tangible opportunity for ethical impact were those with technical expertise to shape the details of project conceptualisation and development. That is not to say that I did not input into discussions. To draw from Katie Shilton's values levers (Shilton 2013), we had valuable discussions which heralded shifts in perspectives. However, my role seemed focused at a high level of commenting on the ethics of an anticipated output, when the bulk of ethical deliberation seemed to lie in the ongoing process of continually drawing and re-drawing the lines of the project and the methods used to accomplish these ends. Given these tensions within my role, recognising the external limitations which shaped the scope of activities, I turned my focus to understanding the positions and perspectives of AI practitioners with regards to ethical decision-making, with the intent to present a

contextually informed account of these activities. The product of this investigative endeavour is surmised within the thesis, as outlined below.

Thesis Outline

This chapter has introduced my research context and motivations for investigating ethics-in-practice, via examination of AI practices, practitioner values and situational contexts.

Following this introduction, Chapter 2 details the theoretical perspective underpinning this research, and the methods and methodology utilised in data collection and analysis. Locating my research, I triangulate theoretical influences amongst the interstitial space of Human-Computer Interaction (HCI), Science and Technology Studies (STS) and Philosophy of Technology. I introduce post-phenomenology as informing my study design, data interpretation and argument. Following this, I describe the qualitative methods employed – ethnography in the forms of expert interviews, an observational study, and auto-ethnographic reflections upon my time spent in the research lab.

Chapter 3 delves into existing literature to consider the facets shaping current approaches to AI ethics, from historical approaches to contemporary influences from other fields. It examines the popularity of frameworks, principles and tools, considers the assumptions upon which they are built, and looks at the growing dialogue on critical and relational approaches to ethics. The placement of this chapter here (after the methods) serves to enable a better bridging of the theoretical and practical terrain with the findings which are subsequently discussed.

Building upon this discussion of the literature, in Chapter 4 I draw from interviews, observations and personal reflections to explore the confluence of contextual factors impacting AI practice, considering the implications of these in terms of ethics and social impact. Traversing from concepts of AI to mundane practice, this chapter looks at how these are shaped by access to resources, epistemic assumptions, ambiguous and uncertain tools and timelines, and corporate capture.

Chapter 5 considers practitioner perceptions of responsibility, how contexts shape these, and the ways in which practitioners respond to these perceptions in their ethical decision-making. Following from this, in Chapter 6, I explore the motivating

values embedded within the narratives that practitioners employed in our discussions, which act as motivators and guidelines for practitioners navigating the complexities of practice.

In Chapter 7 I bring together the observations of context, values and responsibility discussed in the previous chapters to consider their implications for AI ethics. I examine ways in which the ethical theory presented in *the Ethics of Ambiguity* can provide a useful lens through which to approach AI ethics, presenting an alternative way of understanding constraints and values, in the form of embracing ambiguities and engaging reflexively in processes of seemingly mundane work. In doing so, I discuss how *the Ethics of Ambiguity* can help with understanding the nature of responsibility in a highly ambiguous, fluid but fundamentally high-stakes practice.

Chapter 8 draws together the thesis to conclude with a call to engage with the ambiguities which shape AI work as ethical practice, rather than problems to be disambiguated or solved. I then suggest some practical next steps and implications of the research, primarily to inform how AI ethics is conceptualised and engaged with in practice, and in developing approaches to AI governance.

CHAPTER 2: Theory and Methodology

This chapter discusses the theories and methodologies which have shaped the research presented in this thesis. I begin by introducing the literature which directed my research design and analysis. Informed by Philosophy of Technology, Science and Technology Studies (STS) and Human-Computer Interaction (HCI), I explore the entanglement of values, ethics, and AI practice as constituents of the socio-material 'life-worlds' of the practitioners situated with these contexts. I then describe the research methods employed, ethnographic interviews and observations, elaborating on the contexts of these. Finally, I reflect on my own positionality in conceptualising and conducting this research.

Theoretical Positioning

AI practice is exploratory and characterised by uncertainty, underpinned by assumptions regarding its "ethico-onto-epistemology" (Barad 2007, p. 90) - a view towards ethical, ontological, and epistemological domains as entangled, dynamic, and interacting facets, only distinguished by chosen lenses of inquiry. Even the work of a single research team working on computer vision forms part of a much broader constellation of various actors, processes, materials, and relationships, a complex and distributed "sociomaterial assemblage" (Suchman 2007, p. 268). In engaging with "ambiguity as cause and effect" (Suchman 2012, p. 49), "unpacking" the assemblage from this vantage point, I seek to better understand the relationality of different facets making up AI practice. Rather than aiming to locate my research as an analysis of any single location or lab, I instead seek to "disentangle" elements co-constituting this practice, by "analyzing and describing a sociomaterial entanglement in its constituents" (Bratteteig and Verne 2012, p. 17). As such, I examine these elements from different sites of inquiry, just as I apply different theoretical lenses to examine the resultant observations.

In seeking to understand these constituents I draw from a wide range of theoretical perspectives from Philosophy of Technology, STS, HCI, and Feminist studies, employing them as different lenses through which to examine my findings. These have offered up a plethora of angles, sometimes seemingly in opposition, from which to consider the relationship between values, ethics, and practice, and interpret the qualitative data which the research took the form of. Although these bodies of literature intersect in multiple ways, the central touchpoint within my research are forms of phenomenology, highlighted in entanglements and constellations of practices and ethics. This is underpinned by the notion that "there is no social that is not also material, and no material that is not also social" (Orlikowski 2006, p. 29), and informed by agential realism (Barad 2007) and post-phenomenology (Verbeek 2005).

Situated Practice

Scientific experiment is highly shaped by its local contexts (Galison and D'Agostino 1987; Galison 1995), the resultant knowledge produced in fundamentally co-constitutive socio-technical processes (Stengers 2000; Latour and Woolgar 2013). AI is no exception, as illustrated by the multiple, contextually situated ways in which the field itself is understood (Monett and Lewis 2018). Thus, the knowledge negotiated in situated practice is central to understanding the epistemic culture of the domain (Bogner *et al* 2009, p. 27). Likewise, ethics on-the-ground is shaped by situational context, sometimes flourishing despite formal ethics rather than being aided by these tools of compliance, which typically have been shaped by broader political and bureaucratic needs (Heimer 2013). Therefore, knowledge of practice is not just explicit technical knowledge, but includes understanding the tacit negotiation of values and ethics on-the-ground, consisting of "culturally constituted meanings and socially organised practices" (Karasti 2001, p. 215). I have drawn on prior work on "sociomaterial practice" as starting points for investigating these meanings (Orlikowski 2007; Suchman 2007; Van Dijk and Rietveld 2017), including consideration of material practices, spaces, and histories in my analysis.

Entanglements and Moral Mediation

My epistemological and ontological perspectives are informed by post-phenomenology (see below) and Agential Realism, which conceptualises phenomena as intra-acting agencies, mutually co-constitutive (Barad 2007). Technologies impact human practice in ways beyond the implications of their socio-political constitution and socio-historical shaping, as seen in the reflections of Weizenbaum (1976) and McDermott (1976). My perspective on practical ethics is informed by Verbeek's (2006) conceptualisation of moral mediation in technologies (such as AI), itself influenced by the post-phenomenology of Ihde (2009).

Positing human-technology relationships as co-constitutive (Introna 2005), post-phenomenology examines the influence or "mediation" which technologies effect upon their users (and vice versa), considering the relational role which technology plays in these interactions. Moral mediation therefore examines how technologies influence moral decision-making and other ethical considerations, mapping how "distinct agencies…emerge through their intra-actions" (Barad 2007, p. 33). This includes consideration of influences on interpretation (hermeneutics) and of the existential, ontological implications upon human practices and behaviours (Smits *et al* 2022). Technologies mediate the experiences and values of designers and developers, just as with other users. In investigating these, I approach my research as a type of "agential cut" (Barad 2007), examining AI practice and ethics as intertwined, seeking to gain insight via examining these co-constitutive agencies through the life-worlds which the practitioners I interviewed shared with me, and drawing on observations from my time at the lab.

Relational and Feminist Ethics

Given this relational orientation of my ontological and epistemological foundations, it follows on that to understand the ethics of AI is to understand the interplay between the expectations and values of practitioners, and the constitution of AI practice. As such, it is important to understand how the "objects" of AI ethics, whether these be principles for abstract moral behaviour or checklists designed to aid in incorporating fairness into AI design, relate to the everyday practices of AI workers (Madaio *et al* 2020). Anticipating impacts and understanding the values and assumptions informing AI development are

both important and need to be understood in terms of broader practice. In addition to the aforementioned work of Barad, my theoretical perspective has been shaped by other feminist scholarly works on the nature of situated knowledge and understanding and ethics (Haraway 1988). Within the context of AI, feminist methodologies have provided critical concept to examine the biases embedded in technical tools, products of "sociotechnical entanglements" which [re]produce inequalities through AI (Klumbytė *et al* 2022).

An ethics of AI, then, must situate itself in relation to the wider contexts and power dynamics of practice. Given the nature of AI practice as knowledge work, I am particularly concerned with elucidating the logics of knowledge construction underpinning AI practice, and considering ways in which knowledge may be excluded, whether intentionally or not. Formal ethics itself can act as an object of study for this purpose. In an ethnography of an HIV clinic, Mackworth-Young *et al* (2019) investigated ethics-in-practice. They found that while procedural ethics suffered limitations as to its usefulness for practical application, it was useful to map researchers' shortcomings regarding their reflexivity and limits of understanding, and for identifying which communities would be best placed to support researchers in navigating ethical complexities and ambiguities. Additionally, at points in the thesis I employ the concept of "Epistemic Injustice" as a tool for identifying and highlighting instances where certain forms of knowledge may be made invisible, considering the implications of these. Miranda Fricker (2004) first developed the theory of epistemic injustice to describe when knowledge contributions of an individual or demographic are given less or no visibility because of direct or indirect discrimination (Fricker 2017). An example of this is being denied concepts to describe experiences, therefore lacking mechanisms to convey or engage with these experiences (Pohlhaus 2012). As such, the recognition of injustice towards an underprivileged demographic implies a view towards available knowledge as largely developed for/by an overprivileged demographic, which carries epistemological and ethical implications.

In bringing together these influences of situated practice and technological mediation, I aimed to gain a deeper understanding of how the overarching structures of

AI work are produced by the efforts of those working in the field, addressing the gulf between the micro and macro (Berard 2005), to map the "ethico-political arrangement of values, assumptions and propositions" (Amoore 2020, p. 6) characterising AI and AI practice. This rich and growing body of literature provides us with 'tools to think with', concepts which can help situate our conceptions of ethics. However, none of the approaches discussed so far expressly address how to navigate ambiguity and uncertainty, even if they step beyond rationalist approaches to ethics to acknowledge its effect, and the risks of 'flattening' the inherent ambiguities of relationality (Birhane 2021).

The Ethics of Ambiguity

Simone de Beauvoir is best known for her feminist works; however, *the Ethics of Ambiguity (TEA)* provides an insightful take on moral mediation in an uncertain world. De Beauvoir proposes an ethics which builds upon her existentialist *philosophy of ambiguity*, framing ambiguity as a result of meaning-making, which surfaces tensions between values and "facticities" (material constraints). Negotiated relationally, ambiguity is construed as inherent to the process of continually co-constructing meaning in a world where none inherently exists, finding connection to the world through the ambiguous relationship between the self and the other (or the "intra-action"). De Beauvoir lays out the ambiguities of the context within which humans live, resulting from our "plurality of concrete, particular men projecting themselves toward their ends on the basis of situations whose particularity is as radical and as irreducible as subjectivity itself" (de Beauvoir 1947, p. 17). In this view, "morality resides in the painfulness of an indefinite questioning" (de Beauvoir 1947, p. 133), and requires constantly working towards the joint freedom of the self and others. This vision of ethics is framed in terms of its immediate intra-actions and historical contingencies rather than abstracted concepts or anticipated outputs as seen in rule- or consequence-based ethics.

In focusing on the tension between freedom and facticities, de Beauvoir highlights how a choice in one direction might result in harm in another, but rather than viewing the uncertainty of outcome, and pain of potential harm as a failure of ethics, to de Beauvoir this is the core of ethics. Rather than asserting universal values as an end-product, with values "temporarily and precariously grounded in the particular needs and

projects of each particular human community" (Oganowski 2013, p. 6), the goal is instead to enable and protect the freedom of the self and others, especially given the observation that freedom is never static and thus all bear some responsibility in achieving it. De Beauvoir refers to two kinds of freedom; the existentialist kind which people have by virtue of existing, to make their own decisions (subject to contextual pressures), and moral freedom, the ability to engage in ethical decision-making, which has a corresponding accountability attached; does someone engage with ethical decision-making, or do they try to evade their responsibility? *TEA* primarily refers to the moral kind of freedom, recognising that in moral decision-making every decision has the potential to fundamentally impact upon other people.

Furthermore, as a choice which helps one person's freedom potentially curtails another's, "one finds himself in the presence of a paradox that no action can be generated for man without its being immediately generated against men" (de Beauvoir 1947, p. 107). Moral freedom requires that the decision-maker recognise both their "individuality and role in the collective human community" (Oganowski 2013, p. 6), requiring a recognition of practitioners' embodiment and relationality. De Beauvoir not only makes the case for responsibility and reflexivity in the face of ambiguity but also illustrates some roadblocks to engaging with it, through archetypes representing the ways people abdicate and obfuscate responsibility.

This account of ethics serves multiple purposes in this thesis; to critique the simplistic categorisations which can happen in AI, and to provide a tool for viewing the practices of AI which do not reduce them to arbitrary categories for ease of framework design. De Beauvoir sees ambiguity in the nature of humans as concurrently 'bodies' or observable by the Other, and 'lived' or experiencing ourselves moment-by-moment, but this ambiguity is not framed as a problem or duality to be dissected, but rather as a fundamental characteristic of being. "Ambiguity must not be confused with that of absurdity. To declare that existence is absurd is to deny that it can ever be given a meaning; to say that it is ambiguous is to assert that its meaning is never fixed, that it must be constantly won" (de Beauvoir 1947, p. 160). Our existences are disclosed by virtue of our relationality, with ambiguity a natural result of the multiplicity of human

experiences shaped by diverse facticities and values. This foregrounding of relationality, multi-facticity and intra-action informed the framing of the research question and choice of methods, which I describe below.

Methodology and Methods

In Chapter 1, I set out the research questions which this thesis aims to address:

1) '*What kinds of constituent factors and values shape contemporary AI practice?*'

2) '*What are the implications of these for developing a practical ethics of AI*?'

In order to investigate these questions, I undertook an ethnographic approach to qualitative research, informed by the STS-oriented perspective that the facets making up "… the social world (that is, social relations, organizations, division of labor) are not "given", rather, the social world is "produced and reinforced by humans through their action and interactions" (Orlikowski and Baroudi 1991, p. 12).

Understanding the Lifeworld of Practitioners

To draw from de Beauvoir's *the Ethics of Ambiguity*, I investigate the 'facticities' which form the limits of practitioner agency, the "clusters of constraints" which shape practice (Galison 1995, p. 15). In doing so, I seek to enrich understanding of contemporary AI practice by providing a constellation of insights into the embodied cultures and materialities of practice (Ihde 2008). There is a rich body of literature investigating the nature of work practices at the intersection of the micro and macro levels of organisations (Gaggiotti *et al* 2016), building up bodies of knowledge directly grounded in observed work practices (Rennstam and Ashcraft 2014). Such ethnographic accounts of practice typically combine different sources of insight, such as exploratory interviews, direct observation, and document analysis (Lareau 2018). These research methods have been applied specifically in AI, and indeed my methodological design has been influenced by the work of researchers such as Nick Seaver, who published guides to anthropological study of AI (Seaver 2017; 2018). There is a limited body of research employing ethnography to study AI practice (Suchman and Trigg 1993; Hoffman and Friedman 2017; Mackenzie 2017), and none which seem to explicitly place objects of ethics and

values as the objects of studying practice. Rather than seeking to audit the cultural influences impacting design of algorithms, this research investigates the social and contextual facets shaping ethical deliberation in processes of AI design and development. Given this aim to probe at the 'lifeworld' of the practitioner (Brinkmann and Kvale 2019, p. 9), my research methods all possess an ethnographic quality, reflected in how I present my emergent themes in my findings chapters (Chapters 4, 5 and 6).

I engaged with practitioners primarily via qualitative interviews, complementing findings from these with observations from a week-long ethnography where I shadowed an AI practitioner working in a research lab. Specifically, I conducted semi-structured life-world interviews (Brinkmann and Kvale 2018), where I investigated the everyday practices, values, and ethics of AI from the perspective of the practitioners themselves. In addition to allowing for a more nuanced and precise discussion of the topics of interest, taking a qualitative interview approach allowed freedom for topics to emerge free from the confines of a strictly adhered-to interview schedule or survey. Taking this approach also allowed for meanings to be negotiated and re-negotiated, for experiences to be elaborated on and reinterpreted and responded to, over the course of a single interview. This was of particular importance given the highly contextual topics which I was investigating, such as ethics and AI.

A quantitative approach such as gathering larger scale data from surveys, would be far less suitable to the aims of this research. Surveys require participants to respond to rigidly structured pre-existing categories, in the process obfuscating the context of their responses, participant understanding of the questions themselves and thus the situated meaning of responses. A systematic survey can tell us what values participants might choose from a list, but not give the nuance of what motivates them, and indeed the list of motivating values and relevant experiences would have to be pre-decided. Open-ended responses would still be primed by the survey questions with no option for clarification or exploration of themes, contrasting with the flexibility of interviews which allows the researcher to respond to subtle cues, probe answers and clarify concepts. Surveys "at best allow for knowledge at the level of the discursive consciousness

containing rationalist reasoning corresponding with officially accepted standards", in contrast interviews give the opportunity for participants to "reveal a lot more about relevances and maxims connected with their positions and functions: when they carry on talking about their activities, extemporize, give examples, or use other forms of exploration" (Bogner *et al* 2009, p. 31). As such, the latter was most suited to the research aims.

Taking a qualitative interview approach challenged both my skills and my own life-world perspective, as someone who had recently traversed from quantitative to qualitative research, requiring that I renegotiate my understandings of what constitutes data, challenging my perception of what constituted validity, and ultimately my own epistemology. Despite best intentions to approach interviewing as co-constructing understanding rather collecting facts, I constantly found myself constructing new categories in the hope that these would seem more reliable. Or perhaps fretting over the precise degrees to which a sub-theme occurred, despite my awareness of how pointless this was in the context of qualitative interviews.

A silver-lining to the learning-curve of this transition was that the small degree of knowledge I held about AI enabled faster fine-tuning of my position as the interviewer with 'qualified naivete' (Brinkmann and Kvale 2018, p. 13). With regards to the insider-outsider dynamics of qualitative research, I existed in "the space in between" (Dwyer and Buckle 2009), in a similar position to Julian Orr in his investigation of the workings of Xerox technicians (having been a Xerox technician himself). Like Orr, I found my past experiences acted as both "a boon and a curse" (Orr 2016, p. 7), having the benefit of being less intrusive due to requiring less explanation, yet also running the risk of missing details which might be intriguing to others completely unacquainted with the field. In this process of exploration, I found speaking with colleagues invaluable in recognising important details which might otherwise be overlooked. Similarly, my transition from quantitative to qualitative work, although a challenging process, also resulted in discussions and reflections that provided a useful bridge between the abstraction of research and the complexities of practice; forcing me to confront the ambiguities of my

own work formed another aspect of my engagement with the literature introduced in the discussion in Chapter 7.

Thus, acutely aware of the limitations of my own perspective, but with an awareness of the domain including common contextual touchpoints, I was fortunate to be able to facilitate the interview in a way which enabled the participant's own experiences to lead it, with the awareness of when to probe and prompt, if necessary, to provide further insight into the topics of discussion. That is not to say this was an easy process, and certainly my first few interviews suffered from the 'first pancake' phenomenon where the initial set of responses helped me consider and calibrate my investigative approach. Overall, however, I often found that we naturally covered many of the main areas of consideration without ever needing to explicitly ask, which may have been due to a myriad of factors including my previous experiences within the field, my recognition as a colleague (given the context of the study) or even how I was perceived by interviewees. I reflect further on my role as a researcher in the section on *Positionality* at the end of this chapter.

Sampling and Recruitment Strategy

I primarily used key knowledgeable sampling (as described in more detail below) to identify and recruit participants and in doing so I focused upon AI experience. My criteria for choosing participants were that they either had doctoral training in a sub-field of AI, or equivalent work experience. This was necessary as choosing to specifically investigate the perspectives and practices of AI practitioners was a choice to narrow down the view to a very specific angle. The demographics of the practitioners I spoke with further limited this viewpoint, with participants limited to Anglophone countries, and mostly male. Most notably, China has emerged as a leader in AI (Lee 2018), however this is not reflected in either the sample or the content of the interview study discussed here. I interviewed practitioners working on applied and/or theoretical AI, finding that practitioners frequently held roles in both industry and academia, broadly representative of wider trends in the industry (Nyeko and Singh 2015). I purposefully sought to interview practitioners working in a variety of configurations within these spaces, to explore a

broad range of perspectives, including individuals working at multinational organisations, SMEs, and start-ups. The breadth of roles reflects the findings of Kim *et al* (2016), who interviewed practitioners working across 8 Microsoft organisations, investigating the emerging roles existing within data science. Although the terms "data science" and "data scientists" encompass a broad and often vague range of roles (Muller *et al* 2019), the data scientists interviewed for this study largely had backgrounds in AI, statistics, and computer science.

All the participants I contacted (which could include knocking on their door if I was at the lab) agreed to interviews. Several of them also asked others in their teams to speak with me and suggested other practitioners who they believed would have pertinent experience (participant-referral sampling), which kept interviewees suited to the desired criteria. These facets of my sampling approach minimised selection bias, as participants were not self-selected due to interest in recruitment advertisements. However, as with the demographic and linguistic limitations discussed earlier, this perhaps presented a limitation in terms of the breadth of perspectives represented.

Twelve practitioners were chosen due to their expertise in their domain of practice (key knowledgeable sampling; Patton 2014), selected from across several domains. Following this, I recruited via a combination of participant-referral sampling and sharing targeted emails to specific start-up clusters. I stopped recruiting new participants when I reached theoretical saturation (Robinson 2014), that is, when the same subjects recurred sufficiently in different interviews. To elaborate further upon this, although all participants contributed to this theoretical saturation, not all are represented in direct quotes in the body of the thesis. Rather, I have chosen excerpts which best represent the themes at hand.

Four participants were founders of AI start-ups, start-ups which they had set up after finishing their PhD. Of these, there was one medical start-up, one finance start-up, one accessibility start-up and one start-up which worked on military and rail applications of computer vision. Three of my participants held senior roles in academic labs, whilst also consulting for industry. Five were senior staff in multinational tech corporations, their roles including leading research projects, research groups and even being on the board

of research labs, with two holding senior roles in product divisions. Similarly, three participants were senior staff in multinational consultancy firms which offered products using AI. The remaining seven participants held mid-career roles, for example researchers in the labs run by other participants, or team members in industry roles. They had been in the field for decades and attained their position as an expert through experience.

In total, 22 participants were included in this study, with their roles summarised in Table1 below. These roles employed a variety of approaches to AI[3], including supervised, semi-supervised and unsupervised approaches (supervision refers to whether a human has directly overseen classification of the data which the model learns from). Practitioners described their subfields as reinforcement learning (training an AI model based on a "reward" function), deep learning (such as neural networks), probabilistic ML, Data Science, and combinations of these (the exact details of their work are kept fuzzy in order to protect anonymity). None of the participants were trained in ethics beyond the requirements of their lab or workplace. I pseudonymised the participants using an online random name generator, ensuring that none of the names used are those of participants.

Table 1: Participant Information

No.	Participant	Role	Sector
1	Martin	Start-up Co-Founder	Both
2	Cameron	Head of Research Lab	Both
3	Stefan	Head of Research Lab	Both
4	Julie	AI Researcher	Both
5	Lukas	AI Researcher	Both
6	Jason	AI Researcher	Academia
7	Adrian	AI Researcher	Industry
8	Thomas	Head of AI Research	Industry
9	Alec	Head of AI Operations	Industry

[3] As mentioned in Chapter 1, for the purposes of this thesis I largely use AI to refer to the subfield of Machine Learning

10	Dewi	Start-up Co-Founder	Both
11	Joshua	Start-up Co-Founder	Industry
12	Luke	Lead Data Scientist	Industry
13	Eddie	Head of Data Science	Industry
14	George	Head of Data Science	Industry
15	Ross	Data Scientist	Industry
16	Skye	Data Scientist	Industry
17	Ariel	AI Consultant	Industry
18	Lorenzo	Start-up Co-Founder	Both
19	Mark	AI Consultant	Industry
20	Alice	Researcher	Industry
21	Kristoff	PhD Researcher	Academia
22	Louis	Researcher	Academia

Observational Study

Complementing the themes which I constructed from the interviews, I also conducted a focused observation week, where I shadowed a participant, pseudonymised for the purpose of the thesis as Remi, an intern at an industry research lab based in the United Kingdom. This ethnographic study included interviews at the start and end of the week.

In the internship, Remi was working on developing an AI model which could identify everyday items from observing only a few examples in real-world settings, as part of a larger assistive technology project. She had two avenues of focus during the time I was present; firstly, making her AI model more efficient such that it could be trained with less Graphics Processing Unit (GPU) memory, and secondly, acquiring GPUs with more memory. I stayed near Remi for the duration of the study, sitting by her desk while she worked, eating lunch together, and attending most of her meetings. My observations were primarily recorded in the form of jottings, which are notes written down as the observed events are occurring (or immediately after), to preserve as much detail as possible. These jottings were written primarily using the Apple Notes app on a MacBook, with the exception being during meetings where I could not take my laptop and instead used a notepad and pen. I also interviewed the participant about their background and

day-to-day work, using this information to draw up a participant profile, and to complement the notes. I discussed my observations with Remi at length, both verbally in a Zoom meeting, and by sending a draft for her to comment on, incorporating her feedback into the manuscript. This feedback process enabled a better interpretation and contextualisation of the recorded observations, which support the research aims.

Although the time period for the study could be seen as limited, it served its purpose of allowing a deep dive into a project which I had already been embedded within throughout my doctoral research and enriched my insights into the emergent themes of the interviews. Another potential limitation was the high-level approach taken to describing material practices – I focused on broader themes rather than specific details. While this is more suitable to the scope of my thesis, I recognise that this could be a limitation on the depth of analysis. I had also planned and organised workshops with the aligned industry lab, workshops which aimed to further validate findings and investigate tools of ethical practice. These were scheduled for March 2020. Unfortunately, the changes to societal organisation caused by the COVID-19 pandemic put a stop to this plan. While additional data from these intended workshops may have provided additional insights or perhaps lenses of interpretation or points of interrogation, the data collected from the interviews collected prior to this interruption sufficed for investigating the research interests. Furthermore, a core intention in designing these workshops was to pilot an interactive ethical tool, however, the 2020 COVID lockdown provided time to dive deeper into the philosophical conceptual tools I present in this thesis, leading in a new direction which I ultimately feel provides a much-improved contribution to the literature on AI ethics.

Interview Procedure

I conducted 22 interviews, lasting between 38 and 67 minutes (median 51), between November 2017 and May 2019, with these recorded either as face-to-face interviews using an audio recorder or via a video-conferencing platform (Skype). As discussed, I conducted semi-structured interviews, drawing upon the expert interviewing method, and informed by my analysis approach of Reflexive Thematic Analysis. Each interview

was comprised of three stages, though these differed in length and focus according to the participant.

- **Background and motivations:** An introduction during which I asked for an overview of their background and role, and their motivations for their role and domain.
- **Practice and work context**: this section was more loosely structured, consisting of a set of prompts around data, methods, and context.
- **Perspectives and suggestions**: in this part of the interview, I aimed to elicit participant views on responsibility in AI ethics, approaches to ethics, and what they believed were the most pressing challenges facing these areas.

An important note on the terminology of ethics is that although practitioners sometimes initially understood the term 'ethics' as equivalent to compliance, over the course of the interviews, their deeper ethical practices and motivational values became clear. Many of the practitioners were interested in moral philosophy, and understood ethics as the study of this, even if they did not have much deeper knowledge.

Analytical Process

All the interviews were recorded and transcribed manually. The first set of transcripts were then manually coded, so I could familiarise myself with the data. The remaining interviews were coded in NVivo. The codes were validated by a second coder, and we also held regular discussions about the progress of the analysis.

The nature of the interviewing and analysis process taken is co-constitutive, with the integrity of the outputs lying in my own reflexivity in the biases I have brought to the work, and openness about the journey I have taken in my research. That, is, recognition that "meanings are seen to be negotiated between researcher and researched within a particular social context so that another researcher in a different relationship will unfold a different story" (Finley 2002, p. 531). In order to be transparent in my process, I have included numerous quotations in this manuscript, and provided a description of my

coding process in the above section, in addition to further examples of templates and maps in the appendices.

To structure my approach to coding and analysis, I employed Template Analysis (TA) (King 2012). TA is an inductive variant of thematic analysis involving reflexive analysis (Nowell *et al* 2017), in which some a-priori themes are identified early on, but these are then iterated upon, enabling emergent themes to be included in the analysis over the course of the analytic process. TA facilitates exploration of the beliefs and values of participants whilst steering clear of a specific epistemological alignment, working as a technique rather than a methodology. This allowed for a flexible approach which enabled engagement with different theoretical frames (broadly clustered around phenomenology) across the process of constructing analyses, particularly important given the intertwined, multifaceted influences set out earlier in this chapter. I employed the epistemological approach of interpretivism to understand the subjective experiences and perspectives of participants, via interpretation of the meanings of their words and behaviours (Goldkuhl 2012), and the "processes by which these meanings are created, negotiated, sustained, and modified within a specific context of human action" (Schwandt 1994, p. 225).

Research Ethics

Given my chosen research method of qualitative investigation, ethics formed a crucial part of my own research as well as being my topic of study. Before conducting each study, I submitted a risk assessment and consent form to the ethics board in my University department, only proceeding once these were officially approved. In accordance with the research ethics procedures at the Edinburgh College of Art, I shared an information sheet with participants which provided an overview of the study which included research aims, focus, potential outputs, data storage and additional information pertaining to the General Data Protection Regulation (GDPR) act (including rights regarding storage and deletion of data), and relevant University contacts should these be needed. It specified that the audio data and transcripts would be anonymised and viewed only by the primary researcher and second coder, stored securely on University servers before being deleted. This helped ensure that I could obtain their informed consent. To further

facilitate the agency of my participants in consenting to the study, I gave some background about myself and asked if they had any additional questions before beginning interviews. The audio and transcripts of these interviews were stored securely according to the Data Protection Act 2018, GDPR regulations, and the University's own data policies, on a University server.

To protect my participants' anonymity, I have assigned them all pseudonyms and changed any details likely to signal identity (for example, as provided in anecdotes), which both maintained their privacy and enabled greater freedom to give critical perspectives and frank descriptions of experiences. Details regarding the interview participants have been purposefully obfuscated, with careful consideration paid to how I could best minimise the information shared about them whilst maintaining enough richness for the thesis findings.

Researcher Positionality

I would hesitate to call myself an AI practitioner, but I do have some limited background in the field. Following a foray into computational linguistics as an undergraduate student, I had a very brief career in the informatics department of a hospital, moving from data entry to data analysis before deciding to undertake a Masters programme in Artificial Intelligence. More precisely, I began a Masters in Cognitive Science, specialising in Natural Language Processing (NLP). In the latter half of my Masters, a recruiter from a Big Tech company gifted me with what turned out to be one of the best pieces of career advice of my life – "change your degree title to Artificial Intelligence, and you will find you are recruited much more easily". He was very right. Following my Masters, I interviewed for several seemingly swanky, superficial start-ups ("we want to use people's social media info to help decide whether they should get a loan or not"), and Big Tech companies (one of the interviews triggered a panic attack). In the end though, I returned to the hospital, which had recently set up a Data Science unit of its own in collaboration with the local university's AI research group. I loved that job. I worked on all stages of the pipeline – decision-support form design, database architecture, data cleaning, analysis, and visualisation, and even some NLP. I worked closely with a variety of stakeholders

including physiotherapists, finance officers, medical consultants, and AI practitioners. I was fascinated by the potential to apply AI to datasets to speed up and streamline medical diagnosis. Seeing these models being touted in the healthcare domain as solutions was concerning however, given my knowledge of the highly specific and brittle nature of such models, the potential for induction of existing medical bias into such systems, and the marketing of heuristic-driven craft as reliable science. After a short-lived move into a Computer Science PhD, whilst teaching health informatics to my own Masters students, I began the research laid out in this thesis, having fully realised my interest in the social and ethical implications of this sort of high-stakes work.

Over the course of the PhD research, I was involved with the research lab in three roles: as PhD student member of a project, an intern researcher, and as observer in the observational study. In initial meetings of the lab team, I would contribute to discussions and occasionally contribute talks, and I also contributed in this way during my brief internship at the same lab. In contrast, in my role as observer during the observational study, I was explicitly present to observe the team, and tried to constantly maintain consciousness of my own position relative to the people I was shadowing, to avoid presumption on my part and altered behaviour on theirs.

My journey and perspectives are intertwined with neurodivergence and Ehlers Danlos Syndrome (EDS), which have informed my understanding of the situated nature of both knowledge and knowledge production, to draw on Donna Haraway (1988). In terms of research perspective, Attention Deficit and Hyperactivity Disorder (ADHD) is associated with differences in time perception, experienced as diffused and ongoing rather than linear and stepwise. This had synergies with certain findings, contributing to my fondness of de Beauvoir's work on *the Ethics of Ambiguity* (1947) which seemed equally pertinent to the ambiguous world of AI. Similarly, my experience of epistemic injustice in dealing with the chronic health issues caused by EDS enhanced the critical lens through which to consider applications of AI.

This consideration of situated perspectives fundamentally links with a crucial point of reflection – that my PhD research has been funded by an organisation which is influential in the field. In my experience, the lab was both supportive and hands-off. That

is, I was given opportunities to engage, requests for interviews and other ethnographic opportunities were always met with support, and my research direction was not dictated or questioned. However, I have anonymised my work and it does not centre on this organisation at all.

Equally, my identity as a white British person implicitly informed this research. Amongst other considerations, it contributed to the ease of my research access, facing no friction by virtue of my race and citizenship. It also informed the scope of my sample, focusing on practitioners working in organisations in Anglophone countries with cultural proximity to the U.K. While this creates some limitations to the scope of the thesis findings, it nevertheless presents important considerations which are applicable to other contexts, given the widespread and global scope of AI practice and applications, which are considered in the next chapter.

CHAPTER 3: AI Ethics in Context

Just a year after the Cambridge Analytica scandal, which implicated Facebook in allegations of voter manipulation following evidence of mass harvesting of profile data (Cadwalladr and Graham-Harrison 2018), Amazon abandoned development of their recruitment algorithm due to discrimination against female applicants (Dastin 2018). Perhaps the company felt that the risk of bias impacting equity in decision-making outweighed other considerations or were concerned with the impact of negative publicity given contemporary Big Tech scandals. Either way the project was discontinued, providing a touchpoint for consideration of how values such as efficiency and equity influenced the algorithm's development and use.

Values are enmeshed within AI technologies, whether imparted by marketing motivators, the shadow of historic decisions in the development of the field, or influence of the demographics who design and develop AI. This chapter examines the relationship between the values implicit to AI, the contexts of its design and development, and the ethical responses to these. It traces the historical evolution of AI values and ethics together with contemporary accounts of AI practice, then explores the different ways in which AI ethics is currently being approached. In doing so, I aim to illustrate the importance of understanding underlying values, motivations, and historical contexts in developing ethical recommendations.

Research on ethical AI has often focused on abstracted dimensions such as fairness, accountability, transparency, and bias, with the implications of these prompting attempts at unified methods for the general embedding of ethics in AI workflows. Much of the work explicitly labelled as addressing ethics takes an approach influenced by bioethics, in the same vein as parallel discussions of ethics in related fields such as interaction design or software engineering. Yet the field of AI has incorporated several ideals and epistemic assumptions within the course of its development, and an awareness of these is necessary for grounding understandings of the values that emerge in contemporary discourses.

The Evolution of AI (and Computer Ethics)

The development of AI has been shaped by various socio-political contexts, also intertwined with the evolution of Computer Ethics. In this section I sketch out this history and illustrate the current layout of the field.

A Potted History of AI

The most common narrative history of the field goes roughly as follows. In the wake of Alan Turing's 1950 publication *Computing machinery and intelligence*, the term "Artificial Intelligence" was coined in 1956 at the *Dartmouth Summer Research Project on Artificial Intelligence (DSRPAI)*. At this summer school, Marvin Minsky and John McCarthy brought together experts from different fields with the aim of creating machines which could emulate human intelligence (Kaplan and Haenlein 2019). The intervening years have seen the AI hype cycle repeat itself in periods of AI hype followed by periods of limited funding due to failure to live up to this, known as AI summers and winters (Gonsalves 2019). However, as Mendon-Plasek (2021) points out, narratives such as AI winters and summers are created with the goal of foregrounding values seen as desirable, and to market the field. These narratives reframe constraints and contingencies as virtues and intentions, at the same time very purposefully including and excluding certain actors.

A clear example can be seen in efforts to construct a professional, legitimate identity for the field. The sub-field of pattern recognition, which dominates contemporary AI, was purposefully described as data-led rather than human-defined, sold as developing "mechanical identification of significance and the reproduction of human judgment" (Mendon-Plasek 2021, p. 32). In contrast with symbolic representation in which the practitioner sets explicit rules, practitioners framed pattern-recognition as a more flexible, robust, and even objective approach to AI, by focusing on how it 'allowed' the "phenomenon of interest be defined by the statistical properties of the data" (Mendon-Plasek 2021, p. 33). In this vein, the problem of incomplete data was

transformed into the "epistemic virtue of greater contextual sensitivity" (Mendon-Plasek 2021, p. 34), implicitly equating pattern recognition with ethical behaviour.

Such careful consideration of the ethical implications of practices and outputs within AI – and the underpinning practitioner values, often shaped by cultural and political contexts – has a long history which stretches back as far as the history of the field itself, and the recent uptick in interest in AI has been paralleled by an uptick in interest in what is explicitly termed AI Ethics. The proposed capacities of AI (and ethical concerns connected to these) have varied based on how AI itself is understood, often with vague and differing conceptions of "intelligence" which served to popularise the field while creating a lack of consensus on its aims or defining features, producing a fragmented disciplinary domain which holds a multiplicity of practitioner and publicly espoused values (Wang, 2019). This feature of the field has influenced the development of pertinent ethical approaches and considerations, which I consider below.

Values, Applications, and the Emergence of Computer Ethics

After a period of drastically reduced work in the early 1970s, the late 70s and 1980s saw another rise of popularity of expert systems, responding to the new increase in AI funding, and the "grandiose expectations" (Leith 2016, p. 96) of the developers of such systems, in addition to the improvement in computational capabilities. This time AI broke into the commercial domain, seeing a proliferation of Expert Systems. In response to this, Computer Ethics (CE) arose as a more applied approach to the ethics and philosophy of AI.

Deborah Johnson developed an ethics curriculum for computer science students, explicitly naming its focus as Computer Ethics (Johnson 1978). Meanwhile, in *Four Ethical Issues of the Information Age*, Richard Mason proposed that among many ethical concerns, four stood out as most salient - Privacy, Accuracy, Property and Accessibility (Mason 1986). In a broader critique of the domain, Mason also compared the rise of computational technology with the first industrial revolution, in terms of the level of change which would be wrought. He reflected on how "practitioners of artificial intelligence proceed by extracting knowledge from experts, workers and the

knowledgeable and implanting it into computer software" (Mason 1986, p. 9), reminiscent of current concerns about the role of workers in the data pipeline of AI, as tools for knowledge extraction who are undervalued (Sambasivan et al, 2021).

Meanwhile, James Moor set out a need for an ethical framework for computational technology, proposing Computer Ethics as a solution, with several approaches and frameworks evolving from this (Moor 1985). The concerns Moor set out correspond closely to many of those set out about AI, and included in this manifesto was a critique of what Moor identified as the most pressing ethical concerns introduced by computer technology, alongside an explanation of why Computer Science should be considered as a unique form of technology with a corresponding unique set of ethical concerns.

Moor pointed out that although CE might seem like an obvious case of applying an ethical theory in policy, the lack of policies is paralleled by a lack of appropriate concepts and concept-building. Moor also points out that in considering different approaches to ethics, and accepting or rejecting these, the decider demonstrates their own values. Furthermore, as new technology develops, the original values must then be re-examined. Pointing to the ubiquitous nature of emerging technologies, and their capacity for permeating and transforming human activities and institutions, Moor argued for CE as a distinct field of ethics, becoming not just a part of these activities and institutions, but essential to them. Rather than looking at the role of computers in work, the question shifts to "what is the nature of this work?" (Moor 1985, p. 271). Moor was reflecting on the nature on how non-technical roles become performed by computers in tandem with humans, but technology is as much of a formative factor in the practice of those designing technologies; not just the physical computers they use, but the nature of the conceptual materials drawn upon. In the same way that embedding of computers in education might prompt the change of question from "how does technology affect education" to "what is education", as Moor argues, the types of AI currently worked with perhaps prompt the question "what is an ethics of AI practice?", beyond just "how do we apply ethics to AI practice?".

Another argument for the importance of CE cites the "invisibility factor" (Moor 1985, p. 272), or specially three types of such factors. The first is that of "invisible abuse" (Moor

1985, p. 273). One example of invisible abuse is taking advantage of relative invisibility of computational mechanisms in an unethical way, due to lack of access to or comprehension of the workings of a program, which can easily be imaged as the issue of opacity in AI practice. Another example is "invasion of the property or privacy" (Moor 1985, p. 273), or using computer technology to surreptitiously alter information or gain information which is not knowingly agreed to. A third concern was that of surveillance, one which is a major and commonly expressed concern regarding AI (Feldstein 2019), repeating the same concerns as Moor regarding workplace surveillance, just with greater surety on the specific mechanisms of its enactment (McStay 2020).

Moor also expressed concerns over the "invisible programming values" (Moor 1985, p. 273) which can become embedded in systems. He used the metaphor of building a house from specifications, where no matter how specific the instructions may be, inevitably value judgements are involved about matters which have not been specified. He also used the specific example of a flight booking system which is biased to direct the user to certain flights. Moor points out that programmers may even be unaware of the values which they bring to the systems they develop, for example the bugs which remain invisible until a time of crisis.

Similarly, unrealistic expectations might occur as an accidental impact of adopting practices from other fields, for example, the case of practitioners adopting the habits of programmers in other domains such as software engineering. McDermott (1976) discussed how this, in the past, contributed to miscommunication, when practitioners used mnemonics as naming devices for AI projects and data structures, 'wishful mnemonics'. He pointed out that although this practice fit with a style of work which was much more structured and bounded, it did not suit the experimental and open nature of AI, and indeed resulted in misunderstandings on the part of on-lookers (and even the AI researchers themselves). In *Computer Power and Human Reason*, published in 1976, Weizenbaum provided a real-world example of this with the example of how his provocation ELIZA, a program mimicking a psychotherapist, was received seriously by users and press (Weizenbaum 1976). In the words of McDermott, AI "programs to a great degree are problems rather than solutions" (McDermott 1976, p. 4), however, to an

outsider this is not obvious. A third type of invisible concern is the "invisible complex calculation" (Moor 1985, p. 274*),* refers to the construction of calculations which are so complex that they are beyond human comprehension, again reminiscent of current debates around opacity in AI systems, which introduces issues of trust in output. Moor concluded his reflections with the observation that invisibility privileges the value of efficiency at the cost of many ethical considerations.

The Rise of Silicon Valley and Big Tech

Intertwined with AI, the rise of Silicon Valley since its origins in 1969 have transformed Palo Alto from a tiny manufacturing hub to a global centre of technological investment and innovation (Berlin 2017). Today it forms the hub of the information economy, to the extent that in "Silicon Valley" we find another suitcase term, "a cocktail of abstract ideas" (Jones and Sudlow 2022, p. 1121), which itself has spawned neologisms such as "siliconisation", which serve to globally reinforce and redeploy the logics of the Valley (Jones and Sudlow 2022, p. 1122). "An economic space built on social capital" (Cohen and Fields 1999), Silicon Valley houses Meta, Alphabet and Apple, three of the highest valued companies globally and part of the Big Tech group of dominant technology companies (Tarvier 2022). Within Big Tech, there has been a race for leadership, particularly in the domain of generative AI. Most high-profile are Meta, Alphabet and Microsoft, alongside OpenAI which lists many Big Tech and Silicon Valley names amongst its founders, funders, and researchers (Verma 2023). In addition to funding their own in-house AI research teams, Big Tech companies have moved to explicitly place massive investment into leading AI organisations, such as Microsoft's $13 billion investment in OpenAI (Dastin 2023). However, this differs according to the organisational strategy of the company. Apple has traditionally been cautious about its offerings, holding back from public narratives and minimising the risks inherent to emergent technologies, and this holds true for their AI development although current products demonstrate AI capabilities (Bajarin 2023).

Examining the Contemporary AI Sector

Over the past decade, Artificial Intelligence has seen an incredible leap in funding and application in many domains. Marketing contributes heavily to this, with developers employing various values to sell its numerous applications, such as providing novel insights into real-world problems (Zhou 2018). In healthcare, numerous algorithms have been developed which are presented as improving efficiency of diagnosis and providing better outcomes, such as breast cancer screening (Pisano 2020). An entire subfield, AI for Good/AI for Social Good, markets its AI development using appeals to moral values such as sustainability (Huntingford *et al* 2019) or focusing on reducing inequity (Kolenik and Gams 2021). Consultancy firm Price Waterhouse Coopers estimated that in 2020, AI contributed $15.7 trillion to the world economy, stating that it is the biggest factor affecting development of many sectors such as medical decision-making, hiring practices, entertainment, education and even art. AI adoption in business is up as much as 270% from 2015, accompanied by a shortage in AI practitioners, with a recent report estimating 478, 000 people working in AI roles.

However, this broader picture of growth has not translated to social equity within the field. A 2020 report from the British Computer Society (BCS) investigating the diversity characteristics of the UK workforce found that despite increasing overall numbers, women are still severely underrepresented in the field, a relatively unchanged proportion between 16 and 17% from 2015 to 2019 (BCS 2020). This gender disparity was further worsened by precarity, as women were more likely to be employed in part time positions, and on a non-permanent basis, which had implications for facing cuts to employment as were seen during the COVID-19 pandemic. There are also significant disparities due to ethnicity and disability in AI. The same report found that Black and Minority Ethnic workers (BME) in technical roles at technology companies made up 18% of the overall domain in the UK, while the prevalence of disabled workers in technology companies was estimated as 1 in 10 (BCS 2020). Reflections from disabled employees shows that they can feel unable even to disclose their status for fear of repercussions (Thomas 2018), suggesting that disabled people remain underrepresented in AI development. AI might not be developed by individuals in these demographics, but its

impact is seemingly inescapable for them. Some of the recent concerns include hiring algorithms potentially discriminating against disabled workers (Wall and Schellman 2021), and face recognition tools are far less likely to be accurate on darker skin (Buolamwini and Gebru 2018).

Furthermore, the labour which makes up AI pipelines consists of people beyond AI practitioners, hidden labour such as crowdsourced workers who work on crucial tasks from data labelling to platform monitoring (Gray and Suri 2019). Many of these roles are essential for the creation and maintenance of AI outputs, yet remain invisible, in another cycle of the extractionism described by Mason (1986). The recognition of these inequities within the constitution of practice primes considerations of how and for whom AI is developed, given that domain expertise is created, drawn on and redrawn by individuals of certain demographics whose concerns may preclude those rendered invisible by the systems within which they operate. In effect, the AI sector in its current state is poised to builds on and exacerbate epistemic and ethical injustices – especially given its continuous proliferation – despite any framings of objectivity (largely due to its obscured technical and socio-political constraints) or intentions towards "social good". To further understand this, I consider the domains of AI practice as a microcosm within which socio-material dimensions influence outputs.

AI in the Lab

A handful of studies detail qualitative investigations of the practices of AI labs. In an ethnography of two University AI labs based in the US, Hoffman (2017) investigated the practices of one lab which focused upon gaining commercial funding, and another which aimed for federal scientific funding. This study took a birds-eye level view of the lab, looking at the approach which lab heads took in identifying strategy, approaching funding, and conceptualising the goals of the lab. Responding to the constraints of their sources of funding, both labs demonstrated short-sightedness of their research programs, surfaced in different ways. Hoffman identified three sources of ambiguity, pervasive from the level of lab management to AI practice; ontological ambiguity (what is nature of the object we are researching?) which required translating a conceptual

object into a practical reality, epistemological ambiguity (how do we know if our knowledge claims are correct?) due to inconsistency in evidencing knowledge claims, and applications ambiguity (how can we know if this will be useful or not?), concerning the uncertainty as to the value of a method until it is created (Hoffman 2017, p. 713).

Other studies have investigated how practitioners conceptualise specific values. These focused on investigating values tied directly to the behaviour of the systems they design, rather than individual values that motivate the practitioner to conduct this work. Investigating how values can be actively incorporated into system design, Shilton conducted an ethnography of a research lab to investigate the values which data scientists referenced and how they negotiated them in situ (Shilton 2013), identifying values such as privacy and consent as viewed as most important.

Veale *et al* (2018) interviewed 29 AI practitioners working in public services, investigating their perspectives on the challenges of incorporating public values into their work, focusing on fairness and accountability. They found that practitioners were not only keenly aware of many challenges surrounding fairness and accountability, but that they also attempted to engage with these ethical issues, finding some tensions nearly impossible to reconcile. While ethics frameworks or principles include guiding values, practitioners are expected to navigate the numerous tensions required for translation of abstracted values into practice. This illustrates the limitations of proposed institutional values and highlights the importance of investigating practitioners' own values and situated practices, especially if practitioners are expected to guide their actions in situations where ambiguities are encountered.

In a survey comparing how different demographics rated the importance of certain values, Jakesch *et al* (2022) reported a difference in the emphasis which AI practitioners placed on the importance of values when compared to the general public and crowd-workers (not further specified). The authors asked participants to rank values taken from the Schwartz Value Survey (SVS), a standardised values inventory with a background in moral psychology. Following this they asked participants to rate the importance of binary pairs of SVS values, in a specific AI-related scenario. The three highest-rated values amongst AI practitioners were respectively Privacy, Fairness and

Safety, while the other two groups valued Safety, Performance and Privacy (Jakesch *et al* 2022, p. 315). This discrepancy perhaps indicates a focus within the AI field on Fairness, which I consider in greater depth in Chapter 4. Meanwhile, other scholars have investigated the values which can be inferred from analysis of AI research outputs (Birhane *et al* 2022), and ethical tools (Wong *et al* 2022). Birhane *et al* (2022) identified the most common values referenced in AI papers as "Performance, Generalization, Quantitative evidence, Efficiency, Building on past work, and Novelty" (Birhane *et al* 2022, p. 173). Placing these values in context can provide more insight into the intentions and institutional logics underlying the expression of such values, which I investigate more in Chapters 4 and 6. In addition, the domain of "building on past work" is noteworthy considering the theoretical and socio-material circumstances within which the field itself has been defined and developed – any extensions to established ways of working inherently extend existing epistemic and practical issues, and a focus on this value could reduce the scope for addressing issues such as epistemic injustices.

However, there has been a sparsity of research investigating which types of intrinsic values draw AI practitioners to their work. Looking at the broader scientific domain, Lounsbury *et al* (2012) argued that scientists were more motivated by intrinsic (self-motivated) values than extrinsic (external) ones. Fleischman *et al* (2010) used surveys, focus groups and structured interviews to investigate the social, cultural, and moral values of computational modelers in a research lab. This provided insights into which values were likely to be held by those who had read codes of ethics, and those who believe in the usefulness of codes of ethics. Practitioners who placed importance on following a code of ethics scored higher on values such as equity, whilst ethical frameworks had an impact on the culture of the lab (Fleischman *et al* 2010). Fleischmann *et al* (2010) also observed that the level of success of the computational model was contingent on common values between the modeler and end-user. This has implications for situations where values are discordant or unknown, a common issue given the ambiguities surrounding how AI could be applied, and an important consideration given the global market of AI where technologies developed in one context often cross socio-political boundaries.

These findings touch on two important focal points of this thesis, mediation and situated knowledge, indicating the impact of the nuanced interaction of the two upon surfacing of values, and engagement with ethics. There are common emergent themes among all the studies discussed above; values such as "efficiency" and "safety" arise again and again. However, as mentioned in Chapter 2, these approaches do not allow for emergent values, and cannot account for differences in how values are perceived according to the cultural, political, and material contexts of the people investigating them. In addition to accounting for the invisible workers, we also need to be able to understand the invisible programming values informing practice. Otherwise, the insights gained from mapping values remain limited by their abstraction from context.

Approaches to AI Ethics

Often aiming to shape outputs according to the types of values described above, numerous approaches to ethics have arisen in response to the proliferation of AI. The dominant narrative of principles has been criticised as stifling actual ethical deliberation, whether that be because the condensing of debate into static principles represents the antithesis of ethics rather than engagement with it (Resseguier and Rodriguez 2021) or that it constitutes "ethics-washing" to distract from the push for genuine change or external regulation (Wagner 2018). However, this only forms one approach to governing AI. There have been calls to move the focus of ethics from principles to practice (Schiff *et al* 2020) and recognise the limitations of ethical frameworks (Mittelstadt 2019; McLennan *et al* 2020).

This section makes a distinction between what I term *top-down* approaches to ethical governance, and *bottom-up* perspectives. Top-down refers to ethical frameworks or the inclusion of members of a design team to whom ethical responsibility is delegated. Alternately, bottom-up approaches are centred in practice and education, for example using tools such as Shilton's values levers, activities which prompt discourse around specific values (Shilton 2013). Values levers are "practices that open new conversations about social values and encourage consensus around those values as design criteria" (Shilton 2013, p. 1), also described as "practices that pried open discussions about values

in design and helped the team build consensus around social values as design criteria" (Shilton 2013, p. 3). These explicated contemplations of the values which practitioners may embed within their work serve to prompt considerations the repercussions of the design processes and potential outputs. In the absence of such reflections, taken-for-granted beliefs and values often become replicated in the products of technological endeavours.

Exemplars, Education and Ethics from the Bottom-Up

Bottom-up practise refers to the involvement (or, more commonly, illusion) of technological practitioners in ethical engagement. Ethical training might happen in professional practice or increasingly in university education; the history of ethics education in general computing stretches back nearly as far as researchers published reflections on the topic, to 1972 when Nielsen published the paper *Social Responsibility and Computer Education* (Nielson 1972). To this day, Computer Science courses offer standalone ethics modules. However, these standalone ethics modules have proved ineffective in other disciplines, such as Business Studies (Loescher *et al* 2005) whereas integrating ethics into the overall teaching of Business Studies material improved students' awareness of ethical issues and their ethical decision-making abilities (Dzuranin *et al* 2013). Reflecting this, recently there have been calls to integrate teaching of ethics into existing AI courses (Saltz *et al* 2019), and numerous courses developed and critiqued (Fiesler *et al* 2020; Fiesler *et al* 2021; Raji *et al* 2021).

There is a rich Human-Computer Interaction (HCI) literature which considers the role of technology in supporting values, offering up mechanisms for explicit consideration of these values in processes of design (Ackerman and Cranor 1999; Friedman and Nissenbaum 1996; Millett *et al* 2001). This literature introduces some key concepts such as the importance of anticipating the impact of technologies, and that designers be reflexive regarding their role in imbuing qualities which are seen to reinforce core values such as privacy or autonomy. However, HCI approaches such as Value-Sensitive Design (VSD) tend to elide consideration of the multiplicity of possible values (Borning and Muller 2012), the most immediately obvious materialisation of this being prioritisation of the values of those 'in the room', the designers and other

immediate stakeholders, without considering the cognitive biases which influence these (Umbrello 2018) and positing the concept of universal values (Friedman *et al* 2002; Friedman and Kahn 2007). Such approaches can also neglect the implications (both positive and negative) of epistemic values represented in the complex interplay between the system developed and the assumptions of the designers and practitioners designing it. Manders-Huits (2011) points out that taking values at face value "runs the risk of committing the naturalistic fallacy, i.e., by reducing an 'is' to an 'ought'" (Manders-Huits 2011, p. 280).

Another approach to bottom-up practise is the moral examplar or ethics owner (Shilton and Anderson 2017; Moss and Metcalf 2020), project team members who have an excellent history of ethical reflection and decision-making. Shilton and Anderson (2017) examine holistically arising exemplars, touching upon how ethical skills can be acquired via training. Meanwhile Moss and Metcalf (2020) describe the role of ethics owners as intentional roles created within a team. Either way, the outcome is the same, a member of the team who takes the additional role of moral leadership in a project. The advantage of this approach is that moral practice exists seamlessly alongside technical practice, and those with the most knowledge of the systems being built are also leading discussions around the moral implications of work, and the implementation of ethical tools. Another version of this is the inclusion of values advocates on design and development teams. These are people with training in ethics and values who can advise the wider project on social and ethical considerations, drawing from their depth of education and experience in the area. Wynsberghe and Robbins (2017) see this as a necessity, "to say that either the ethicist can do engineering with minimal instruction, or the engineer can do ethics with minimal instruction would be an undervaluing of either/both disciplines. Secondly, the skills necessary for ethical analysis are not necessarily transferable from one project to another for the engineer whereas this is precisely the kind of training the ethicist will have" (Wynsberghe and Robbins 2017, p. 10).

However, the ethicist on the team is subject to the same knowledge constraints discussed for participatory design, which may limit their ability to contribute. Joining as

an outsider can be beneficial to the role though, by providing new perspectives on the culture and values of a team (Shilton and Anderson 2017, p. 74). The outsider may be equally restrictive in their contribution of values, however, especially with regards to the values of minoritized groups and international cultures (Borning and Muller 2012; Alsheikh *et al* 2011). In addition to cultural distance, physical distance also has an impact upon the integration of the advocate into the team, with distance from the team reducing trust (Shilton *et al* 2014).

Another way of approaching ethics in system design has been incorporating it into the process itself, by involving practitioners in development and focusing it on the specific contexts of practice. Madaio *et al* (2020) sought to ground ethical principles, embedded in a checklist, in practitioners needs, focusing on principles which addressed issues of fairness. They used interviews and co-design workshops to investigate the practitioners' perspectives. The study found that practitioners often saw fairness as an issue of personal importance, alongside being critical for maintaining the reputation of the organization they worked at. Despite the importance of fairness to the success of the company, practitioners noted that it was down to individuals to act to ensure fairness. The fast-paced nature of practitioners' work stood in tension with desire to embed fairness into their models, resulting in 'social cost' for the individuals who did decide to advocate, working as an organisational barrier. Madaio *et al* (2020) noted that the individual action inherent in these processes resulted in an ad-hoc approach to ethics. Practitioners suggested checklists as a useful method for explicitly integrating fairness into their work if these were matched to the existing workflows. Checklists might make practitioners freer to challenge development by providing risk-free opportunities to raise concerns, prompting critical discussions around fairness ramifications. Certain practitioners were concerned however that a checklist would be yet another compliance list to deal with and trivialise fairness concerns. Finally, the importance of organizational culture was highlighted, with the authors finding that checklists would need significant modification between organisations in order to account for culture and goals.

Other bottom-up approaches include encouraging practitioners to use narrative cell tools in code development environments such as Jupyter notebook (a web

application for creating and sharing computational documents, which could be used for process audits and ethical review), although there is evidence that such tools are underutilised, partly due to the exploratory nature of data analysis resulting in "messy" notebooks (Rule *et al* 2018). Compared to the top-down approaches of frameworks, principle-sets, guidelines and so on, normative bottom-up ethics has had a paucity of literature, and the ones that did exist often suffer from a lack of usability due to difficulties translating them into practice (Morley *et al* 2021). This thesis addresses this gap in literature by investigating the values motivating practitioners involved in AI work "on the ground", the contexts constraining them, and the role which ambiguity can play in ethical governance.

Frameworks, Principles, and the Top-down View

Ethical frameworks represent one of the key top-down approaches to ethical decision-making and governance, proving numerous enough to allow analyses of 112 such documents published over the space of three years (Schiff *et al* 2021). These frameworks were commissioned by public, private, educational and third sector institutions, finding a greater breadth of ethical considerations in public and NGO frameworks as compared to private ones. These differences in frameworks potentially reflected different underlying beliefs about an organization's responsibilities, and which viewpoints were important to include in the design of such technologies (Schiff *et al* 2021). The findings also suggested tension between values of social and economic good, an important consideration given the competitive global marketplace in which AI enterprises are situated, and the social ramifications of many AI applications.

Similarly, Floridi and Cowls (2022) conducted a broad review of frameworks, with an eye for variety although the authors noted that they focused on including high profile documents (Floridi and Cowls 2022). They only included normative principle sets, and excluded either calls that were specific enough to qualify as policy objectives, or too vague. They excluded principle sets that focused on a specific type of AI (e.g., facial recognition) but included those tailored to specific domains for example criminal justice or healthcare, finding that while the former differed in scope, the latter were consistent

across domains. The authors presented 8 themes that cut across all principle sets – "privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values" (Floridi and Cowls 2020, p. 15). These values lack meaning when abstracted from contexts and the relationships co-constituting them. Given this vacuum, they then tend to be interpreted narrowly and exclude the socio-political dimensions shaping the ethical issues which such values purport to address. Indeed, a focus on evaluating the "ethicality" of AI outputs via abstracted values can be unintentionally harmful when not qualified by examination of the socio-historical contexts shaping underlying assumptions (Bennett *et al* 2023). For example, narrow interpretation of the value of fairness "risks reinforcing existing power dynamics" (Bennett and Keyes 2020, p. 1), further marginalising the already marginalised.

Limitations of Current Approaches

This proliferation of top-down approaches presents challenges to practical ethics. Vague, abstracted principles can be difficult to adapt for application to real-world projects, particularly in absence of recognition that these "*exist in complex societal contexts, rife with biases, disparities, and ethical issues— requiring deeper commitment than a general adoption of ethical frameworks and principles" (*McLennan *et al* 2020, p. 21). There is a concerning tendency of these abstracted approaches to stymie ethical reflection, by focusing attention on the outcome rather than the process of ethics itself. Additionally, as touched upon, the assumptions underlying frameworks have pitfalls of their own, obfuscated in being transferred across fields, employed as tools of legitimisation. In this way, formal ethics can be subject to similar shortcomings raised by the epistemic agnosticism of AI practice (Moss 2021). Leonelli (2016) argues that such frameworks risk the pitfalls faced in other fields such as clinical trials and may *"reproduce the mistakes made around the regulation of clinical trials, where the delegation of all responsibility for ethical assessment to professionals has effectively encouraged the elimination of situated ethical reflection among researchers"* (Leonelli 2016, p. 3). Similarly, Resseguier and Rodrigues highlighted how the legal/compliance approach to

ethics which has dominated the domain, has several drawbacks (Resseguier and Rodrigues 2020). I briefly discuss some of these drawbacks in the following sections.

Increasing Complexity

In a field of increasing complexity and ambiguity, frameworks are often solutionist in nature, limiting their scope and flexibility. Mittelstadt *et al* discuss the ethical ramifications of the status quo of algorithms and algorithm-embedded system, positing that the continually increasing complexity of algorithms and richness of data is leading to a "gap between the design and operation of algorithms and our understanding of their ethical implications" which may lead to "severe consequences affecting individuals, groups and whole segments of a society." (Mittelstadt *et al* 2016, p. 2).

Transferring Assumptions

AI ethics frameworks also suffer from importing assumptions from other fields, primarily bioethics, again abstracting values from their contextual narratives of origin. For example, the principle of preserving autonomy draws from the rule-based bio-ethics casuists Beachamp and Childress (2001) who introduced four principles of bioethics: beneficence, non-maleficence, justice, and autonomy. Floridi and Cowls (2019) directly identified these four principles in their analysis of frameworks, and additionally argued for inclusion of a new principle of *explicability*. Meanwhile, Jobin *et al* (2019) found 11 overarching principles "by frequency of the number of sources in which they were featured: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity" (Jobin *et* al, p. 391), and no single principle which occurred in all 84 documents. They observed how the tendency to borrow from bioethics proves subpar in the domain of AI ethics because AI development "lacks (1) common aims and fiduciary duties, (2) professional history and norms, (3) proven methods to translate principles into practice, and (4) robust legal and professional accountability mechanisms" (Mittelstadt 2019, p. 501).

Recurring throughout these are verbatim bio-ethics principles, de-contextualised, and just assumed to be useful and moral. The purveyors of these

approaches often justify their choices by pointing to the popularity of these more widely, with the implication being that this means such approaches are the best. However, even within the field of medicine, these principles of bioethics have been critiqued for failing to account for the uncertainties inherent to practice, promoting an avoidance of ambiguity among practitioners who are (unreasonably) expected to anticipate and respond to unknowable outcomes with some degree of certainty (Domen 2016). Furthermore, as Evans (2000) discussed in his critique of principlism, taking this perspective disregards the impact of the sociology of knowledge, that these approaches fail because they are the best, but because they happened to rise to the top given the social conditions they were created within. Like bioethics, much of AI ethics draws from the assumptions of deontological (ethics which places importance on duty and obligation) and consequentialist (ethics which places importance upon outcome) schools of ethical thinking. Indeed, I posit that the same critique which applies to bioethics applies to AI ethics; that outside of more comprehensive approaches to ethics, principles such as autonomy have little value (Clouser and Gert 1990). The conception of such values or principles, and the very questions asked of/about AI, partially depends upon the conceptualisation of AI which one takes, and its interdependencies with various actors. For example, one might view AI as simply a tool, one which results in an intended outcome if used, and that as a tool it destines a certain outcome (technological determinism), reflected in discussions of frameworks for regulating AI development.

The issue with translating directly from medical ethics to AI ethics is in part to due this difference between the maturity and organisation of the professions. However, construing the failure of AI ethics purely as due to a lack of professional standards and licensing oversimplifies the situation, positing bioethics as a wholly successful endeavour within the medical profession, and erasing the complexities of conducting ethics. There are a few ways in which bioethics has fallen short of achieving its purpose. It assumes the vantage point of its predecessors and current experts is correct. In *The Tyranny of Expertise*, Carl Elliott reflects on the dangers that according too much authority to bio-ethicists poses to medical practitioners, *"Bioethicists are treated as experts whose judgments on ethical matters must be solicited, quoted, paid for, deferred to, and*

perhaps occasionally refuted or criticized, but in all cases, given the proper respect" (Elliott 2007, p. 372*)*. These ethicists are in their roles not just due to their knowledge but also "because of their links to the dominant power structures, such as industry, government, and professional bodies" (Elliott 2007, p. 46). Elliott describes how bioethicists possess a unique power not replicated in fields such as engineering ethics and legal ethics, but in the light of numerous suggestions that the answer to AI ethics is to bring ethicists on to teams to handle ethical decisions, it is important to be aware of the implications of this especially since ethical agency of practitioners is already at risk.

> "You have not described the case correctly unless an ethicist says you have; you have not made the correct moral judgment unless it has been confirmed by an ethicist; you have not produced a reliable set of ethics guidelines unless it has gotten the bioethics stamp of approval" (Elliott 2007, p. 46).

By contrast (or perhaps an extension), Black bioethicists describe a gatekeeping of topics they propose as regarded as not "real bioethics" (Ray 2021), demonstrating their separation from the dominant power structures due to racism and intra-disciplinary marginalisation which has shaped the epistemic landscape. This "narrow vision of the field" built on the perspectives of earlier (primarily white) domain experts consequently relegates Black bioethicists' contributions to a sort of niche domain which is not afforded similar legitimacy. Similarly, Emmerich (2016) conceptualises the field as a form of trading zone, where knowledge is exchanged at the intersection of multiple disciplines, in the form of interactional expertise (Collins and Evans 2014). Emmerich argues that bioethicists must engage with a relational view of their work, situating it in relation to governance, policymaking and so on. Returning to AI ethics, we find a similar way forward. From the vantage point of the post-phenomenological positioning introduced in Chapter 2, AI brings about certain impacts not just due to intentional engineering, but also due to the practices of developing it which inscribe a script of use, intentionally or unintentionally.

The Importance of Context and Relationality

Miceli *et al* (2022) have critiqued the restriction of ethical considerations such as bias to narrow definitions addressed at the level of the individual dataset. In framing socio-ethical considerations as problems, we imply existence of corresponding solutions, all the while excluding consideration of the impacts of power structures and socio-historical conditions upon ethical points of consideration. Meanwhile, Leonelli (2016) mapped the problems which data science poses for pinpointing accountability, identifying many of the impacts of early design decisions which are still central in the literature today, issues such as the lock-in inherent to choosing certain data formats and tools, and how choices of data and approach impact the focus on data scientists. Leonelli extended the definition of distributed responsibility to consider the impact of context upon the distribution of responsibility. Similarly, Jacobs proposed AI governance through epistemic mechanisms of measurement, bringing these wider considerations of data framing into the purview of AI ethics (Jacobs 2021).

A recognition of the relational nature of AI practice is also important to broadening this framing of AI ethics. Relational theories of ethics, such as Care Ethics, have centred on the role of reciprocal relationships (whether individual to individual, individual to community, individual to machine and so on), conceptualising values as situated in reciprocal relations between agents (rather than being centred on one specific agent), rather than universals and rights (Van Wynsberghe 2013). Given acknowledged power structures which influence the design of AI and shape dominant approaches to AI ethics, a more reflexive framework would consider the most marginalised perspectives in its foundational conceptions. Birhane (2021) proposed Afro-feminism and enactive cognitive science as theories of ethics which account for this relationality. Proponents of Afro-feminism construe knowing and understanding as embodied processes arising from direct experience, rather than the distanced, objective model of rationalist theories (Birhane 2021, p. 4). Enactive cognitive science posits knowing as an activity, an embodied process between the knower and known, with this knowing process necessarily informed by the various factors comprising their being and history. Drawing from these relational theories, Birhane presents an ethics of AI centred

on human relations, with the individuals who are disproportionately impacted and marginalised communities at its focus. It should aim to acknowledge structural inequalities and the hidden labour inherent in the structures within which technologies are developed. Considering concerns such as inequity as narrow, discrete problems abstracted from the power dynamics which create them in the first-place risks false illusions of a quick fix (Hampton 2021). Practical instantiations of Black feminist ethics include designing mechanisms for refusal, for example Benjamin's mapping schema (Benjamin 2020) which centres those most affected by an AI system to identify power asymmetries and recognise how these systems demonstrate wider socio-political issues, "thereby locating points of resistance, refusal and the redistribution of power".

Similarly, in a critique of how "fairness" has been used to address issues of disability in design of AI, Bennett and Keyes (2020) considered how this ideal risked reinforcing existing power dynamics, such as through reinforcing medical gatekeepers who defined the scope of disabilities, or promoting tools and techniques that benefit a subset of disabled people who are drawn from a heterogeneous group. Much like with technologies which have been biased against women, people of colour, and members of the LGBT community (Whittaker *et al* 2019), the failures to consider structural injustices in design and the ethical framing of "fairness" favoured processes suited to more readily identifiable (and otherwise-privileged) disabled people, while harming and reinforcing the marginalisation of those considered outliers. Without recognising relationality and the inequities within and between populations, these attempts to achieve "fairness" risk becoming new modalities of oppression.

This chapter has charted an overview of the history and composition of the AI sector, and some of the dominant approaches to AI ethics. It has presented my perspective on the landscape and shortcomings of these approaches and charted some possible ways forward which I consider subsequently in the thesis. The various permutations of Artificial Intelligence are a product of complex interactions between histories, disciplines, stakeholders, marketing and so on. As this chapter has explored, various aspects of these interactions find their way into the way the ethics of AI is conceptualised, whether intentional or not. Although several recent studies have begun to investigate practices

of AI, there is still little understanding about the interactions between immediate contexts of AI development and practitioners' values, and the wider implications of these for an ethics of AI. A consideration of these contexts of practice forms the starting point for the discussion of my findings from the interviews of this study, which I discuss in the following chapter.

CHAPTER 4: The Shape of Practice

[Airy and well-lit, the research lab cultivated a sense of openness in its design. Even the security gates were understated and bounded by clear glass barriers which offered an open view of the atrium. Housing a cafeteria and coffee bar, along with a grand piano and several colourful murals, the ground floor provided a space for connection and sustenance. Whilst the large tables fostered a social atmosphere, there were booths at the back for more private meetings and escape from the bustle of the cafe.

The other floors contained offices of various departments, with a mainly open-plan layout, and glass walls maintaining a comfortable level of natural light. However, the main source of natural light was taken up by the single or double-occupant offices at the outside edge of the spaces. These material boundaries indicated the positions of the occupants on either side. While offices housed permanent, and more senior, members of staff, the open-plan areas housed interns and post-docs, the transient yet essential lifeblood of the lab. Open-plan desks were easy to fill and to vacate, much more flexible than the limited private office space. They facilitated discussion, making it easier for anyone within the building to reach the occupants of these spaces, promoting an awareness of the self within the space, in the knowledge that at any moment you might be unknowingly observed or unexpectedly engaged in conversation. Only people with approved access can be allowed into the lab, yet once they have access, this access is nearly completely unfettered.]

My initial impressions of the lab, of carefully crafted openness to shape perceptions of the places where the privileged few conceptualised and developed AI, reflected a broader state of the field, as this chapter explores. The boundaries of AI are permeable and overlapping, from the standpoint of social engineering as well as definition of the field or topic itself. Scholars have brought attention to the vast yet invisible workforce which underpins the field, most notably in the form of gig economy workers who label data or direct these interactions (Gray and Suri 2019; Miceli et al 2022; Schmidt 2019), oft overshadowed by the contrasting pristine office spaces where decisions are made. This workforce can include (but is not limited to) domain experts providing input to

project framing, data preparation workers, data architects, data labellers, software engineers, User Interaction designers and so on, often distributed over a global landscape characterised by material and socio-political inequities (Altenreid 2020; Amershi *et al* 2019; Armbrust 2021; Catanzariti *et al* 2021; Sambasivan and Veeraraghavan 2022; Toxtli *et al* 2021; Van den Broek 2021; Yang *et al* 2020). There is an increasing body of research which explores the nature of this work. For example, there is increasing scholarship on the experiences of User Interaction designers in working with AI, who report that AI complicates design to a considerable degree (Morrison 2021; Wang 2022; Xu 2019). Highlighting this domain of interactions between people and their circumstances allows a consideration of how the intrinsic values of AI practitioners both shape and are shaped by the environments in which they work, which include factors such as the resource constraints impacting progress, and cultural values directing which efforts are best rewarded.

To draw from de Beauvoir's Ethics of Ambiguity, this chapter investigates the 'facticities' which form the limits of practitioner agency, the "clusters of constraints" which shape practice (Galison 1995, p. 15). In doing so, this chapter seeks to enrich understanding of contemporary AI practice by providing a constellation of insights into the embodied cultures and materialities of practice (Ihde 2008). Reflecting upon the findings of the interviews and observations making up the ethnographic research, I illustrate and discuss the ways in which culture and material context intertwine to shape practice.

This chapter considers this intertwine from two perspectives, which could be described as "top down" and "bottom up"; top-down considers how existing practices and industry-specific expectations shape practices, while bottom-up perspective considers how the practices themselves produce the outputs that come to be associated with AI, and how both are connected and constrained by the limitations of access to resources. Then, I illustrate how the contexts of practice were marked by uncertainties and ambiguities, from process to culture. This reflects similar findings about other types of scientific research, which demonstrate that the presence of uncertainty is endemic in scientific careers, impacting decision-making in numerous sites of scientific research

(Fochler and Sigl 2018), with practitioners expected to navigate ambiguities, and manage conflicting expectations (Byers 2010; Grinnell 2009).

Constructing Concepts of Artificial Intelligence

In the linguistic choices and rhetorical devices which practitioners employ in the process of defining AI, we can find traces of cultural and material influences which also come to bear upon practice. Over the course of the interview study, I came to understand the nuanced but crucial distinction between Artificial Intelligence as a concept and Artificial Intelligence as a field or tool. Furthermore, the extent to which either was discussed was closely related with the role and training of the practitioner. Through various discussions with practitioners, I describe a spectrum of configurations of AI along axes of industry and academia, and theoretical and applied. This investigation of how AI is understood and defined does not seek to identify a unified understanding of what AI is, or even to map out the different nuances of this in order to suggest a framework for definition. Rather, it explores how the different ways in which the term is defined, used, and critiqued can shed light on the nature of the experiences and environments which shape these understandings.

Reflecting the diverse backgrounds and disciplines which shaped practitioners' journeys to the points at which we spoke, AI was conceptualised in numerous ways. As discussed in Chapter 2, AI is a suitcase word, or a word which carries many implications according to the situation and person using it. Unsurprisingly, one contention which surfaced over the course of the interviews was the question of what "counts" as AI, with this having run-on implications for whether roles are considered AI work. Reflecting the nature of AI as a suitcase word, practitioners held nuanced conceptions of the term, including individuals holding multiple understandings of what "AI" refers to. Through the ongoing discussions, I found that start-up founders had the most training in theory and demonstrated the most nuanced considerations regarding what "counts" as AI. In contrast to this, people based in product-oriented roles in larger organisations had the least engagement with theory and were less concerned with elucidating the nature of AI. Although certain practitioners working primarily on theory

projects seemed dismissive of applied non-AI work, the same ones also wished to work on projects with real world value, which are by nature more applied. As this illustrates, perceptions of what counted as AI were characterised by multiple tensions.

Applied and Theoretical

Across the course of my discussions with practitioners about the boundaries of AI, I was struck by the focus of practitioner delineations on applied versus theoretical AI, and the contextual nature of these boundaries, oft intertwined with themes of meritocracy. Dewi was the founder of a computer vision start-up focused on healthcare, designing AI models to aid in interpretation of medical imaging. To Dewi, AI held multiple meanings depending on whether one was referring to the colloquial use in reference to tools, or a 'true', more specific, definition. On the one hand, he described AI in the context of his start-up as a tool to aid clinicians in clinical practice. On the other hand, a deeper discussion on ethics in AI led to the following reflection.

> "I guess when we're talking about true AI… true AI learns continuously and by experience…often people think that machine learning is AI…it is not. It is a methodology used for AI. AI is really a machine that is able to learn from an environment and improve with time."

Dewi's concerns for such an AI centred on the possibility for creating data loops, where unjust structures once learned, are then reinforced as they shape the society which draws upon the resultant outputs, feeding back into the system (Beer 2021). His broad definition focused more on autonomy than potential for sentience, the latter of which was the concern of practitioners such as Mark, a senior AI researcher in a Big Tech lab. Mark was overtly unconcerned with these risks of designing "foundational" AI which recreated potentially unjust structures, but rather was primarily occupied with discussing the ethical implications of AI outputs which were themselves considered agents (Artificial General Intelligence or AGI). He expressed how his primary concerns for AI ethics lay in the moral decisions of the machine itself, and whether a hypothetical "true AI" system could be designed to behave in an ethical manner. Stefan worked at the intersection of multiple types of AI work; he was the head of an AI research lab at a university, co-founder

of a start-up, and consulted for a large technology company. Stefan's research was also largely based in theory, and he felt that this "value-alignment problem" could be addressed by training models to pick up the values of the humans interacting with them, describing why he preferred to use one AI approach over another:

> "The reason that I find learning from demonstration such a compelling framework is that it actually tries to address this value alignment problem, so you try to learn a behaviour that is consistent with a human expert, so you sort of infer their values from data."

In contrast, Dewi's reflections belied an overlap between concerns for safety and data justice, recognising the risks of reinforcing epistemic injustices. Rather than intrinsic moral capacity, he was concerned with contextual make-up, serious ramifications potentially presented in an inequitable instantiation of AGI when hypothetically employed in a medical context. Brought together, these discussions perhaps suggest a view that the relevance of ethics-in-practice was tied to the degree to which AI research was applied. Lukas, a PhD researcher at an academic AI lab, reflected on the relationship between the field which a practitioner might be in, and the type of work they would do, commenting on whether there was a distinction between academia and industry in this sense. He explained that in his view the moral implications of AI work differed more between types of work (or between types of projects) than the type of domain a practitioner is based in.

> "I wouldn't draw the line between academia and industry, especially in this field. You have a lot of applied research happening in university… I know a professor in applying machine learning to medical data which would have a lot of ethical implications, would guess. And at the same time, you have different companies, Google, Facebook Microsoft etc which apply the algorithms but also as far as I know, also just do fundamental research without having a particular application in mind at least to some extent. So, I would draw the line more between different topics of research than between where it happens."

In Lukas view, abstraction from a specific end-goal of a project meant equal abstraction from ethical deliberation and anticipation. This reflects a view of opacity and distributed responsibility as diluting ethical responsibility, discussed in more depth in Chapter 5. Although a less popular topic of discussion over the course of the interviews, the flip side of this type of abstraction also surfaced in some conversations. Alec was head of an AI team and managed research and implementation at a large AI company. He spoke of concerns about the implications of drawing broad inferences from datasets scraped from the web, as intruding on the lives of those represented within the data with a birds-eye view never before possible, feeling that this encroached on the masses providing this information. However, it was the bread and butter of a lot of the AI work which he oversaw.

At various points in the interviews, this broad-ranging eye of data, "parameterised" via AI, was referenced as providing a more objective view (Hoffmann 2017). Thus, applied work was also subject to abstraction, implicitly constructing an "Imperfect Demon", an objective perspective on highly subjective aspects of human existence (Agarwal 2019). This concept of the Imperfect Demon draws upon the work of the marquis de Laplace, best known for his view that social affairs could be better implemented using his probabilistic methods, illustrated with the example of a computational judge, in his paper *A philosophical essay on probabilities*. His vision was founded upon the perspective that probability allows transcending "the ignorance and the weakness of the human mind" (m. de Laplace 1902, p. 196), by concretising the "exactitude that which exact minds feel by a sort of instinct without being able ofttimes to give a reason for it" (m. de Laplace 1902, p.196). Laplace constructed a thought experiment, which came to be known as Laplace's demon, where he proposed an omniscient intelligence which "having access to all data, can map all the behaviours of the universe, past present or future" (m. de Laplace 1902). This perspective is directly replicated in the language used by central figures of AI such as Geoffrey Hinton, who expressed a view of data as the key to improved judgement, as the key to domains such as medicine, claiming that AI models trained on massive amounts of data not only are a viable approach, but superior to current practice (Hinton, 2018).

A useful illustration of the practical instantiation of this Demon is the computational judge, designed with the belief that predictions of an algorithm can improve upon judicial decisions (Kleinberg et al 2018). This approach neglects the fundamentally embodied and affective dimensions characterising interactions within Medicine or Law. Radiographers exposed to the practical realities of Hinton's Imperfect Medical Demon were "…upset, and pointed out that they didn't just read these things, they also interacted with patients, and that was going to be harder to automate "(Vance 2021). While this conceptualisation of AI modifies the Demon to account for a nondeterministic world, it still asserts that given enough data it is possible to predict future events, even when disembodied from material contexts of development and application. Borrowing Laplace's language of disaffected prophecy is not hyperbole. Hinton spoke of AI as having predictive capabilities, an Imperfect Demon with limited capabilities, but a Demon, nonetheless.

> "You can predict the future quite well for a few years, and you might be able to predict what's gonna happen in five years' time, but as soon as you start making predictions about what's going to happen in 20 years' time, almost always you end up hopelessly wrong."[4]

In this way practitioners engage in a practice of abstraction, by asserting neutrality on the part of the algorithm. This ideal of neutrality also takes for granted the assumptions and presuppositions that have shaped events so far, obscuring these as "factual" outcomes of an unbiased social environment. However, such biases affect how practitioners understand and position themselves within the field of AI and can contribute to their engagement with – and applications of – ethics. Individuals associated with the "default settings", the common demographics of the field, may be more inclined to see ethical issues more abstractly and thus relegated to "soft" concerns which interfere with interests in application. Experiences of marginalisation shape differing perceptions of neutrality and value. After the final talk of an AI summer school at the lab, I heard the female AI PhD researchers packing up behind me say:

[4] Accessed at https://ashleevance.substack.com/p/oralhistorywithgeoffhinton

> "I think it's so awful that the female speakers were talking about ethics and responsible innovation, what a stereotype, such a soft subject. As if it isn't already hard enough for us in this field".

My first reaction was to bristle a bit, as a researcher working in that 'soft' field, which almost immediately subsided to understanding. After all, as I have shown in the preceding discussions of this chapter, AI ethics can indeed be viewed as a soft subject for the aspiring AI researcher who just couldn't quite make it. Women are indeed massively underrepresented in the field, often faced with misogyny from peers. This position can play out in a number of ways. Having such a standpoint and experiences of marginalisation or devaluation can contribute to nuanced understanding of issues of fairness and justice, and interest in righting experienced wrongs. However, this standpoint also creates a position of conditional acceptance, and an incentive to distance from concerns and concepts which might bring this acceptance and respect into question. These dynamics speak to the dimensions of power inherent in how practitioners navigate the field, which are further explored in the next section.

Power, Hierarchy, and the Political Economy of AI

Although their status may be fraught with tensions, nevertheless AI practitioners occupy a powerful station as a "coding elite" (Burrell and Fourcade 2021, p. 213). An important and recurring influence on the shape of practice lay in the power dimensions and concerns of practitioners within the wider landscape of the sector. Recent literature has described the corporate capture[5] of AI by a small number of powerful players (Whittaker 2021), which of note given how common it is for practitioners to work in multiple sectors, often holding some sort of role in Big Tech alongside other types of work.

Academia, Industry and Hierarchy

[5] Corporate capture refers to the influence and control which dominant corporate entities exert on the means and processes of decision-making, knowledge production and regulation to further their own interests.

The field of AI is significantly variegated, in the domains of knowledge it encompasses, sites and forms of practice, disciplinary interests, and expertise of practitioners. Early-career researchers might work purely on theoretical work at an academic institution, whereas the more senior researchers seem likely to actively collaborate with partners outside academia. Largely though, practitioners did not tend to see any meaningful distinction between AI work in academia and industry, pointing out how many academic researchers are employed in industry roles, be that working with start-ups, or consulting with large tech firms, and pointing to a distinction between applied and theoretical work. Although not all academic work was seen as abstract/formal, Stefan described the formal work as the academic work:

> "I wouldn't put the split at academia versus industry…because in academia there is also applied research, like this telepresence robotics project was an academic project. For the really academic stuff it's really about inventing new algorithms and inventing new, like, it's about equations…In a more applied project, like [] or this [] project, or the stuff that my company is doing…we're already thinking about the end user from the beginning".

Stefan's neutral phrasing of "really academic" evokes the prior descriptor of "soft", marking a distinction from applied work which was framed as more important or impactful to the industry and public. Hence while academic work was acknowledged to contribute to the development of the field through numerous inventions, its value (and the value of those engaged in it) was typically sidelined in favour of the outputs that could be derived from it, a position likely reinforced by Stefan's own concerns from running a company. The construction of hierarchies of AI practice was evident across my sites of research. During an internship at a Big Tech lab, I noticed that certain AI interns working on formal methods and theory differentiated themselves from those doing applied work, describing themselves as doing the "difficult" work on "real AI".

These construed hierarchies of practices influenced practitioners' attitudes towards their own work and others considered to be in the same field. When discussing the ramifications of an ethics-related scandal involving a Big Tech organisation which

happened a few months prior to the interview, Lukas made clear that he viewed the AI methods implicated in the scandal as distinct from "real" AI, viewing them as too basic to count as "real". Besides the hierarchisation evoked in this perspective, it seems that this devaluation was more intended to support a narrative that the greater technicalities or complexities of "real AI" were less likely to result in similar unethical uses, thus locating any ethical issues within the perceived quality of methods. This view towards which methods contributed to "real AI" were typically bound up with legitimising ideals of being more objective, value-neutral, and ethical, which has implications for the perceived role and value of ethical considerations within practices. However, reflecting the ad-hoc historic development of the field, this sort of view was far from universal. George was a lead scientist in the corporate division of a Big Tech corporation, describing himself as a *"self-taught Machine Learning scientist"*, a departure from the narrative of practitioners with academic training in AI. George viewed the monikers of Machine Learning, AI, and data science as roughly interchangeable descriptors for related domains, hinting at the complex, multi-faceted nature of AI which this name concealed:

> "In the last few years people started referring to it as such, it was around for quite some time under different names, and they just finally came up with a unifying title for the domain".

As described in earlier sections, AI practitioners, particularly with theoretical expertise, are highly sought-after in Silicon Valley and beyond (Hoffman and Friedman 2018). Within this hierarchy, it seemed that ethics and related research sat at the bottom, as summarised by Kristoff (PhD researcher in an academic institution), who told me that he had heard *"people talking about people who work around the field of AI as 'people who wish they were in the field of AI', and they're too stupid…literally, quote 'too stupid'."* This casts further light on the reaction of the AI PhD researchers discussed in the previous section, perhaps. This view of ethics as low-status work, combined with pre-existing notions of women as less capable, could potentiate desire to avoid associations of gender with being "too stupid" to work on "proper" AI. In this way, hidden hierarchies of

identity and vocational merit play into conceptualisation of AI practice, concurrently shaping a narrative of who has the freedom to engage with the topics which challenge the values underpinning these hierarchies.

Conversely, the pressures introduced by market forces were influential in maintaining ambiguous definitions of AI, as this vague terminology can be exploited to justify and gain funding for work. Martin, the founder of a start-up, recently emerging from a start-up "accelerator" which developed academic outputs for industry applications, confessed that sometimes hiding uncertainty was necessary for the survival of the practitioner:

> "Reality poses itself before you can go to the ethics and then decline grants and then you don't have a client any more… you don't have time to…so it has been ethical maybe…in business you have this problem, that you are trying to sell things that you don't know 100% sure that you will be able to, but you are betting on it, you are putting resources into it, to make it true."

These market pressures impacted upon the types of work which practitioners felt able to propose and develop. Lorenzo, co-founder of a computer vision start-up, described these as dictating the focus of his work. He spoke about how concerns regarding funding had influenced his move from adaptive technology, which had been the inspiration for even founding his computer vision start-up, to the more profitable market of life sciences, which nonetheless still posed remarkable pressures. In the light of this, it was essential for to market the capability of his AI models as well as he could:

> "The area of adaptive technology was not a very easy one to navigate as a technology business, mainly because we were building extremely expensive technology, for a very low-profit market, and decided that we needed something that could drive it forward."

This is not a new phenomenon in AI. McDermott (1976) wrote that addressing a problem too concretely removes its appeal, with practitioners preferring to focus on the concept rather than the application. A recent study of modern Data Science systems found that

business teams were remarkably talented at reframing failures as successes, transforming even what Data Science teams reported as unsuccessful situations into marketable products (Passi and Sengers *et al* 2020). However, this observation also fits with that of more mature scientific fields. For example, in an analysis of how molecular biologists use concepts, Neto (2020) discusses how imprecise concepts can help scientists to "formulate and organise various activities" (Neto 2020, p. 58). In particular, and in keeping with how some of the practitioners I interviewed used terms, the imprecision of these concepts was useful for scientists to be able to account for the work of others and allow researchers to recognise the "transferring a principle from one area of biology to others" (Neto 2020, p. 58). Similarly, the imprecision of AI may allow for recognition that the principles and methods remain similar across applications and fields.

Corporate (and Regulatory) Capture

The overlaps between domains of practice were linked to corporate capture, the influence of Big Tech funding as seen across industry, academia, and the public sector, with AI work distributed across many sites. The market domination of large companies resulted in a sense of inevitability to decision-making. Cameron, who was helping develop a collaborative research centre which aimed to bring together industry and academic partners to work on AI and related domains, expanded on this by describing how:

> "So, now of course market capitalism would say if these rates are too high someone will come in and offer a smaller rate. But we know that these companies dominate the market. So, I think there's all that debate to be had and there's also debate to had about, um, once the service becomes kind of essential, so, you know if Google has all the data in the world it's very hard for anyone to compete with them so you have to take the price that they offer."

This was a particularly salient concern for those working in small enterprises. Lorenzo reflected on the reality of staying afloat as a start-up in the field, reflecting on the

downsides of being based in the UK due to its competitive disadvantage compared to other countries:

> … you have this field, which is extremely highly funded in the US, and extremely highly competitive as well with China, and if you aren't as well-funded as your competitors you will die".

The influence of a small group of powerful companies directly and indirectly mediated access to financing and access to resources, even the ethical culture of the field. Luke was leading a product Data Science team which utilised AI methods. He expressed strong opinions on the impact of corporate capture on the culture of smaller organisations, going on to specify that he particularly felt that a toxic positivity culture applied to companies with a Silicon Valley origin:

> "All companies nowadays have got values and they're usually very cheesy and they're usually very similar, um, and are box ticking… certain companies could be very savvy in their branding to look warm, fluffy, friendly, Facebook is one of the best examples…everyone thought it was their best friends, one of the best examples to their social lives and the company and it came across as social which always gives off good vibes and things but really what they would do with people's data was pretty difficult if you think of some of the ways that they were profiling people …whether it's Cambridge Analytica or the myriad other things that they were making money out of."

Indeed, ethics is another aspect of the AI sphere which has been explicitly subject to corporate capture. As I touched upon in the preceding chapter, amidst concerns about the implications and impacts of AI, Big Tech has argued against stronger regulation, rather proposing that developing in-house ethics approaches or engaging ad-hoc ethics consultants provides a sufficient answer to concerns (Jobin *et al* 2019). This has led to a proliferation of "soft ethics", transforming ethics into another form of asset which Big Tech can both influence and profit from (Phan *et al* 2021). Previous research on practical ethics has found that official forms of ethics may even cause disruption to ethics "on-the-

ground" and frustrate the ability of those in care roles to provide the services that are needed:

> "Rather than sensitizing people to the full panoply of ethical questions, official ethics focuses attention on some questions and deflects attention from others" (Heimer 2013, p. 377).

This issue of which questions are attended to is worthy of consideration given the evidence for regulatory capture by dominant organisations in the AI sector. The questions highlighted in practical ethics may differ from official ethics. Heimer concluded that practical ethics is "unlike official ethics", observing that practical ethics is not allied to sets of principles or official documents, and "may not even be explicitly identified as ethics…" (Heimer 2013, p. 377). He noted that access to resources played into this, with limited resources complicating engagement with official ethics, even if teams had dedicated ethics roles[6]. This finding is of relevance to the AI ethics sphere given recommendations for "ethics owners" in AI teams (Moss and Metcalf 2020). Thus, although compliance tools provide a veneer of legitimisation, reassuring governing bodies that adequate steps are being taken, in practice they may be toothless or even actively harmful, if not thoughtfully adapted to account for the relational context. In the time since I conducted these interviews, several ethics scandals have provided meat to the concerns of 'ethics' being used as a tool of legitimisation without action. The most high-profile of these was the firing of Timnit Gebru from Google, in response to Gebru raising concerns about ethical implications of Large Language Models (LLMs). In the words of AI ethics researcher Meredith Whittaker, "What Google just said to anyone who wants to do this critical research is, 'We're not going to tolerate it'" (Simonite 2021). These pressures, whether explicit or expressed more subtly, have impacted on the intellectual freedoms and focal points of AI ethics researchers and AI practitioners, as I discuss throughout this thesis (Ebell *et al* 2021).

As alluded to in the previous paragraph, corporate influence beyond direct implementations of "soft ethics" may include foregrounding specific approaches to

[6] The authors clarify that *demonstration of compliance*, rather than *practical ethics*, was resource-intensive.

practical ethics. Julie was a PhD researcher working in collaboration with an industry lab to develop AI models which they could then implement in products. She discussed how she had seen a lot of interest in and discussion of the issue of biases in datasets since it became a focus of key AI conferences, and commented on how this was having an impact on the mindset of practitioners:

> "Obviously bias in algorithms and in your data… this has been discussed a lot and it's really cool to see that the Machine Learning community cares about this a lot now…because…there have been a lot of instances where things have gone wrong and showed that bias in your dataset, which you didn't think about at all, has really negative consequences…it's something we should really think about more."

The issue of bias had been increasingly reported on since at least 2013 in humanities domains (Noble 2013) and in the popular media (Crawford 2016), with a very prominent report published in 2016 by ProPublica, which critiqued the racial biases encoded in the COMPAS recidivism prediction algorithm (Washington 2018). Within the field, the biggest push for consideration of these concerns came as response to the possible regulation of Big Tech, with organisations seeking out alternatives which would ameliorate the risk of external oversight being imposed (Phan *et al* 2022). This focus on fairness and issues of social bias has since been spurred other efforts by practitioners in subfields of AI (Zou and Schiebinger 2018; Corbett-Davies and Goel 2018). Many individual practitioners were acutely aware of the impact of broader structural injustices - *"It's biased towards the mechanism with which the data was gathered, prevailing cultural practices at the time"* (Adrian), - and aware of the limitations of approaches to fairness. However, the influence of corporate capture raises questions about which questions are focused upon in the first place, how they are framed and whose interests they serve.

Indeed, I noticed that practitioners were generally very engaged with addressing the ethical and social considerations which they had been exposed to in their field, the key phrase here being "exposed to". Julie explained one way in which one could address mitigating problems with data, was by anticipating the demographics which the data

would be collected from, but that this was complicated by pre-existing structural injustices such as access to technology:

> "There are probably problems where you want very diverse behaviours demonstrated and not just from one type of person. Say, chatbots, you don't just wanna collect data from white rich men for example, collecting data from a certain app because those people use that app a lot. Because your chatbot will be biased."

Considerations of the role of power and corporate capture surface here again, to help us answer the question, "who gets to decide which issues are on the agenda?" Thus, we see how epistemic injustice shows up both in practice and context. Epistemic injustice limits which ethical dimensions of AI are even discussed, resulting in downstream impacts which dictate whose knowledge is then represented in the data and models, in a recursive loop between data, AI and the social world (Beer 2022). In the words of Ruha Benjamin, "that new tools are coded in old biases is surprising only if we equate technological innovation with social progress" (Benjamin 2019, p. 108).

Materialities of Practice

Having noted some of the primary drivers of and influences on AI research, this section explores what it is to engage in AI practice at a more local level, exploring the nature of the methods, as unpredictable and characterised by contingencies, the nuances of datasets which models are reliant upon, and practitioners own limited insight into their methods.

The Impact of Resource Limitations

To draw again upon the work of Galison and D'Agostino (1987), I examine some of the constraints which shape practice, and by extension ethical decision-making, here illustrating the constraint of resource limitations with a vignette from the observational study:

[The pace of work at the lab was modulated by a constant tussle of hardware limitations with human objectives. Due to computational constraints, Remi often worked on the end-of-internship report, rather than her AI model. She had been struggling with optimising her algorithm so that her computer graphics processor (GPU) could run it.

After running experiments to test whether her model would run, and failing, Remi went for a brief coffee break to decompress and chat with a colleague at the machines. After returning to the desk, she explained that there were issues with the Graphical Processing Unit (GPU), and that despite varied efforts, the model was taking up too much memory from the external (but within the physical lab) GPU that the computer utilised and would fail when they tried to run it. Remi explained that although most machine learning researchers would happily use a Virtual Machine to access adequate GPUs, those were not sufficient for the purposes of this model. Indeed, such hardware limitations were a common issue in video processing due to the huge amounts of information a model would need to analyse. For context, most iPhones default to recording video at 30 frames per second, and with for example 4 seconds of video, this means the model is processing 120 frames per video, 360 frames per object, and so on. Compare this with a single frame for researchers using static images, who might have 3 frames per object, and the problem becomes clear. Remi reflected on how this hardware need provided a huge obstacle to labs which did not have large financial backing, using the example of university labs, creating inequity in the AI world.

By early Thursday afternoon, it was clear the GPU still had not been upgraded, and this had a considerable impact on her work. With the hardware capacity stalling Remi's ability to make progress with AI work, the focus switched onto other tasks, taking advantage of this time to fine-tune the final-week presentation. Every intern is required to present on their research in their final week of the internship, and Remi had the added motivational focus of this presentation doubling as an interview for a research role at the lab. By the end of Thursday Remi had still not been given an update on the GPU access she needed, so she continued to work on the report, presentation, and other non-coding tasks. She tackled the hardware issue by decreasing the amount of memory the model needed in order to run, working on implementing the changes a senior researcher had suggested, looking at the papers he had cited and working out how to apply the methods to this model. By Friday morning Remi could finally access better GPUs and run the necessary processes. Finally able to run iterations of the code, she spent the morning working on the model.]

Remi's experience was not unique or unusual, with hardware constraints a common concern when developing AI models. The computational power necessary to train certain models is considerable enough to even warrant concern around environmental impact (Bender *et al* 2021). Practitioners described how these constraints reflected the previously discussed inequalities in the field, posing a huge problem to researchers working at academic labs or industry labs which were not part of one of the dominating corporations. These inequalities in access to resources had significant impacts on what sorts of projects could be developed within the working environments, and hence how much input practitioners had in shaping outputs. Indeed, Remi reflected on the implications of the experience of that week at the lab. After all, if hardware constraints could be seen at a Big Tech lab, the impact must be far greater for practitioners working at smaller companies with much less access to the necessary resources. Lukas described how the pressure placed upon him by hardware constraints was indeed one of the primary concerns in developing and testing their models:

> "Given all the constraints imposed on you, time, computational resources you oftentimes you don't have the abilities to test it out as much as you would like to, so you have this internal mental challenge of trying to test your ideas, but not to, but trying to be scientifically as rigorous as possible."

Similarly, Jason (an AI researcher in an academic lab) described how the limitations of hardware had a foundational impact on the research decisions he made in terms of direction, shifting his focus towards theoretical work in order to minimise the impacts of lack of access to necessary computational resources, and told me how he preferred:

> "A more theoretical approach, because I think using a more aggressive approach, you can kind of break new ground in new areas without needing all this computational power, who has access to that power?"

This acknowledgement of the impact of material constraints invites considerations of the resultant affordances influencing practitioner decision-making. To borrow from HCI, an

affordance is a design feature (intentional or not) which permits or encourages certain types of interaction with an artifact at the expense of others (Blin, 2016). Material resources are potentially co-constitutive with the deliberation and decision-making around AI practice. Perhaps this recognition of the "brick and mortar" of AI itself acts as a sort of ethical affordance, directing the deliberation itself through limitations it places on practitioners' agency, creativity, and reflexivity. (Schoenherr, 2022). Yet, beyond influence of material resources, ethical concerns and discussions were also bound in the co-constructed data sources on which practices and outputs were based, as explored in the section below.

Data and Data Work

AI is knowledge work, built upon data (Mackenzie 2017), with the quality of these models only as good as the data they are predicated upon (Redman 2018). Over the past few years there has been a large uptick in studies investigating and critiquing the nature of data work, reporting "serious gaps in what AI practitioners were trained and equipped to handle" (Sambasivan 2021, p 77). Jason explained how datasets and models interact, reflecting upon how building a 'black box' model upon a biased dataset compounds both issues of opacity and of fairness:

> "It's particularly important for machine learning systems because they're often employed in the real world as these black boxes that are hard to understand. And people often give a lot of authority to these kind of systems because they are quantitative but actually the models are often …the models are always biased… but it's the data that you input to make sure that the users can understand and can interpret what the models spit out in a good way and so they can decide what to, how to use the model in the way they direct it to"

Unfortunately, the data work which forms the basis of the system is often devalued as a practice (Sambasivan *et al* 2021), with data acquisition, data cleaning and integration taking up to 90% of practitioners time according to Muller et al (2019). In addition to being viewed as a mundane practice, perhaps this negative view is partly because it is

this point which most clearly reveals the fundamental ambiguity and experimentation which belie claims of rationality and objectivity, the Imperfect Demon falling foul of self-scrutiny. Biases form an essential facet of ground-truth data, as "necessary, yet contingent, external referents… operate as supervisors of learning processes" (Jaton 2021, p. 2).

However, Muller *et al* (2019) found that practitioners clung to the notion of the "truth" of their data, with truth used as an equivalent to "fact", pointing out that this subtle equation of truth and fact was a display of the power inherent in the role of practitioner. Despite vigilance to avoid incorrect patterns, the notion of ground-truth in datasets (for example labelling from surveys or domain experts) was often not scrutinised. Furthermore, as mused upon in the previous section, this equivocation of truth and fact allowed practitioners to disengage with the actuality and contexts of the data and employ proxy tools and methods which relied on the notion of the status of the data as "true". Power again figured in the labelling of this data, with the authors giving the example of the case of the practitioner who labelled his own submissions, "I am the ground truth" (Muller *et al* 2019, p. 10).

Grove *et al* (2021) describe "digital data points" as spores containing very partial information, which are nevertheless manipulated by technological narratives "as entities that can function as agents of mediation" (Grove *et al* 2021, p. 2). Again, the veracity of the "ground-truth" is not necessarily of note, and often they will be developed and utilised with limited evidence. Drawing on Ingold (2015), the authors conceptualise the togetherness of data as a mesh of ongoing entanglements which connect and diverge at various points, rather than form a stable network (Grove *et al* 2021), matching the descriptions given in this thesis. Indeed, the "ground-truth", and the assumptions which underly it, often form ground zero for the problematisation of a space (Jaton 2021), that is, the framing of a domain in a way which motivates the application of AI within it. Indeed, data acquisition and annotation have subtle overlaps which place both into the realm of ground-truth selection. Ground-truth is often represented by human-labelled training datasets, however, even unlabelled datasets for unsupervised algorithms behave as ground-truths. That is, the choice to use them to train AI models reveals underlying belief

in their representation of the real-world. However, this unchecked belief in neutrality reflects and encodes broader structural injustices (Mhlambi and Tiribilli 2023). AI models are often evaluated in reference to a benchmark dataset for which performance is commonly known, with different domains developing benchmark datasets for example, to test the performance of medical (Ponomarenko and Bourne 2007) or semantic web models (Ritoski *et al* 2016). However, ground-truths go beyond labelling or benchmark datasets to the "choices and actions contributing to the aggregation, probing, organization, and cleaning" (Jaton 2021, p. 9). Lukas described how resource constraints and dataset quirks came together to form tensions at the centre of practice:

> "…you believe it's useful because you've seen it succeed on the dataset but you kind of neglect that you tried it out on five others, and it didn't work. Which is fine because not every algorithm has to work perfectly all of them, but given all the constraints imposed on you, time, computational resources, you often times you don't have the abilities to test it out as much as you would like to, so you have this internal mental challenge of trying to test your ideas, but trying to be scientifically as rigorous as possible."

Here, we see a description of compounded materialities which are navigated with the best of intentions, but risk unintentionally allocating the status of ground truth due to diverted attention or seeking a model that "works" given the computational resources. This observation seems obvious to anyone who has engaged in machine learning and is sort of embedded in the culture, yet these realities are obscured when the datasets created through these processes become the basis for additional modelling or foundational to developing real-world applications. In other words, if subject to a context shift, it results in ethical debt, since these "working models" are used "without proactively identifying potential ethical concerns" (Petrozzino 2021, p 205).

Despite the tensions, practitioners enjoyed the challenge posed by constraints, finding reward in the creativity required to navigate them. Alec described the main thrust of his practice as being to "put pieces together, different data, to try and build a picture", using language of craft and art to explain the complexities and nuances of his work.

Muller *et al* (2019) found that data work utilised a combination of the practitioner's intuition, domain knowledge, and a trial-and-error process of designing predictive features, with one practitioner even referring to their processes as a "kind of art" (Muller *et al* 2019, p. 8). Julie reflected on how data collection for projects required a conscious engagement with the intentions of use for the data, anticipating issues that might arise and need to be "mitigated":

> "…one thing is the implementation side but then there is also the data-collection side, collecting the data in such a way as you have imperfect data collection too…you have to make these decisions consciously of what kind of algorithms you're going to use and what kind of problems you can already mitigate during data collection."

While this technically makes sense, this somewhat artistic approach to the craft of data work has fundamental epistemic implications (Thomer *et al*_2022). Missing data is a problem in constructing ground-truths, and practitioners might have to impute data themselves in order to run a model using other models or their own intuition, resulting in greater complexities of data curation. For example, expert-labelled ground-truths can be difficult to obtain due to the time costs involved, with knock-on effects (Jaton 2021; Kang 2023). Julie reflected on how the potential for data to be utilised using AI to generate predictions also had risks from a lack of insight into these large datasets, the biases they encode, and what modelling them might entail. She was particularly concerned that by employing other to model their data, data-owners would believe that this was equivalent to understanding it, even without access to any ground-truths:

> "People have massive amounts of data, which they don't really understand, can now use it with deep learning. But on the other hand, you still don't understand your data, then you have this black box, which is telling you something about this data, but you still don't understand what's going on."

Indeed, returning to the Muller *et al* (2019) study, the authors observed that practitioners had to be creative in identifying alternate sources of information, for example CCTV

coverage of the behaviours which they were interested in. According to their study, both surveys and CCTV footage had their downsides, which practitioners reasoned about in terms of value trade-offs, weighing feasibility against accuracy and validity (Muller *et al* 2019). The study reported that ground-truths were impossible to access and had to be simulated, or even guessed. This approach has proved risky, one example being when the IBM Watson team tried to develop a model to predict personalised cancer treatments based on analysis of the oncology literature (Ross and Swetlitz 2018). Here we see tangible impacts of the collapse of human complexity into subsets of data-points. The IBM Watson team discovered that when reading articles, physicians often utilise information which is not the primary point of the study, adapting their care in a way which is qualitatively obvious but when considered based on data alone, is far from obvious. For example, when the FDA released a drug personalised to specific genes, only 4 out of 55 participants in the study they cited had lung cancer. Physicians now knew to routinely screen lung cancer patients for this gene, but such a small percentage would likely not be picked up by the AI model. To address this issue, the team created fictional profiles to train the algorithm (Strickland 2017), which proved untenable in such a high-profile, high-stakes domain. However, this employment of fictions to train a data model showcased how experimental processes were applied to demonstrate the utility of AI, complicating purportedly clear ideals of workflows as described below.

Models and Workflows

In contrast with traditional software development workflows (although, even these are subject to change), AI workflows were characterised as not having linear workflows to reach a very particular end-goal. Rather, generally stated end-goals were described as rather vague, and periodically updated based on the results of various iterations of work. Joshua (co-founder of a computer vision start-up) described the process as *"…fairly free-form and experimental"* and difficult to plan for, because *"as with all machine learning models it's very hard to predict whether or not a change will be positive or not"*. He summarised this way of working as *"you basically try stuff, throw stuff at the wall and see*

what sticks." Alec went further, to suggest that the entire process is unclear and unpredictable:

> "Often at the start of a data type project you don't know where you'll get led by the data, and you look into what's possible rather than, if you're building a product you might start with a list of features which you steadily add on".

Even in multinational technology corporations, and in the product divisions of these, the nature of AI techniques as vague and difficult to predict outcomes impacted the way in which in which projects were approached and planned. George's department operated according to a "matrix-driven" ethos, where practitioners were given goals over a time period (such as a year) and just told *"you must increase x by two points this year, something like that."* This statement alluded to demands to achieve concrete (and often perhaps arbitrary) end goals, priming practitioners to demonstrate some sort of progress with processes that were difficult to predict and subject to constraints as described earlier. Similarly, in academia, Lukas described how his research on AI theory inherently involved uncertainty, especially given the black-box nature of this approach:

> "Because I know that it's kind of fiddly, and unreliable, and black box to some extent, I design my workflow around that that obviously means trying my ideas as early as possible and ideally as rigorously as possibly, especially on toy domains where you can control all the influencing factors… could get a lot better at that but that's the idea at least."

Reflecting on this inherent unpredictability, Lukas wondered about how this impacted the safety of his work and the algorithms his team worked on, saying that *"we don't know what will work what won't…sometimes we just try out things until it works basically"*, describing how this meant that a lot of his work was based on building intuition over time, necessitating practitioners gain *"a little bit of understanding and intuition of the things which might work and what things might not work"*. When pulled together with the uncertainty of the process, perhaps even experts in the field would struggle to thoroughly document and explain their work:

> "We have a better understanding but it's far from perfect, and so it's impossible to predict and it would be really difficult to write a guideline on how to use an algorithm. Cause you…it's hard even to come up with all the cases of when it might work or when it might not work, it's a very difficult problem I think."

Even iterative approaches to software development seemed too structured for AI work. Alec described how he got to the point of losing patience with the well-known project management tool his teams had been using, and had even recently led a boycott of it, as it was *"good for making products but not very good for doing science or doing data science, when you don't know where you're gonna go"*. In the end his team moved to using a time management and tracking app *"for monitoring how many hours we spend on things"*. This experience extends to other aspects of AI practice, with data curators needing to engage in "craftwork" to get Jira to work for their processes (Thomer *et al* 2022). Indeed, George remarked how his industry team's data science work differed from standard software engineering:

> "There's much less of this sort of day-to-day coordination like you would with a traditional software development project where everybody is working on the same thing, and you have a design that you are implementing in pieces".

Several studies have tried to capture the common structures of the AI workflow. Patel *et al* described a workflow consisting of "formulation of a learning problem, collection of appropriate training data, the extraction of features from the data, the selection of a modeling algorithm, and experimentation to determine whether the resulting system meets the needs of the application" (Patel *et al* 2008*,* p. 669). However, interestingly, the practitioners they interviewed caveated this workflow by emphasising "the fact that the actual development of a system is much more exploratory than such linear dependencies suggest" (Patel *et al*, p. 669). Providing insight into the perspectives of other potential stakeholders in AI design and development, Yang *et al* conducted a study of 50 HCI designers and 12 AI researchers, investigating sources of complexity in human-AI interaction design, and found that the two main sources of this complexity were

capability uncertainty (defined as "the functionality AI systems can afford (e.g. detect spam emails, rank news feeds, find optimal driving routes), how well the system performs, and the kinds of errors it produces" (Yang *et al* 2020, p. 6), and output complexity (defined as "what an AI system produces as a possible output" (Yang *et al*, 2020, p. 7). They concluded that designing human-AI interaction presented much more of a challenge than typical HCI work.

These reflections echo observations from across the history of the field, characterising AI practice as experimental and hard to predict (McDermott 1976), as "craftwork" (Suchman and Trigg 1993), and "highly iterative and exploratory" (Patel *et al* 2008, p. 669). They also surface a tension in contemporary AI, particularly in applied contexts, a tension representing its uneasy place between science and engineering, between studying a phenomenon and building an output (Parnas 1999).

The findings I discussed in this chapter scratched the surface of a very complicated set of hierarchies and interruptions which characterises AI work, and illustrated that practitioners often worked in different domains simultaneously, with the boundary placed at applied versus theoretical work. The terms Artificial Intelligence and Machine Learning are ambiguous or at the very least under-defined. However, as we have seen with the way practitioners employ these terms, this ambiguity allows for flexibility and nuance of context. There were huge power imbalances in both industry and academia based on the resources available to labs and companies, with hardware constraints forming perhaps the largest obstacle to practice with the knock-on effect of forming a barrier to moral reflection and engagement. Yet beyond the impacts of power structures and material constraints, the capacity for moral reflection and engagement could also be located within AI practitioners' framings of responsibility for ethical deliberation. This domain formed the second major theme from discussions with participants in this study and is explored in the following chapter.

CHAPTER 5: Locating Responsibility

"Freedom cannot will itself without aiming at an open future"

Simone de Beauvoir

The constraints and contexts of AI, where the values of the scientist meet the implications of the engineer, create a messy, complex set of socio-technical entanglements and ambiguous notions of responsibility. Working in an ethics role within an AI team can prompt an uneasy feeling; there is a constant tension of feeling the desire to contribute meaningfully, pushing against the knowledge of one's own limited ability to provide insight into a system which is in flux, especially at the early stages. In the prior chapters I described how this uncertainty informed my research direction, prompting investigation into the ways AI practitioners might conceptualise their responsibilities with regards to their work. Indeed, it was while discussing this topic that the complexity of AI practice, and the type of moral reflections it mediated, came into sharper relief. Such a fundamentally embodied process mediates conceptions of responsibility, reminiscent of Dreyfus' remarks on how specifics of coding languages can create unconscious misunderstandings of work (Dreyfus 1965). As such, consideration of responsibility within AI bridges the facticities which shape practice (as discussed in the previous chapter), and the values which guide it (which is the focus of the upcoming chapter).

Modes and Sites of Responsibility

This chapter pulls at some of the threads of responsibility[7] and accountability in designing and developing AI, considering their fundamental connection with ambiguity and uncertainty. I examine the ways in which AI practitioners framed responsibility; this is organised into the facets making up distributed responsibility which result from the pipelines making up knowledge contributions, and those of obscured responsibility or the epistemic standpoint of practitioners within this. I employ the concept of epistemic

[7] I draw upon the definition of Smith (2015) - a morally responsible agent can "'answer for' her attitudes and conduct" (Smith 2015, p. 103).

responsibility as understood by Simon (2015), to examine looping or intra-action (Suchman 2011) from two perspectives; bottom-up "what does it mean to be ethical in knowing?" and top-down "what does it take to enable responsibility in knowing?" (Simon 2015, p. 146).

The fundamental complexities and ambiguities characterising responsibility in AI practice can serve to reduce concepts of agency, when viewed at an abstract, absolute level. This can occur either intentionally or unintentionally. For example, ambiguity between abstract research and implementation in the real-world can potentially obscure the potential impact of practitioners' work, or be easily framed as obscuring it, and thus affect their sense of moral responsibility, chiming with the observation of Louis (senior researcher in an industry AI lab) that "*a significant portion of researchers…if they don't work on real world data think, well it doesn't really apply to me*". However, examining responsibility as pertaining to situated practice clears some of the fog.

Often, participants would touch upon issues pertaining to responsibility and accountability within the first few questions of the interviews, questions which did not mention these topics. I broadly separate framings of responsibility into three rough types; obscured, distributed, and diffused. Obscured responsibility refers to the obfuscation of who exactly is responsible, especially in combination with opaque, complex, and unpredictable models and processes, complicating attempts to understand the implications of their work and mitigate harms. Distributed responsibility refers to the actions or outcomes of a system when spread out between multiple individuals, teams and larger units, the data and approaches they use, and the specific contexts of their work. Diffused responsibility occurs when obfuscation results in avoidance of accountability due to this presence of others or mechanisms which can also be held responsible, a phenomenon also seen in content moderation, where "increasing visibility of a content moderator may inadvertently inhibit bystander intervention" (Bhandari *et al* 2021, p. 284).

Diffusion, Distribution, and the Political Economy of AI

Agency, Expertise and Diffusion

Certain applications of AI prompted reflection on the nature of responsibility and "moral good" in the context of AI practice. To be more precise, the framing of these applications prompted this reflexivity. Alec reflected on his discomfort with a personalisation project he worked on, combining large datasets scraped from the web with other sources shared by companies, to build models which segmented groups based on shared characteristics. He told me that *"… I don't know how I feel about some of it… it's sort of systematic, on a big scale that you're doing stuff about people, if you did it in real life it might seem a bit creepy"*.

Phan and Wark comment on this as experienced by the individual being profiled as possessing a "creepy-factor" (Phan and Wark 2021, p. 4) which gives perception of personalisation a "dull sheen" (Phan and Wark 2021, p. 5). They contrast this dull, perhaps neutral, view of personalisation at the individual level with its "profound shaping effects on our societies" (Phan and Wark 2021, p. 5). This awareness of creepiness can serve to desensitise the perceiver to the more disturbing implications of such emergent socio-technical assemblages. This applies to the perspective of the practitioner; the systematic scale of the model is concerning, but this concern is examined at the individual level, defanging it of ethical impetus. The diluting effect of framing ethics at the individual level also illustrates a potential flaw of the subjectively-based ethical heuristics which are explored further in Chapter 6, in diffusing some of the more concerning considerations which might be posed in a higher-level analysis. Similarly, certain AI approaches were felt to be more amenable to be employed as a mechanism for distancing architects from negative outcomes of their models, underpinned by claims of objectivity and distance. Jason expressed concern about how using AI approaches such as "Reinforcement Learning" introduces limited accountability:

> "I think more often than not with that, because we're using black-box methods and there's very little accountability in there, I think they also have the power to be incredibly dangerous".

He noted how the use of certain AI approaches can even present a method of rationalising undesirable outcomes and a form of plausible deniability for the practitioner; *"you can inject previous biases into datasets, and then blame the neural network for being this kind of 'objective' mathematical algorithm acting on what is actually inherent bias in the dataset"*. However, the uncertainty of practice has an associated risk, which should necessitate the practitioner to be *"explicit about risk. Because these things change, it needs to kind of say on the label that we don't know"*, in Cameron's view.

At the same time, a factor impacting whether specialists engaged with moral decision-making was their level of confidence in their own understanding of ethics. Ethics beyond compliance was sometimes perceived as reserved for elite ethicists, as important in theory but not relevant to the day-to-day mundane workings of practitioners, and as something they wish to be trained in but lack the time or resources. Skye (data scientist at a multinational AI organisation) made a nuanced distinction between lectures on ethics, perhaps akin to professional issues and compliance, contrasted with being taught how to reflect on right and wrong; on the *"ethical"* process rather than the prescriptions of *"ethics"*. Furthermore, this lack of formal ethical training sometimes resulted in feeling unable to contribute properly to ethical discussions despite having a keen interest in the subject. There were sustained discussions on what an ethical education for AI practitioners should comprise of. For example, Luke told me about his anxieties about being qualified to do ethical reflection and decision-making:

> "I am still learning. I don't have a formal education in philosophy, and I am concerned that the education that I have obtained for myself has holes because I can't just jump into a conversation without people bringing out terms that I'm not familiar with although in every field people like to use words they don't need to use. It's something I do have a strong interest in."

In fact, although Luke described how he felt like although he had a strong interest in engaging with ethical practice, often he also felt excluded from the conversation due to *"people bringing out terms that I'm not familiar with, although in every field people like to use words they don't need to use"*. As discussed in the Chapter 3 (on Approaches to

AI Ethics), other practitioners also described how ethics training in university courses did not translate well into practice because high level concepts were not translated in ways which made sense to a practitioner trying to apply them. The result of a lack of confidence in ethical abilities on day-to-day AI work is characterized by Stefan, who prefers to defer anticipatory analysis of impacts to ethicists:

> "I don't feel like I could make useful predictions about how our technology could be used, probably someone could...an ethicist or someone…people who are speculating about the long-term consequence, probably they could sit down and think systematically about it."

However, placing ethicists as experts not only created the sense that the responsibility for action belonged to them, but changed the focus to lie upon outcomes rather than process, resulting in overestimation of the ability of the ethicist and misunderstanding of their role. Furthermore, the association of ignorance of ethics with inability to engage in ethical reflection or discussion can be harmful. These statements can be argued to represent a deflection of responsibility by way of purposive ignorance, as a kind of weaponised epistemic lack (Tilton 2022). As Proctor (2008) argues, an attitude towards ignorance which is often rooted in its common understanding as a native and naïve state to be corrected with adequate knowledge can equally become a chosen position born from selective [in]attention or claims of gnostic inadequacy. In essence, a claim of not knowing enough to act serves both deflect responsibility, and to maintain a position of innocence or absolution from blame in the event of any emergent harms that result from ignorant actions (Proctor and Schiebinger 2008).

This potential deflection was further complicated by the concern that AI systems were too complex for non-experts to understand sufficiently to be able to comment sensibly on, centralising power in the hands of practitioners. Julie felt that using AI to generate predictions from data involved risks posed by a lack of insight into the nature of the datasets, the biases they encode, and what modelling them might entail. Indeed, data-holders ran the risk of believing that by employing other to model their data, this was equivalent to understanding it; *"you have this black box, which is telling you*

something about this data, but you still don't understand what's going on". Furthermore, the specialisation of knowledge could also act as a barrier. There was a concern that the complexity of the technical work Skye was involved was sufficient that even someone trained in an adjacent topic would struggle to understand how it worked; *"someone who isn't intimately familiar with it will struggle to understand what it's all about."* Thomas, who was the Head of an AI subsidiary of a commercial organisation, expanded this consideration with a reflection on power, pointing out how this imbalance disadvantages end-user groups:

> "You're dependent on something you don't have any control over or know how it works, how you understand it, how it can be used if the people who developed the algorithms who essentially have that power and they're relying on it."

However, this could also be turned on its head to diffuse responsibility, by implying that people who were cautious to accept AI outputs might just be insecure and fear change:

> "Typically, as humans there's kind of a natural thing that, we are kind of unsure and insecure about things that we don't understand… with AI it's just a computer makes a decision, I don't understand how it works, how did it come to that decision? And I think the fear of the unknown plays a big part in this."

This reframed what could be a legitimate concern (especially given previous scandals) into a general cognitive bias against the unknown, which enabled a dismissal of the underlying issues which necessitated these fears. This perhaps sheds light on how the value of "trust" is interpreted in relation to AI, as willingness to use a system is influenced by oft-interchangeable underlying factors (Glikson and Woolley 2020), with this "willingness" pushed to the fore. If you consider this view of trust in light of a predisposition to see user qualms as unwarranted, then building trust is oriented towards building a trustworthy aesthetic – a kind of paternalistic positioning of AI that legitimises the obscuring of the processes. In addition to making a further case for a process-based ethics which can be embedded in processes of design and development, these topics also raise the crucial consideration of the ways in which power dynamics implicitly and

explicitly shape conceptions of responsibility, and responses to it. Considering this, I turn attention to the roles of governance and regulation as modalities through which ethical responsibilities are construed and managed, as discussed in the next section.

Governance, Regulation and Culture

I noticed a fundamental scepticism regarding external governance, whether this be due to the ease with which the realities of practice can be obfuscated, or the slow and limited capacity seen as characterising regulation. This scepticism was expressed to differing degrees. George believed that regulation could potentially be a way to ensure ethical design and development, but was pessimistic about its actual practical impact, commenting on whether "*this should be policed or regulated… that would be great, but I look at the practical aspects of that, and…I just don't see how it could ever work…*" Similarly, he argued that regulation such as GDPR was only of use to very basic levels of data protection, but lacked the necessary nuance in addition to the type of power which would be necessary to leverage in order to result in actual change, describing regulation as *"a very blunt tool to enforce things … it only works if it's got a very big stick behind it."* In a similar study, Orr & Davis (2020) investigated AI practitioner perspectives on the loci of ethical responsibility in AI systems, and the responsibility practitioners believed they should bear. Noting the complex network of relationships impacting practitioner ability to transform guidelines into practice, they found that practitioners largely placed responsibility upon policymakers and regulators, whilst acknowledging the importance of their own expertise. This rings true to the research reported here, however, this was tempered by concern about the ability of policymakers to respond to innovations in the field. Indeed, attempts to mandate accountability might be undermined due to underlying attitudes and seen as an attack or attempt to limit the freedom of practitioners which would respond by developing new models, as Jason described:

> "it's this paradox where, because this technology is breaking new ground, where existing laws can't keep up with it, the only people who have intuition as to how it could be regulated in a positive way, are the people who might not be

> benefitting from such regulation… ultimately I think that if people aren't willing to self-police, then it's difficult to regulate because a lot about these algorithms are very slippery in terms of… using that example of discrimination, and neural networks…they quite cleverly avoid what can be pinned down, and it's so easy for them to develop around these regulations, make another algorithm to circumvent them"

These observations align with recent external analyses of the limited usefulness of governance in directly dictating model development. Bechmann and Bowker argue that "The speed by which processing takes place…makes it difficult to govern these algorithms and services favor effective and fast models and processes over standards, balancing tests and documentation" (Bechmann and Bowker 2019, p.7). In keeping with this, Skye framed regulation as a threat which could be dealt with or avoided by changing the details of practice and developing new types of models, depending on the circumstance, contrasting it with creating a culture of open ethical reflection*; "with regulations it's always big scary things that you try and avoid or tiptoe round or dodge them but with a culture, if everyone's bought into the culture, it's a lot better"*. Distributed responsibility added a further spanner to the works, casting doubt of the very possibility of pinpointing responsibility, obscuring accountability. Therefore, whilst regulation was generally perceived as desirable in theory, its practical value was seen as shrouded in ambiguities at best.

This fear is nothing new to AI. Simon pinpointed how "difficulty to attribute responsibility and to locate accountability" (Simon 2015, p. 145) can result in fear around sociotechnical systems, arguing that the complexities of computational systems represent new kinds of epistemic responsibilities, or responsibilities concerning knowledge and processes of knowledge construction and attainment. The locus of this responsibility was usually deferred and abstracted away from the practitioners themselves, pushed up to the upper echelons of the social-economic structure. At the highest level, AI was construed as a public service, with responsibility for algorithmic systems belonging outside companies and in the realm of government. To use a specific example, Lukas argued that public interactions with Big Tech platforms should count as

a public utility and be regulated as such, although where they would allocate responsibility is unclear. He compared large technology corporations to utilities, comparing Facebook to water, an essential resource which cannot be avoided and therefore must be tightly regulated by an accredited body:

> "To some extent it's something like a commodity like, that's a stretch but like a water supplier right…you need to get water, yes you can choose not to buy the water if water is dirty, but you need to buy it, and so the government makes sure the water is clean. But I think there should be something that should apply to those large websites like Facebook or Google."

Indeed, in some cases AI was indeed inputting to public services, a concept which Stefan found incredibly worrying in terms of the potential risks and harmful societal implications:

> "There's a load of potential doing damage [with AI]. Yeah… I mean you can so easily destroy someone's life. Take away all their opportunities on the basis of the Social Credit system they happen to randomly fall foul of."

He expanded upon with a belief that the social implications of these applications of AI equalled the atom bomb in terms of potential risk and damage. Despite this potential for harmful ramifications, Stefan was very sceptical that responsibility would be taken seriously within the wider industry and field. Other practitioners were also pessimistic about responsibility in the field, expressing views that those who most need to engage with "the spirit" of any self-policing or regulation, would instead be at odds with it. In essence, practitioners felt that ethical governance was essential, especially given the discrepancy between often high-stakes of application and the ad-hoc nature of implementation but were sceptical about the ability of regulators and policymakers to provide and enforce this. Complementing this they felt that a focus upon [re]building a more ethically responsible and epistemically rigorous culture was necessary, and that this required cultural change which was spearheaded from leadership as moral exemplars.

Direct and Indirect Impact

Zooming out to the realm of the indirect stakeholder, another site of distributed and obscured responsibility was in the relationship between the team creating a model and their client. Joshua discussed the delicate discussions which his leadership team had around who to partner with in future: *"there might be a conflict where they say, well you could get this partner on board…but ooh they could do something horrible with it"*. For certain businesses, the AI system is a bespoke product created for a third-party, who then employ it for decision-making about their consumers (retail), applicants (hiring), patients (medical) and so on. Eddie (AI team lead in an industry organisation) discussed how this affects the degree of power he held over the process, expressing concern over the ambiguity it introduces to the process of assigning responsibility to the parties involved. They contrasted between duty to communicate if a model is unambiguously unethical, and being accorded responsibility if it behaves in an unethical way or is used in an unanticipated way:

> "I haven't decided who would be responsible for the model. You can imagine a situation where a model is awful, like ethically awful or it's unambiguous saying that this group of people should have something happen to them – Is communicating that to the customer enough for us to wipe our hands clean – I don't think I would be comfortable with that. Does that mean I think [company] should be held personally responsible for the thing even if we've done our best to avoid it or to communicate it back – it is their business. They're the ones applying the model surely, they should be accountable for that, especially if we have communicated to them."

These sorts of observations revealed a legitimate ambiguity and concern. They also suggested a view of the world that as long as the status quo was upheld, then no blame could be assigned, and that challenging existing biases was not considered as part of their purview or even understanding of the world. This contributed to the previously mentioned sense of inevitability of harm. Dewi, reflecting on a recent Big Tech scandal, and whether something similar might happen in the healthcare domain, looked visibly

affected and told me that *"it has probably happened already, and it will happen in the future…unfortunately humans are humans"*. This was not helped by the blurring of lines between technical artifacts and the humans interacting (or intra-acting) with them. Cameron reflected on the difficulties this posed for being able to evaluate the ethical implications of an algorithm:

> "One problem I had, I tried to come up with a list of, a very practical way of evaluating an algorithm ethically. The problem is, you end up evaluating the people, not the algorithm."

This socio-technical constellation included humans at all stages of the process, beyond users or affected groups to organisations who might potentially misuse the model once already created. Thomas summarised the socio-political implications which would potentially result from this sort of data coil:

> "Essentially, more and more people who do have access to the information have more and more power, and then value is then entered into this feedback loop of how people use them, and how people come to rely upon them."

Similarly, Orr and Davis (2020) identified the locus of responsibility as distributed across complex and dynamic user-AI constellations, suggesting that the nature of AI ethics in constant flux. This distributed responsibility might extend to the out-of-sight end-users; Cameron brought up the balance of responsibility between the creators of a technology and the users of this technology. They suggested that the company should not be expected to take all the blame for the way certain users behave on a platform, raising the point that a system can simultaneously be harmful and greatly beneficial, and that changes have ramifications beyond those borne by the company itself. These are impacts which are very difficult to adequately account for:

> "The other part of this is shared responsibility. So, we have an interesting situation where, so assume Facebook became a terrible threat to human life

> because of the way that people behave on it… the company would take some blame, but also the people."

Indeed, ultimately, many of the discussions I had on this topic agreed with this statement that responsibility should be a shared matter. However, as discussed, the details of how exactly this might be shared in practice was mired in different considerations and deflections. Furthermore, often this omits consideration of the sort of social looping effect discussed above, and consideration of the power dynamics shaping creation and maintenance of these loops.

Approaches to Navigating Responsibility in AI

Practitioners wanted to ensure that AI, both the systems they built and the concept more broadly, was designed in a way which was fair, and even empowered people affected by them. In this way, their values were the basis to engage with and potentially restructure existing AI practices, a perspective rooted in their recognition of and response to the impacts of such practices. Their perspectives on and engagement with responsibility emerged through this mediation of values in practice, both shaped by broader societal contexts and constraints.

Fairness and Ethics

A narrative of concern for fairness ran clearly throughout the numerous practitioner accounts detailed thus far, alongside deep-seated concerns about the impact of practitioners and system constraints on equity and bias. Reflection on fairness was often regarding concern about societal impacts of the models themselves. However, these expressions of concern for equity were tempered by recognition of the impact of power imbalances on the development of AI, described using metaphors of currency such as Thomas described *"in terms of who has access to these algorithms, to run them and interpret them; who has the power"*. Indeed, many of the topics discussed in the interviews and observations were interwoven with concerns about the impact of AI systems on society, such, exemplified by Mark's view:

> "I do think that machine learning could have major social implications - I think those are the more important, from my perspective, those are the most important questions".

Despite strong beliefs in avoiding harm, practitioners simply might find it extremely challenging to enact this, illustrated here by a discussion of the complexity of negotiating ethics in small teams. The co-founder of an AI start-up, Lorenzo spoke about his experience of discussing whether to do business with organizations they themselves view as unethical, but others on their team were more concerned with the potential for a lucrative contract:

> [I was] "recently being approached by [military organisation] to say, you know, we see what you're doing, we've heard a lot about you. Can we have a chat about your work and what applications it has for military use?"

This personal belief, that it would be unethical to partner with a military organisation, became difficult to materialise when brought into the context of the wider team they worked within; "*people have different views, and different tolerances, if I can call it that, for what they view as ethical*". Joshua described a similar experience of how a military organisation approached him, and then reflected on his ensuing process of moral deliberation; *"[military organization] offered me a project. It was basically for a defence system, and I wasn't really interested in contributing any intellectual value in that space. So, I rejected that and ended up working on a project in distribution systems optimisation"*. As a solo worker, Joshua expressed more freedom to act in accordance with his values than Lorenzo had felt able, having greater capacity to enact his moral freedoms. This raises the consideration of how practitioners balance different types of responsibility, especially when in positions of leadership, with responsibility for immediate concerns potentially mediating a diffusion of perceived social responsibility. It also highlights the interplay between contexts of practice and the moral values of the practitioner, a theme which I explore further in the following chapter.

Moral Exemplars, Training and Discussion

Heimer *et al* (2013) observed that fairness and moral relevance were key factors in the practice of ethics "on-the-ground", with participants citing exemplars of moral leadership as inspiration for action. In addition to be guided by exemplars, Heimer *et al* observed that this ethics on the ground was conducted through personal interactions and internal organisational cultures rather than being recorded officially. They also found that ethics-on-the-ground was focused on *doing* ethics, usually in the form of conversations, rather than just signalling it, which was the case for official ethics. Official ethics focused on the needs of the project, whilst ethics on the ground focused on the needs of the patient and the multiplicity of their relationships and roles - official ethics "abstracts out this complexity" (Heimer *et al* 2013, p. 375).

One way in which practitioners felt an AI ethics-on-the ground could be built up was embedding ethical reflection through training materials, an approach since adopted by Harvard in response to student demand (Grosz *et al* 2019). Another idea was regular workshops habituating industry practitioners in good moral practice. This was seen to be an interactive, relational practice. Julie reflected*, "I don't think that we as machine learning researchers can have the entire burden of this ethical discussion, but we should be poked and start a conversation with people that know about ethics"*. It was for this reason that Van Wynsberghe and Robbins (2014) suggested placing ethicists in the research lab alongside practitioners.

Threaded through the discussions of practice, and supported by the related literature, these findings indicate how ambiguities pervade practice and shape the realities of ethical reflection and engagement at all stages of development. This is not altogether surprising. Ambiguity is pervasive across all domains of practice, from finance (Best, 2012) to medicine (Domen 2016; Luther and Crandall 2011) and has been construed as a central facet of organisational practice. Aspects of it have already been referenced indirectly in various ways in AI. Discussions of the risks of black-box systems often touch upon ambiguity, as do attempts to tackle distributed responsibility. Similarly, the issues caused in placing ethicists as experts, changing the focus to mainly outcomes rather than process, and resulting in overestimation of and misunderstanding of their role, is also seen elsewhere.

In this chapter I examined how practitioners conceptualised the nature of their own responsibility, and how they understood the implications of the socio-technical entanglements that they created and worked within. Practitioners were keen to better understand responsibility, trying to pinpoint their own roles to play in ethical decision-making, while navigating the complex situations and sometimes competing motivations which were discussed in the previous chapters. However, this was complicated by several factors, including individual apathy, tactical ignorance, industry logics and unpredictable material practices. Our discussions of responsibility and accountability often touched on issues of accessing, processing, and communicating data, and associated knowledge and information, however, such responsibilities might be seen as abstracted from practitioners. Complexity and ambiguity often obfuscated the locus of responsibility, especially given the distributed nature of AI work, including stakeholders beyond practitioners, in various permutations and arrangements. Furthermore, combinations of contexts, constraints and values mediated responses to responsibility, serving to diffuse notions of accountability. Practitioners believed ethical leadership was important, focused upon actions and building a cohesive and coherent ethical culture, rather than empty value statements. Meanwhile, concerns about a lack of understanding of ethics impacted practitioners' sense of being able to respond to moral issues in their practice. However, practitioners were hopeful for cultural change, proposing a process-based relational approach to engaging with ethics in practice. The potential for changes to be realised from this relational approach nevertheless hinges on practitioners' own values which shape their engagement with their work (a directed expression of their moral freedoms) which is the focus of the next chapter.

CHAPTER 6: Navigating Values

> **[** During one of our many coffee break chats in the week at the research lab, Remi spoke of a passion project. She was helping to organise an AI summer school, with the vision of creating a space where people living in the Global South are taught AI by world leading experts. The hope was to build competencies in AI which would facilitate local agency in decisions about the models built and employed in Global South regions. Remi spoke at length about how entire regions were prominent as sites of AI yet had little active involvement in how these developments were framed, developed or deployed. As we finished up our coffee, Remi also reflected on how diversity in the Global North also left much to be desired. On a later date, she met with other organisers of the summer school, who were also working at this lab, to plan the upcoming events of conference, encouraged by the project leads managing her internship.]

Designing and developing algorithms is engaging in meaning-making (Ruthven 2019), constructing artefacts with implications for human decision-making and freedom to act. In this way, we have a fundamental responsibility to consider which meanings we create, and the implications of these. As I explored in Chapters 4 and 5, values and practices are intricately interwoven. However, although we have aspects of our "…existence that are situational and factual" (Riggs 2019, p. 6), our actions also surface a desire to overcome these constraints to pursue our motivations and realise our values. For this reason, in this chapter I draw on de Beauvoir's conceptualisation of values as internal and personal, with externally enforced values having little meaningful impact on the person engaging with them unless said values are also intrinsically motivational[8]. From this perspective, the subjectivity of values means that no universal code of values can be sufficient. Indeed, empirical studies of micro-ethics, as discussed in Chapter 3, have illustrated the divergence and tension between official frameworks of ethics, and the personal

[8] However, as discussed in Chapter 4, external forces impact which values are foregrounded.

engagement with ethics on the field, with external frameworks having a negative impact if personal practices are not accounted for (Heimer 2013, p. 377).

There is an embodied materiality to AI practice, for all the narratives invoking abstraction and objectivity. Examining the confluence of practice and values reveals an approach characterised by curiosity, challenge, and craft. Taking this into consideration, I employ three different lenses to explore value motivations and mediations, examining the artistic practices, guiding narratives, and reflexive tendencies of practitioners. I then draw together findings from Chapter 4 and 5 with the present discussion of values, to consider similarities between art and AI practice.

Values can be end-goals in themselves, or instruments in service of these terminal values. Equally, AI can be the site of a value, or a way of engaging with or achieving its aim. I unpick some of the ways which practitioners utilised AI to mediate realisation of their values, and other ways in which AI was the focus of these values, discussing interrelated themes of social good, problem-solving, intellectual curiosity and self-direction. To illustrate this (and given the interrelated nature of the main themes of this thesis), I draw upon findings from the previous chapters which looked at the contexts and constraints which shape AI practice, and reconsider this in Chapter 7 which re-examines practitioner conceptions of responsibility and ethics-in-practice through the lens of *The Ethics of Ambiguity*.

The values motivating practitioners are small pieces of a messy, uncertain puzzle, used to guide ethical heuristics to try to navigate the distributed nature of the responsibility webs they are embedded within. Beyond understandings of values in the moral sense which, simply put, consider how attitudes and beliefs of practitioners were aligned with ideals of how their work may be deemed good or bad, this chapter also examines epistemic values which focus on how practitioners' engagements in their work were framed to contribute to advancements in knowledge and skills within their field. Whilst there have been historic attempts to separate out 'rational' from 'moral' values, feminist epistemologies do not create this boundary, considering that our epistemologies are fundamentally situated within our individual, contextual standpoints.

In this way, epistemic values are even difficult to truly separate out from moral values, when one accounts for the nature of epistemic responsibility.

The Artist and the AI Material

The complexities and ambiguities of technologies have long prompted comparisons with art, after all, "discovery requires aesthetically-motivated curiosity, rather than logic" (Smith 1977, p. 144). This aesthetic refers to the experience of the practitioner, the [post]phenomenology of technologically-mediated exploration. Indeed, these motivations of intellectual curiosity, exploration and challenge characterise the discussions I had with practitioners. As we saw in Chapter 4, the practitioners I spoke with primarily used the language of art in descriptions of their practice. Alec was *"constructing pictures"* from the data he worked with, whilst Lukas sheepishly told me that his work on foundation models was driven in large part by "*intuition"*.

The Values of Art and AI Practice

In the same vein as art, AI "contributes to knowledge production by exemplifying aspects of the world that would otherwise go overlooked…inviting novel juxtapositions" (Gorichanaz 2020, p. 2). However, art goes beyond this to explicitly engage with "exposing and even challenging societal assumptions" (Gorichanaz 2020, p. 2), whilst AI practice largely conceals this, often in service of a legitimising narrative of objectivity. This implicit fusion of moral and epistemic values, often in the form of innovation, creation, and knowledge production, formed a core motivation to pursue the roles and types of work which practitioners undertook. Dewi was motivated by *"coming up with innovative methodologies",* and Julie described how she enjoyed *"coming up with new ideas of how to do things"*.

These intuitive processes were paired with desire for self-direction (freedom of thought and/or action). Stengers (2000) examined the efforts of scientific communities to preserve autonomy and demarcate boundaries to abstract their work from political and social concerns, preserving the aesthetic of objectivity. Stengers noted that this results in inherent tensions which renders the scientist as a "vector of a creativity" which

would be antithetical to a more critical stance (Stengers 2000, p. 5). In a study of AI artists, Stark and Crawford observed a similar dynamic of perceived moral exemption for the AI artist in pursuing their open-ended, independent aims (Stark and Crawford 2019). Indeed, self-direction heavily influenced practitioner choice of roles and projects; they sought freedom of expression, facilitation of creativity, and to explore the problems which they are most interested in. This could be materialised in the form of founding a start-up or pursuing academia. Lorenzo shared how he was motivated to set up his computer vision start-up because he desired self-direction in both his thoughts and his actions, *"what motivates me is …creating a place where you're working on something that interests you."*

As discussed in Chapter 4, comparisons of AI practice (including data work) to craftwork span several decades (Suchman and Trigg 1993; Thomer *et al* 2022). There was perhaps a tension between enjoyment of this craft of AI, of sculpting data and models, with perceptions of data work as a distinct task, I suspect impacted by "residual conventions and perceptions in AI/ML drawn from worlds of 'big data'…and of viewing data as grunt work in ML workflows" (Sambasivan *et al* 2021, p. 2). In essence, in the chimera created by combined understanding of AI as both art and science, there is a tendency to see the products of the technological systems through the lens of human transcendence while obscuring the processes behind them. This can be unintentional, through use of ambiguous language, invocation of unidentified underlying assumptions, or the [mis]use of descriptors intended to invoke ideas beyond the actual capacities of such systems (Phan and Wark 2021). Furthermore, as I have touched upon, there is a cultural tendency in the field to enjoy this artistic process but feel the need to conceal its existence to provide a singular narrative of the output. Such framings of AI as a singular output can serve to obfuscate and diffuse accountability in a process better characterised as creative combination and crafting in a process of exploration.

Artist approaches to the ethics of aesthetics can perhaps shed some light on alternative modes of engagement. After all, the art domain has itself seen a journey from dominant, even "propagandistic", narratives to the critical, heterogenous characteristics of the present (Gorichanaz 2020). Conversely, art can shape engagement with ethics.

Though a direct attribution of moral improvement to aesthetics is an over-simplification, it can potentially serve to broaden our moral horizons by extending our understanding (Kieran 1996). Stark and Crawford (2019) reported on AI/data artists describing a thread of ambiguity running through their practice, from conception through curation to the situated interpretations of their audiences. They employed two conceptual lenses to examine how AI artists navigated resultant facets of ethical responsibility, noting how they balanced the "moral good" of educating audiences via disruption of existing narratives, with the unintended impact this might have (Stark and Crawford 2019). Noting the implications of ambiguity and attitudes of being a "moral exception" for AI artist engagement with ethics, Stark and Crawford employed Walter Benjamin's critique of the "aestheticization" of politics via technologies to reflect on the boundaries between which the interviewed artists walked, pointing out the risk of heightened narratives in facilitating fascistic politics (Benjamin 2008). The same can be argued for some of the grand narratives of AI referenced in Chapter 4. However, as Stark and Crawford noted, narratives can also serve an important function in facilitating ethical reflection.

Heuristic Mediation of Values

In their intuitive approach to ethics, practitioners crafted imaginaries, cultivating their processes of ethical deliberation, and understanding, reminiscent of the role that art can play in broadening moral imagination (Kieran 1996). These heuristics enabled practitioners to transcend the limitations imposed by the autonomous character of the field, to approximate the situational contexts deemed necessary for ethical deliberation. Alec's expression of the roles of curiosity, self-direction and ambiguity in his work led to a discussion of how values might be made easier to reflect on in practice, and thus incorporate into model design and development. He emphasised that despite the complexities posed by ambiguity, the answer was not to add more software, *"I'm not sure I'm always a fan of having software, 'a tool to solve everything', whether it's more like, you establish values by doing stuff, by acting in a certain way and not being a dick, you know"*. He used the example of human decision-making to drive home the view that rigid ethics was detrimental to practice, pointing out the risks of such an approach. In such an unpredictable, exploratory space, Alec felt that:

> "It might just paralyse everything - if you weren't allowed to use the nasty tricks that your brain uses then you just wouldn't be able to do anything. Or you'd end up with that kind of, like a real kind of computer box-ticking exercise where if you don't exactly fit the criteria then you can't get through, which is a bit too rigid. You need to have workarounds, and fall-backs and heuristics which always work or whatever."

Rather than inflexible and abstracted approaches to ethical deliberation, he raised the concept of ethical heuristics. This concept of ethical heuristics arose across different conversations with practitioners, employed in the absence of more concrete approaches to ethical reflection and decision-making, and/or as a complement to professional ethics and compliance. Here I focus upon the most salient of these heuristics, an approach involving placing oneself in another's shoes (cognitive perspective-taking), perhaps to anticipate the emotional experience of the other (affective perspective-taking), although this emotional element was not necessarily an explicit aim. Skye described analysing how her decisions in designing and building models would impact her personally:

> "There's a lot of considerations that as I'm building things…always in the back of my mind always is if I was the user of this, would I like the way that it's being built, and like would it benefit me, would it disadvantage me, um…so I think that's yeah, always playing on the back of my mind."

This perspective-taking approach, speculating in a way which stimulated empathic concern, involved anticipating harm by making harms and benefits emotionally salient. Practitioners did this by putting themselves or others in the hypothetical situation (Young and Koenigs 2007; Huebner *et al* 2009; Conway and Gawronski 2013). It could take two broad forms, either empathising (trying to place oneself in a situation which affects another) or sympathising (trying to understand the experience of another in a given situation). Thus, I employ the terms 'empathetic anticipation' and 'sympathetic

anticipation' to describe the instigation of concern for others welfare (Decety and Cowell 2014). Dewi – in an example of sympathetic anticipation – described imagining family or friends as the beneficiaries of his work, making clear that he used this method to engage in moral decision-making:

> "…my parents are patients, my friends are patients. So, it's constantly thinking about what is the ultimate goal of what we are doing. Considering all of these things, it's a second nature, it's kind of natural that, you know, ethics to me is…is something obvious I guess."

In constructing an imagined scenario, usually based on the experiences of the practitioner, practitioners created an opportunity for anticipation and a point of critical reflection on ethical practices in AI system development. This method of empathizing with an unseen third party has been linked in previous research to ethical decision-making (Hoffman 2001). Mencl and May (2009) demonstrated that ethical decision-making was influenced by psychological proximity, and its impacts on empathy, as well as physical proximity.

These approaches demonstrated eagerness for engaging with ethical reflection and other ongoing processes of decision-making in a way which explicitly considered impacts. Moreover, these were embodied approaches to deliberation, which can confer certain benefits. Impersonal choices are argued to activate utilitarian responses, whereas empathy can short-circuit this and elicit more situated decision-making, with empathic concern perhaps even being crucial to moral decision-making (utilitarian tendencies may be indicative of diminished empathic concern) (Gleighgerrcht and Young 2013). This impact of distancing on ethical decision-making is of particular importance when we consider the constraints already imposed by distributed modes of responsibility.

However, empathic anticipation is imbued with the biases of the empathizer; Kristoff reflected on how the outcomes of the algorithms he builds are a direct representation of his worldview and moral character in stating:

> "a lot of it is from my own personal experience which is, yeah comes back to like when I'm building a model, essentially what model the algorithms outcomes will

> be, is probably a very similar reflection of what my decisions would be, same as any other designer or developer of those programs.".

This reliance on "*personal experience*" has been demonstrated as a potential impact of employing empathy in ethical decision-making, "introducing partiality, for instance by favouring in-group members" (Decety and Cowell 2013, p. 337). It serves as a useful, usually intuitive, tool in the day-to-day arsenal of the machine learning practitioner working on a system, but it has flaws from the limited perspectives of the ethical reflectee. Consequently, it is limited in scope; by the limited project overview a practitioner may have when working in a larger team, and the limits of empathy itself. These limitations were picked up on by Skye who valued external input:

> "...also giving you know buy in from the entire team rather than just be my personal judgment because you know we all have our own individual biases and the more minds you can get on something the better decision you can arrive at."

As examined earlier in this chapter, practitioner values seemed often activated towards creating novel systems to "promote good" and minimize harm, within an existing structure, rather than critiquing the potential for good and for harm of the system being created, and challenging underlying assumptions of existing practices. Empathic morality is a prominent example of how interlocking factors of values and ethical frameworks can serve to reinforce existing norms, resulting in a reductive force which impacts the framing and scoping of AI projects. When legitimised, this amalgam risks formalizing individual or group biases, instantiating and concretising system rules which are inequitable and perpetuating epistemic injustices (Abebe *et al* 2020, Bennett *et. al* 2023). This is compounded by the lack of diversity in the industry, where researchers are primarily white (West *et al* 2019), and females making up less than 14% of AI researchers in the UK (Stathoulopoulos and Mateos-Garcia 2019). As a result, while such systems aim to be beneficial, the attempt to circumvent the privileges of being-in-the-room through leveraging empathy to account for the missing voices can both lead to moral

harms due to oversights and entrenches epistemic biases (and injustices) which are built on in subsequent practices.

Discovery and Integrity

Although moral and intellectual values are often framed as distinct, and even in tension, perhaps this is an artificial boundary which is unhelpful. It seems that the connection between epistemic and moral values is very straightforward in AI where the epistemic values of the practitioner result in knowledge which has vast potential for direct impact on the wellbeing of others. In examining the relationship between narratives of science and art, Shapin (2018) compared the language of discovery employed in science, with that of inventing, constructing, and making seen in art (Shapin 2018, p. 177). Indeed, the abductive approach characterising AI has been described as a "logic of discovery" (Niiniluoto 2018), the afore-mentioned problem-solving and curiosity utilising the materials of AI to discover knowledge or improve the process. Skye was "*fascinated about scientific decision making and uncovering the truth of how things work*". However, this truth was subject to the artistic process described in the previous section. Jason described his typical process:

> "You come up with some mathematical idea, but then you try it out on some benchmark tests and if it gets better then it gets published obviously. But in order to get published it needs to perform better which means, well, no algorithm ever works better on the first try, so you keep working on the algorithm and you keep maybe even changing the benchmark…"

He reflected on the integrity of this approach, often feeling concerned that it involved some sort of self-deceit, and therefore a broader deceit to those encountering such work in other contexts:

> "…you could say its, well it is all to make the algorithm better and figure out if it's better, but at the same time it is important to reflect about whether you're still doing solid reproducible research or trying to I wouldn't say cheat, but it's kind

> of overfitting to what you're doing… you believe it's useful because you've seen it succeed on the dataset but you kind of neglect that you tried it out on 5 others and it didn't work."

This sort of accidental self-deceit was a concern of McDermott (1976), who used the example of misleading naming of AI approaches. In addition to misleading those outside the field, the practitioner themselves began to believe the implications of their naming, due to associations which were carried over from other uses of the terms; in McDermott's words, "concepts borrowed from human language must shake off a lot of surface-structure dust before they become clear" (p. 5), instantiating further ambiguity. In choosing how to frame practice, epistemic values were interwoven with the moral values which spurred this work; for example, the drive to produce a moral good (outputs that improved society) paired with epistemic good (improving knowledge and ways of working with AI, which could potentially lead to improvements of outputs). Therefore, in obfuscating the epistemic value through obscuring the process, it carried the potential to impede the moral value of the work.

Practitioners wished to work on tasks which were intellectually challenging, engaging in "thinking that aims to overcome barriers and to reach goals in situations that are complex, dynamic, and non-transparent", (Guss, Burger and Domer 2017, p. 851). Joshua told me "I like solving problems… I've always… [liked solving problems]", and Ariel (AI consultant in a leading consultancy firm) explained how a perk of working in AI is that "*you always have the reassurance that you've pushed the boundaries of human knowledge*", where improving knowledge meant improving the ease at which it could be generated and accessed by others. The desire to contribute to knowledge and facilitate processes not only informed Ariel's career choices but were integral to her interest in the domain, "I always wanted to find a better way that doesn't fail so much".

Meanwhile, Skye modified her description of enjoying problem-solving to include that she liked "*trying to find optimal solutions*". This focus upon optimisation made clarity desirable, generating a discomfort with any complexity of problem or context that was not considered readily tractable. George summarised this by saying "*I like the fact that*

it's very black and white, it either works or it doesn't, there's not too many, yeah like compared to other domains where you do work but it's all a bit grey". That is, that although the journey and outcomes are uncertain, the immediate feedback was clear. Here we see an example of a practitioner navigating the ontological ambiguity observed by Hoffman, the ambiguity faced by practitioners in trying to apply the abstract concept into a tangible output (Hoffman 2017). This difficulty was often managed by favouring or insisting on the pragmatic applications of the outputs created by the unclear processes or problem definitions that preceded them, a sort of retrofitting of the product to make it seem like a clear "black and white" solution to a post-hoc defined problem.

This approach to tackling ontological ambiguity persisted into certain practitioner perspectives on how "official" ethics should look, where I noticed examples of thinking transferred from AI practice, viewing the ambiguous as something that needed to be combatted with solutionism. This included discussing documentation of ethics in terms of concepts more readily associated with scientific practice, and with approaches to dealing with epistemological ambiguity concerning validity (Hoffman 2017) such as weighting, classification, mitigation, boundary-pushing and focus on constructing testable hypotheses, as seen in the comments of Adrian (an AI researcher in a multinational AI company):

> "I'm not very scientific about it [ethics]... I don't currently follow a structured process, but I would ideally like to have some kind of, like, an ethics checklist for a project and put a weighting to the mountain of my various concerns that are possible, and explicitly write about mitigating factors."

This might be partially due to lacking appropriate vocabulary to otherwise describe these sorts of considerations. Skye discussed her discomfort at lacking the terminology to feel able to tackle ethics in a way which she felt held the same weight of the methods she had been trained to employ to manage the ambiguities inherent to AI practice.

> "These would be helpful because we're not trained… it's, like, you do a degree in data science, and you learn all about the coding and how to build a model. But you need to be able to… I had a few lectures and stuff on ethics, and like

> obviously it's drilled into you that you need to use tools responsibly, but yeah you never really get told what's right and what's wrong, what's ethical…what's not."

Such descriptions might suggest a subtle dynamic of abstracting the responsibility from the practitioner, that rather than making inherently ethical (whether judged as "right" or "wrong") decisions during their practice, these decisions can be distanced as impossible to engage with if not suitably couched in specific terms. The epistemic and the moral overlap here, with the attempted collapse of complex human ontologies, power dynamics, relations and in the process moral implications, into data points, terminology, and checkpoints. Still, given the emphasis of many approaches to ethics on adaptation of abstracted, universal principles into concrete, tractable problems, it would be unfair to lay this solely on the backs of the practitioners themselves. Abstract terms couched on epistemological lineages pose a form of gatekeeping effect to engagement with ethics; it contributes to a presentation of ethical reflection as belonging to a domain that requires specific forms of knowledge or conceptualisation, which practitioners may hesitate to approach given the potential for harm if they fall short of expected domain expertise (and the capacity for diffuse responsibility by deferring to experts). Yet these tensions between the roles of AI practitioners and ethical practitioners plays into views of how "right" and "wrong" practices were to be evaluated – often through privileging the issues deemed to fit within their own domains of intellectual expertise. This informed another site for outlining practitioners' values, which I discuss below.

Epistemic Vice and Intellectual Honesty

Concern about intellectual values extended beyond the personal ethics of practitioners, to form part of a commentary on the perceived ethics of their colleagues (within teams, labs, and companies), and yet further to others in the field. Perhaps indicative of the hierarchies and power dynamics mentioned in Chapter 4, Kristoff was concerned about a sense of apathy or even derision of ethics within certain parts of the field, which he described as motivated by "arrogance", a view that people working in AI-adjacent fields such as ethics were "wanna-bes" who did not make the cut for "real" AI:

> "I recognise that a lot of people have huge egos, and I've heard people talking about people who work around the field of AI as people who wish they were in the field of AI, and they're too stupid…literally, quote 'too stupid', to know exactly what is going on, and therefore they're just 'attention-seeking'."

This intellectual hubris existed within a space which was also described as a ticking time-bomb due to academic fraud, which has even seen articles professing that "exposing the behaviour of a community of unethical individuals will encourage others to exert social pressure that will help bring colluders into line" (Littman 2021, p. 43). Responding to this, Buckman (2021) expanded the critique of the community to explicitly address the content of AI papers, rather than focusing on the relationships between specific reviewees as Littman did. Curious to learn more about these concerns, I spoke with another concerned member of the AI academic community via Zoom shortly after these articles were published (name withheld to protect their identity), who expressed a strong desire that the epistemic culture of the field change for the better, and even felt that the only response to the dire state of publishing was to "burn the whole thing to the ground and start again". They cited the example of the "replicability crisis" in the field of psychology (which the field was still reeling from), characterised by issues such as widespread cherry-picking of data to achieve statistical significance (Wiggins and Christopherson 2019).

Martin was keen to discuss what he viewed as a widespread deficit in epistemic values, speaking of how "*most of the things I see in people in the market are not true.*". He felt that others in the field were concealing faults, overemphasising features/successes and even misleading the clients who bought their AI models (as Jason alluded to earlier in the chapter, when reflecting on his own practice). Furthermore, practitioners felt that their colleagues' perspectives on ethics presented a barrier to engagement. Kristoff also spoke about the culture of dismissing non-technical expertise, where arrogance and epistemic injustice (favouring certain forms of knowledge and practice over others due to systematic biases and social injustices) resulted in dismissal of ethical concerns.

> "yeah, I think that the biggest hurdle in applied ethics, is making the community appreciate the need to look into applied ethics more and to educate, and to get over that barrier of arrogance that is particularly predominant in the field."

As I continued to interview practitioners, I came to expect the recurring refrain of "ethics is important to me, even if I can't see direct ethical implications in my work", complemented by the profession that "I feel alone in this though, other practitioners aren't very bothered by it". Indeed, the practitioners who described themselves as motivated by truth-seeking expressed concern about the actions of other practitioners, describing the deficits in intellectual traits which they attributed to fellow researchers, akin to epistemic vices which are "character traits that impede effective and responsible inquiry" (Cassam 2016, p. 159). Perhaps this was partly due to recognition of their own susceptibility to contextual pressures and constraints.

Narratives, Values and Ethics

As illustrated throughout the thesis, and particularly in this chapter, narratives surface values (Diochon and Anderson 2011) and are employed as tools to mediate them. Guiding lights in an ambiguous space, narratives provide a tangible vision of the values shaping practice, and a powerful method of transcending facticities. They work by "…establishing a given endpoint and endowing it with value, and in populating the narrative with certain actors and certain facts as opposed to others" and, having done this, "the narrator enters the world of moral and political evaluation." (Gergen 2005, p. 7). In the previous chapters I sought to convey the complexities and nuances with which practitioners' described AI. These narratives formed different shapes, influenced by the contexts and values of their creators.

Harm and Good

Value considerations were highlighted in practitioners' accounts of how they navigated their field, in terms of where and how they chose to engage or disengage with different forms of work. We saw in the vignette how Remi noticed the considerable inequities in the field, seeking to address these inequities by getting involved in projects which would

provide skills to people in areas which had been historically marginalised. This motivation to work for some sort of tangible impact, expressed as a form of "social good", frequently influenced AI practitioners in choosing which projects to pursue, and which roles to take. This "good" could take several different forms including improving "wellbeing", perhaps being *"ultimately about how it can empower us to live a better quality of life"* (Alice – researcher in a Big Tech company). It equally could be realised with a broader scope, as in the case of Lorenzo's motivations:

> "For me, the essence of taking something out of the lab, and actually creating some value to society, uh, in a wider context, is a very motivating thing."

McDermott (1976) noted the importance of such narratives for strategic advantage, with these broader narrative framings intended to preserve the notion of working on an important question, "if a researcher tries to think of his problem as natural-language question answering, he is hurt by the requirement that the answers be the results of straightforward data-base queries" (ibid, p. 7). This is reflected in Lorenzo's rhetorical distinction between "the lab" and a societal "wider context", with an implication that these domains have differing requirements of the value of his work – the lab requiring a demonstration and development of specific processes largely to serve intellectual value, while the societal requirements were of a widely appreciable humanistic value. In employing this narrative of "taking something out", he is positioned as a mediator between these domains, leveraging the technical for the moral.

Meanwhile, Stefan valued avoiding harm to the extent he had consistently refused to work with companies and research groups which he saw as contrary to social good (for example those who just "make the rich richer"). This snapshot of him as an individual who shaped his career to reflect his values was put in sharp relief by his description of the case of a prominent AI practitioner who had a strong social media presence critiquing the ethical risks and pitfalls of AI. This social media personality was *"quite outspoken, especially during this whole Facebook crisis, about the risks of the misuse of technology…",* and in the example Stefan gave, had recently tweeted about the harms of Facebooks technologies, even though his actual actions were seen to contradict his

words. Towards the end of the tweet chain Stefan told me about, he recalled this practitioner saying:

> "'I'm so glad that I work for Google because Google doesn't have any of those problems, because Google's business aims are aligned with those of their customers'…which is just like such delusional horseshit…it's like, it's kind of baffling how people are able to rationalise these choices."

He interpreted this as a lack of critical reflection, as "a blind spot" which was described as unfortunately common in the community. He also identified a tension between high level values guiding discrete choices and the unpredictability of developing AI models. Values were useful for black or white scenarios but in an uncertain world, *"how is that going to be used by society, is that gonna end up being better, in the long term, or worse*", but Stefan was left reflecting that *"I don't know where that leaves me ethically"*. In fact, the impact of context could transform a tool which seemed to represent a morally good value into a negative:

> "Facebook was gonna be good because it, like, contributed to the Arab Spring and everything, and um, supporting device of democracy and giving power to the masses and so on, and now it looks like it's going to be a tool, a tool for the elite to control the masses".

Although pessimistic about the ability of individuals to have an impact even if they do try to make ethical choices, Stefan expressed hope that the culture in the community would change to be more reflexive, in *"a cultural shift, a shift in perspective among AI researchers and practitioners."* However, espousing values was not enough for practitioners to be seen as ethical, facilitating group awareness of ethical concerns was not enough and perhaps even counterproductive. Stefan acknowledged a risk of stopping at just discussing and acknowledging an issue, especially in spaces such as conferences, illustrated by a large panel session he had attended at a conference. Populated by academics and industry researchers, this session had focused on the ethics of working with autonomous weapons, where *"everyone still walked away with a sense*

of, oh we're taking the ethical implications of AI really seriously and we're really responsible", perhaps hypocritical given the context of a product designed to cause harm. However, this seeming hypocrisy might be less incongruent than at first glance. Rather, this scenario illustrated the empty nature of certain concepts such as "being responsible" in the abstract, which practitioners then project their own values onto. For example, an influential, emergent narrative of ethics has prioritised development of AGI, which is programmed to follow ethical theory, at the expense of addressing the current material inequities which AI applications potentially create and perpetuate. The "strategic ambiguity" of such vague concerns has then served as a "glittering generality" over values prioritising technological advancement over human life (Gleiberman 2023, p. 14).

Meanwhile, Dewi was passionate about his start-up, with enthusiastic energy radiating from his face and voice, even over Skype. Having worked at the intersection of healthcare and AI from the start of his PhD, working with a well-respected hospital as part of its data science division, he was well-versed in the processes of the U.K. healthcare system, and wanted to be involved in work which had a wider impact than just academia. To this end, he had been awarded funding to develop a start-up which used AI to aid in clinician decision-making, using AI to identify cancer from scans.

> "My motivation of why I'm doing this, is it gives value and that's what I'm passionate about. And I think the ultimate thing I see [] doing is to build a system, build an AI that is being used for example at [] hospital, and has improved patient outcomes, improved patient safety benchmarks, and that would be something that I would say ok we have been successful, and not a monetary value where saying, oh we sold the company for 1 billion."

In their personal guiding narratives, practitioners' might invoke values which lacked inherent meaning in isolation but were in service of a "broader strokes" aim. For example, efficiency might be referenced as illustrating the social good of a project. Dewi felt that safety represented a core facet of creating social good:

> "When we're talking about AI the whole purpose of it…one it needs to improve outcomes, now we're not only talking about outcomes, we're talking about positives we're talking about safety and the whole lot, it needs to be improved, it needs to show why we should be adopting AI otherwise there is no point in AI, why don't we have humans doing it."

Embedded within Dewi's narrative is a conceptualisation of AI as producing more efficient and objective results, disembodied from the material situation which it would be applied within. Here we again run into the risk of the empty value, which "seeks to optimise every possible human operation without knowing how to ask what is optimal, or even why optimising is good" (Vallor 2021). Dewi described how optimisation is at the heart of his work, which he envisioned as serving to free up time for the clinician to perform tasks which are more befitting to their abilities:

> "That aim of improving productivity, improving the time in which a treatment can be made, freeing up clinicians time, so that clinicians can dedicate their valuable time to something more, more impactful, whether that is doing research, whether that is treating more patients, doing surgery and so on…I see them all as subpoints of outcomes".

He extended this to suggest that his AI work was intended to improve on existing human abilities, also, to "make clinicians more effective decision makers, to kind of improve their decision-making". However, this faces similar pitfalls to Hinton's Medical Demon, introduced in Chapter 4, of a system which is more resource-intensive in practice. Extending prior analyses of the gap between plans and situated actions to the unconscious decisions prior to this, Roth argued that "scientists do not know what they are doing until they have achieved a certain degree of certainty about what their practical actions have produced" (Roth 2009, p. 317). These uncertainties potentially contribute to a "magical" view of AI, such as the myth of Enchanted Determinism (Campolo and Crawford 2020). Enchanted Determinism combines the magical language characterising AI discourse with the opaque, unaccountable nature of models deployed in social contexts, models with potential to impact the agency of affected groups, and minimise

the human decision-making and labour involved in building AI models. The enchantment moniker draws from literature on disenchantment, "means that principally there are no mysterious incalculable forces that come into play, but rather that one can, in principle, master all things by calculation" Weber (1946 p. 139, cf. Campolo and Crawford 2020 p. 5).

The ethical heuristics described in the previous section perhaps served as a way of anchoring narratives in a more immediate context. This anticipation might have a relational aspect, considering how the implementation of technologies would affect interactions and existing relationships. Indeed, a familiar way of engaging in affective perspective-taking is by reading fiction, in which one is placed in the experiences of another, with research suggesting that reading fiction can temporarily enhance affective perspective-taking skills (Kidd and Castano 2013). Alec described how he imagined scenarios to try and understand what these impacts of his decisions might look like:

> "If you imagined this system in use in the world, the thing that makes it useful and usable is the fact that it does has an influence on the social interactions with the people around them. And that is having an impact on the family, their friends, random passers-by in the street, everyone else. And so, what is that, how is the, what is the impact of the technology on the, what is the word for…umm collateral damage, impact on the people around the thing. And you can actually think of that on multiple levels."

Perhaps, then, the sort of heuristics discussed here represent a method of anchoring narratives in practice, providing practice with a direction and narrative with a grounding. Yet, as many excerpts from this chapter demonstrate, these narratives are inevitably grounded in the beliefs and ideological positions that practitioners have regarding themselves, their contemporaries, and the implications of their work, often reflecting other conflicts and uncertainties. In choosing to engage with AI development, values become a medium through which issues are exposed, addressed, or justified.

Building on previous work which examines the values explicated in finished outputs or dictated in AI policy and ethics frameworks (Birhane *et al* 2022), this chapter considered which values might intrinsically motivate practitioners to engage in

processes of meaning-making. It explored what sorts of motivations the interviewed practitioners ascribed as drawing them to their work, in addition to investigating which motivations informed how they engaged with this work and derived meaning from their professional activities. Thus, rather than separating them out, I delved into specific examples of AI practitioners aims, motivations and concerns, attempting to unpick the types of values which can be interpreted from these conversations.

Given awareness of individual limitations, practitioners tried to broaden empathic morality to provide a more inclusive, critical appraisal, by engaging their teams in open-ended group discussions. Despite being well-intentioned, this broader form of empathic morality could also result in the same harms as that of the individual, in various degrees and modes of expression (Costanza-Chock 2018). It still had an underlying assumption that teams of practitioners already include individuals with the breadth of experiences and honed critical thinking skills which are necessary to inform truly ethical design. Furthermore, the proposed benefits of being in a team potentially contributes to issues related to responsibility for ethical decision-making and effecting systematic change discussed in the previous chapter. Considering these issues, I propose a shift away from creating modalities aimed at achieving concrete ethical solutions towards a tacit recognition of and engagement with the ambiguities inherent in AI practices. With this, in the next chapter, I turn a focus towards a view to ethics which centres on the complex, messy and relational role of AI practitioners, and discuss how this inform ethics-on-the-ground.

CHAPTER 7: Negotiating Ethics

> "Science condemns itself to failure when, yielding to the infatuation of the serious, it aspires to attain being, to contain it, and to possess it; but it finds its truth if it considers itself as a free engagement of thought in the given." *Simone de Beauvoir*

Although ambiguity presents challenges for ethical AI practice, it can equally be positioned as representing opportunities for ethical reflection, perhaps even seen as an essential component of ethical practice. I propose that Simone de Beauvoir's The Ethics of Ambiguity (TEA) can provide a lens through which to understand and navigate the moral implications of ambiguity. Drawing upon this, I suggest that rather than attempting to artificially solve or define ambiguities, these should be embraced as an opportunity for examining assumptions, identifying impacts, and introducing reflexive practices. Throughout the previous chapters I have demonstrated how contexts and values of AI practice are fundamentally intertwined, with twin threads of responsibility and reflexivity weaving between motivations and materialities. Building upon these observations, in this chapter I elaborate on the ambiguities and nuances of practice and motivations and put forward ways in which to conceptualise and engage with the contingencies of these.

Unpredictability and uncertainty pervade development of AI models. Compounded by the nature of the AI material practice, arguably qualitatively distinct from sister domains such as software engineering, these characteristics are reflective of contingencies inherent to the "challenges of social complexity or the unpredictability of the future" (Best 2012, p. 88). In engaging with ambiguity, we are dealing with an "inescapability of interpretation" (Best 2012, p. 88), requiring consideration of a plurality of perspectives in be able to truly facilitate responsibility and accountability (Renn et al 2011). An account of ethics and responsibility must acknowledge the role ambiguity plays in practice, instead of merely try to segment it, or these grey areas end up simply excluded. Having traversed the many and complex facets making up these conceptions and contexts of responsibility, it might be tempting to ask: "well then, is there any place

for agency or hope in a context where harm seems inevitable and impacts so murky?" To respond to this very reasonable concern, I engage with Simone de Beauvoir's The Ethics of Ambiguity to examine how responsibility plays out in AI practice, weaving in discourse from the domains of philosophy, feminist studies, and politics.

The ethical implications of ambiguity have often been approached from a fairly descriptive perspective; this includes investigating misuse of strategic ambiguity in business (Paul and Strbiak 1997; Sim and Fernando 2010), and how tolerance of ambiguity amongst different demographics is linked to moral decision-making (Weisbrod 2009; Moardi et al 2016). More recently, scholars have explored the relevance of relational care to navigating ambiguity in care settings (Sabie and Parik 2019; Bregnbaek 2021).

Simone de Beauvoir's ethics provides a normative basis to build upon, furthermore, the existentialist and phenomenological basis upon which she predicated her arguments is evocative of the practitioner experiences illustrated in the previous chapters. This ethics is built upon core concepts of ambiguity, reciprocity, relationality, and the interconnection of freedom with ethics (Oganowski 2013). Whilst indicative of the relationality of practice, the diffusive effect of heuristics such as Empathetic and Sympathetic Deliberation illustrate a need to move beyond limited (and individualistic) methods of ethical reflection and anticipation. Here, I suggest that thinking-with de Beauvoir's concepts can help identify, even centre, the fundamental ambiguities of AI practice, without framing these as a problem in need of a solution.

In Chapter 5, I investigated a plurality of practitioner attitudes towards responsibility. This chapter returns to these, to consider them in the light of de Beauvoir's archetypes illustrating ways of engaging with moral responsibility. These archetypes can provide a practical way to consider the impetus for and limitations of responsibility, whilst respecting the multiplicity of values within a field which spans continents, roles, and domains. Importantly, they help surface an underlying value driving instrumental principles such as bias, privacy or transparency; all are concerned with maintaining and protecting freedoms to act. However, without explicitly recognising this, technical fixes risk short-sightedness and attempts for inclusivity fall short.

Finally, employing TEA recognises the fundamental inability of ethics to predict the future, and in fact the nature of all decisions as impacting some groups positively and some negatively. That is, it centres the notion of reflexivity, the concerns, limitations, and experiences of fundamentally interconnected stakeholders. Thus, it provides a useful tool for escaping linear, overly future-focused, solutionist thinking around ethics. This approach has been discussed in the context of medical education, where educators face an uphill task in training medical students to cope with the ambiguity and uncertainty which form an inevitable part of medical practice. Domen notes how even attempts to dispel ambiguity in medical research, "to clear up some ambiguous areas", cannot escape this, as "even more uncertainty seems to be the end result" (Domen 2016, p. 2). Low tolerance for ambiguity not only results in frustration and anxiety for the medical practitioner, but also impacts their "attitudes towards underserved or marginalised groups" (Domen 2016, p.2). Noting a conspicuous gap in materials addressing ambiguity in the medical literature, Domen suggests including it as an explicit item of consideration rather than something to avoid, suggesting the ethics of ambiguity as a way of fostering tolerance to ambiguity and respect for patients. Scholars have also proposed employing TEA as a conceptual tool to facilitate inclusivity of marginalised identities in developing political strategy, proposing that the ambiguity arising in the gaps between situated experiences act as a site for collaboration rather than strife (Nicholas 2021).

The Ethics of Ambiguity

AI practice is multifaceted, ambiguous, uncertain and involves numerous actors in varying degrees of distribution and obfuscation. Negotiating ethics involves navigating material constraints, interwoven with impacts of the political economy of the sector, whilst following one's own motivations and engaging with the external values imposed in numerous ways. There are different tools for doing this, from ethical heuristics to training and compliance mechanisms. However, attempt to parcel off aspects these off into tractable problems risks flattening, solutionism and diffusion of ethical impetus, all while legitimizing the outcomes of any set processes. The main sources of ambiguity and

certainty came from the elements of practice itself, cloaked by the nature of AI as a suitcase phrase, and extending to the blurring of lines between AI outputs and the humans intra-acting with them. This sociotechnical constellation includes humans at stages of the process, beyond just users or affected groups. The additional complexity introduced by these factors complicates ethical deliberation but at the same time can serve as a jumping-off point for reflexivity.

"Reality poses itself before you get to the ethics".

Indeed, rigid approaches to ethics seem mismatched with AI practice, framed as detrimental and almost alien to the messy, iterative processes of AI design and development. The exploratory process, the ambiguity, is an essential feature of AI which stands at odds with attempts to make it linear and the implication is that attempts at linear ethics also stand at odds with the nature of AI practice. Proceeding from this view, certain types of ethics itself act as a constraint to the ethical design and development of AI. As Martin said in Chapter 4, sometimes hiding uncertainty is necessary for the survival of the practitioner, *"reality poses itself before you can go to the ethics"*. Resources were a key part and cause of this, with access to resources intertwined with the power structures defining the field. A practical example of this is how uncertain/limited access to resources can redirect focus to efficiency and optimisation. A practitioner working at a larger organisation has the potential to access resources which minimise need for workarounds necessitated by resource limitations. Resources hold less sway over the mediation of values in such organisations, although even here the impacts of limited resources cannot be completely removed, due to inherent material limitations.

Although AI practitioners occupy a position of power with regards to their knowledge and place in its construction, a perceived lack of ethical knowledge can mediate reduced agency in ethical deliberation and decision-making. This results in the sense that practitioners' ethical reflection has little worth compared to that of an ethicist or regulator, in effect disempowering the specialist. Indeed, ethics-related terms also carry ambiguity, and even serve as a means of gate-keeping much in the same way as the ambiguity of 'AI' can be used to exclude others from certain knowledge. I would argue that ethical concepts also can risk adoption as "Serious Objects", even (see

following paragraph). Luke described how he felt like although he had a strong interest in engaging with ethical practice, often he was excluded from the conversation due to "people bringing out terms that I'm not familiar with, although in every field people like to use words they don't need to use". Similarly, cultural differences in how practitioners approached metrics, construed by George as just malicious attempts to game said metrics, might represent a genuine ambiguity in meaning; a different interpretation based on cultural norms.

Situated Narratives of Technology

At the broadest level of AI narratives, we see concerns about a singular vision of 'AI' as an existential risk to humanity (with 'humanity' employed as an amorphous concept). De Beauvoir critiqued this tendency of certain practitioners in science and technology (translated as "technics"), to cast their field/outputs as Serious Objects, as disembodied, objective truths. Rather, de Beauvoir positioned technics as embodying a potential for greater freedom in human invention, even if often imbued with goals/motivations/values framed as "absolutes", which lack meaning upon a deeper analysis (e.g., saving time/work).

Rather than taking these face-value offerings of technology, de Beauvoir instead claims that "Man is a being of the distances, a movement toward the future, a project" (de Beauvoir 1947, p.108). Technology fails if it just attempts to make up for empty values which aim to address a perceived lack, but escapes failure if it aims for situated discovery, continuously reflecting on, and communicating its weaknesses and failures. Fictions of AI can serve a purpose in this regard. To quote Wheelright (1962), "The metaphor and myth are necessary expressions of the human psyche's most central energy-tension; without it … mankind would succumb to the fate that the Forgotten Enemy holds ever in store for us, falling from the ambiguous grace of being human into the uni-signative security of the reacting mechanism" (Wheelwright 1962, p.134). That is, fictions of AI are necessary to scaffold and direct our curiosity and trajectory, to keep from only reacting to facticities. However, in doing so, we must account for our relationality and power in designing the narratives which will impact the freedoms of many.

Towards Moral Freedom

In decision-making, we set ourselves up as agents, devising goals which reveal our distance from the ideals these goals represent whilst simultaneously creating meaning in an apathetic world. In the context of the practice of AI this translates to goals in designing and developing models/systems. To de Beauvoir (1962), the qualities of vitality, sensitivity and intelligence are a result of this engagement with the world, of disclosing being. These are tempered by the nature of our embodied existence; however, materiality does not define these qualities but rather is our source of relationship with the outside world. More important is how we respond to the world given our capabilities.

Given this fundamental state of ambiguity, de Beauvoir presents an ethics which grapples with the possibility of failure rather than focusing on achieving whatever measure may indicate success. This is a phenomenological ethics which "originates in the individual consciousness and is made concrete through engagement with the world" (O'flynn 2009, p. 76). It necessarily requires recognition of the intertwined ambiguity and relationality of practice, where "genuine recognition – or moral freedom – is marked by uncertainty, by the possibility of failure, and by the relinquishing of individual control or mastery" (O'flynn 2009, p. 78). Freedom is the central aim which we should endeavour towards. De Beauvoir's use of freedom is not the ontological kind, as we are all "technically" free to make our own decisions, within the constraints of facticities. Rather, she means the moral kind of the freedom, where we can make moral decisions, engaging with (or avoiding) the responsibility and reflexivity inherent in decision-making. Engaging with moral freedom means acknowledging the existence and impact of ambiguity and the inescapability of harm that results from limited insight, in addition to actively protecting and seeking to improve the freedom of others. Our moral freedom, in this case, refers to recognising the ambiguity of the practices and contexts and application of values, and working to protect and further the freedom of others.

Responsibility Anchored in Relationality

The freedom of others to engage in constructive activity can be constrained by external factors. De Beauvoir argues that this constraint, being restricted from setting and

following s goals, can only be introduced by other people. Though material constraints also occur (she uses the examples of a flood or earthquake[9]), in the same way that only we can imbue our lives with meaning, also only we can deny it meaning, and as humans are interdependent, this confers a relational mechanism of removing meaning.

That is, "It is other men who open the future to me" (de Beauvoir 1962, p. 88), the future upon which freedom depends. If someone is cut off from being able to engage with moral freedom, they are cut off from the future and therefore "transformed into a thing" (de Beauvoir 1962, p. 89). In being cut off from moral freedom, those affected are dehumanised and transformed into the role of an object. We are left with two groups. One consists of those with the power and freedom to identify goals and work towards them. The other group consists of people who "are condemned to mark time hopelessly in order merely support the collectivity" (de Beauvoir 1962, p. 89), off whom "the oppressor feeds himself on their transcendence and refuses to extend it" (de Beauvoir 1962, p. 89).

The oppressing group tries to negate negative reaction or protest this state by appealing to nature, defending the situations he maintains by referring to them as natural phenomenon. In this way a struggle is set up between the exploiter and the exploited, one with concrete implications. The oppressor, in succeeding, realises his own freedom, whereas the oppressed is relegated to an object. While the terms "oppressor" and "oppressed" may seem damning in this research context, they serve to locate the impact of "objective" views of AI ethics, utilising processes which (due to inherent power imbalances) frame the intended users of AI outputs as objects. As such, recognising the capacity for this objectification – often through leveraged values or heuristics to manage ambiguity, as discussed in previous chapters - necessitates a different approach to ethical discourse, as considered in the following section.

Envisioning an Ethics of Ambiguous AI

Practitioners' decisions have the potential to impact freedoms in various ways, as indicated in prior discussions of responsibility and accountability. Previous chapters

[9] Climate change muddies this distinction.

examined the subjective nature of AI practice, shaped by the constraints of hardware and the idiosyncrasies of the positions which projects and practitioners occupy within the greater patchwork of resources and power. AI practice can cast individuals and groups as Things, in service of a larger goal; even, and particularly, if this goal is a noble one. Given this, I consider different ways in which the attitudes towards ethics discussed in TEA can manifest in AI practice and look at how seemingly straightforward solutions can lead to ongoing problems. Avoiding restrictions on freedom is impossible, as any decision will have an impact on some groups. By making the choice to focus on one domain or problem, we are implicitly excluding others, as TEA points out. Thus, AI can curtail freedoms in various ways, whether intentionally or not. Additionally, there is a middle ground - AI can easily be ambiguous given the intent of practitioners to develop models which have a broad applicability.

Negative Spaces

Over the course of my doctoral research, in work both directly thesis-related and emerging from engagement with the other projects (as described in Chapter 2), I frequently noticed how the negative space created by decision-making is rarely examined, especially the moral implications of these. Ross reflected on the potential for negative change, on how "there's a load of potential doing damage [with AI]". He was concerned about conscious applications of AI in ways which would reduce freedoms, especially given how invisible the workings of many systems can be:

> "…Take away all their opportunities on the basis of the Social Credit system they happen to randomly fall foul of. You know I think that's the scariest idea I have encountered in terms of the use of technology since…well since the atom bomb."

Though such overt issues are easily considered, there is often less consideration of covert harms, particularly in how AI reinforces the status-quo of existing societal inequities. Recently, an increasing literature has grown mapping how typically less visible aspects comprising AI models can contribute to unconscious limiting of freedom. Datasets and the iterative processes of work involved in gathering, cleaning, preparing these for modelling, are an important part of this. De Vries *et al* (2019) demonstrated that

computer vision models trained on high-income households were much less accurate in identifying objects in low-income households (de Vries *et al* 2019). Popular approaches to Natural Language Processing such as modelling using word embeddings, where practitioners model the relationship between words using vectors to model how closely they share a semantic association, are of course heavily influenced by the socio-historical structures shaping the context in which the words in the training datasets are used (Garg *et al* 2018). From the prior discussions of data and practice, we can see that practitioners were aware of their own involvement in data modelling, attempting to be reflexive about the impacts of their decisions, thanks in large part to the discourse around data bias. The practitioners I spoke to also viewed many ethical concerns in AI as due to a lack of reflexivity, a lack of practitioner reflection on their decisions and how they impact others. However, they were typically less conscious of the influence of their own subjectivity on more mundane practices such as data work, even given the downstream ethical implications. This was partly due to the implicit understanding of AI as more objective, as the Imperfect Demons discussed in Chapter 4, a perspective which led to a favourable interpretation of AI outputs.

Underlying Implications

In this way AI can create distance between the self and the other, reducing reflexivity, as discussed in the previous section. In the words of de Beauvoir, those placed in power are inadvertently framed "as a transcendence" or as having superior knowledge/capabilities, thus considering "others as pure immanences", and in doing so assume "the right to treat them like cattle" (de Beauvoir 1962, p. 110). By converting people into things, we reduce or discard their worth and we make it far easier to disregard details that might be important. This is seen in common conceptualisations of data, in the parameters extracted from data-points representing complex human and social systems, which are central in the process of AI design and development. Transformed into vectors, data forms the "moving substrate of Machine Learning" (Mackenzie 2017, p. 72). The vectorization of goals, as observed in studies of the life sciences, "modifies the objects visibility" (Lynch 1988, p. 229) in the same vein of Dreyfus' argument about how naming conventions mediate perception. Stark and Crawford (2019) note a similar dynamic in

the "defamiliarization" which "data artists" employ to convey the aesthetics of data-driven surveillance to their audiences, yet in this context this is employed as an intentional device to prompt reflection. This translation work from goal to output involves qualitative judgements about how to measure a phenomenon. Barocas and Selbst (2016), for example, illustrated the impacts that seemingly arbitrary decisions made in choosing which variables to include, which they argue could happen "in such a way that happens to systematically disadvantage protected classes" (Barocas and Selbst 2016, p. 678).

As mentioned above, datasets are intrinsically impacted by the conditions of their collection, which includes the socio-historical facets. AI projects are fundamentally situated within this wider network of dependencies, to borrow from de Beauvoir "no project can be defined except by its interferences with other projects" (p. 76). However, there are two problems which face any attempts to develop more ethical datasets, models etc. One is the issue of one solution resulting in further unfairness, which I will discuss first. The other is the limit of responsibility of the practitioner, in their context of practice, which I will then discuss.

Re-examining Fairness and Bias

Discussions of fairness and bias have been threaded throughout this thesis and considered from various angles. Over the past few years, the field of Fairness, Accountability, Transparency and Ethics (FATE) has emerged, trying to develop technical fixes to these concerns. However, measures to tackle such issues can result in greater problems down the road. Strauß (2021) highlighted how this recent focus on fairness, transparency and accountability measures in AI can contribute to user interaction biases by reducing reflexivity due to greater faith in the systems output, when in fact the complexity of automation of deep learning can result in these measures even contributing to output which should be questioned. Tackling fairness is a crucial concern, however envisioning this as a narrow, discrete problem, abstracted from the power dynamics which generate bias in the first place, risks false illusions of a quick fix (Hampton 2021). Here, the attempts to preserve or increase freedom come face to face with the issue that in addressing one problem we inevitably create another, seeing here what de Beauvoir's means in her assertion of the inescapable ambiguity of the ethical process.

Drawing upon Rittel and Webber (1963), Strauß framed this creation of new issues as a Wicked Problem, a "class of social system problems which are ill-formulated, where the information is confusing, where there are many clients and decision-makers with conflicting values, and where the ramifications in the whole system are thoroughly confusing" (Buchanan 1992, p. 15).

However, a Wicked Problem is framed as a highly specified situation; Rittel and Webber spell out ten characteristics of a Wicked Problem which include "non-definitive formulation" (Rittel and Webber 1973, p. 164) and "there is no tame solution" (Rittel and Webber 1973, p. 163) to such a problem. The framing of TEA is broader than this; there are no tame and wicked problems, rather, all decisions involve ambiguity thus all must be engaged with regarding this. That is, rather than a framing of a problem, it suggests a reframing of a perspective. Taking a closer look again, we can see the superficial similarities between Wicked Problems and de Beauvoir's conceptualisation of ambiguous existence; that there can be no good solution, because of the ambiguity of the situation, and every solution will have ramifications down the line due to this.

> "The formulation of a wicked problem is the problem! The process of formulating the problem and of conceiving a solution (or re-solution) are identical, since every specification of the problem is a specification of the direction in which a treatment is considered." (Rittel and Webber 1973, p. 161)

The second issue is the ambiguity of the degree of responsibility the practitioner has in addressing the socio-historic origins of issues such as bias. There are two issues at play here; the ability of the practitioner to change the flaws of the underlying system, and their agency to act in the given situation. Of course, it is absurd to suggest that practitioners can make changes to the contexts of power structures in which they are subordinates; as discussed in the previous two chapters, the answer to this must be cultural change from the top down. Indeed, de Beauvoir frames this as a political rather than moral question, arguing that "we must end by abolishing all suppressions" (de Beauvoir 1947, p. 95). Whilst working towards this higher-level goal, we have an ongoing moral responsibility to consider the potential implications of our decisions upon the moral freedom of others.

Power, Agency, and the Nature of Responsibility

Making an ethical choice which moves towards the freedom of one group may lessen the freedom of others, hence the fundamentally "painful" nature of ethics, as was introduced at the start of this chapter. De Beauvoir illustrates this using the example of political revolutions. When collectives work together to overthrow an oppressor in order to gain their freedom, they necessarily relegate the oppressor to a "Thing", with relegating others to a Thing having a negative impact on the self.

There are a few ways in which a practitioner can engage with ethics in their practice. One is recognition of the numerous contextual factors impacting decision-making, recognising the ways in which the Object of focus might be unconsciously placed above the freedom of those who contribute to its creation; ensuring that they are "not blinded by the goal…to the point of falling into the fanaticism of seriousness or passion" (de Beauvoir 1947 p. 96). This requires awareness of the tensions which are formed in the space between lived experience and the perception and actions of oneself by the Other, tensions which risk rote casting the Other as a homogenous group, and erasing their own complex lived experience (Parker 2015). Embodiment inherently involves being constantly in motion, engendering difficulty in making a representative record of lived experience. De Beauvoir asserts that this tension, although uncomfortable, allows for richness of experience, rather than representing an irreconcilable state of being.

It also requires recognising that values are not universal; they change based on group and context, and as seen in the discussions in this thesis, they are impacted by context; *"these kinds of problems are dynamic and changing"*, in the words of Stefan. De Beauvoir recognises this multiplicity of values and contexts, warning against assuming we know best how to handle a situation in which we do not have lived experience by dictating values; "There is nothing more arbitrary than intervening as a stranger in a destiny which is not ours" (de Beauvoir 1947, p. 92). To de Beauvoir, in the same way in which it does not make sense to try to attain our own ontological freedom because we already have it, we cannot exactly attempt to procure this freedom for others, as they also have theirs. Taking this at face value results in an attitude of distance, which views no one

solution as better than another, where present occurrences have the same status as past events, as "impartially contingent facts" (de Beauvoir 1947, p. 81), with choice, then, being an illusion. She calls this the 'aesthetic' attitude, an attitude of withdrawal and discouragement rather than a truly moral view. Contrary to this view, the present involves choices - making no decision is imbued with moral implications to the same degree as making an active decision. De Beauvoir instead focuses on protecting the freedom of others to act according to their values; to her, this is the core of ethics, and the freedom of the Self and the Other are intertwined. That is, "To be free is not to have the power to do anything you like; it is to be able to surpass the given toward an open future; the existence of others as a freedom defines my situation and is even the condition of my own freedom" (de Beauvoir 1947, p. 97).

In an analysis of TEA, Parker (2015) draws attention to de Beauvoir's use of singularity in the text, arguing that she uses it to reference "the inherent multiplicity of existence" (Parker 2015, p. 2), as "resistance to conceptuality and categorization" (Parker 2015, p. 2). At the same time, relationality is also core to the philosophy, with de Beauvoir recognising that an individual's capacity to make independent decisions is conditioned on their material situation in addition to will. This interpretation means that "To intervene on behalf of the other can be just as problematic as believing that I have no obligation to do so" (Parker 2015, p. 6). The inherent distance between the self and the Other gives a necessary pause for reflexivity, rather than representing a chasm to be breached:

> "Thus, we see that no existence can be validly fulfilled if it is limited to itself. It appeals to the existence of others. The idea of such a dependence is frightening, and the separation and multiplicity of existents raises highly disturbing problems" (de Beauvoir, p. 85).

Parker (2015) identifies a tension in asserting ambiguity as a value; "what to do in each moment is and ought to be affirmed as a matter of indefinite questioning; my next step is not inevitable and as a lived agency I live this noninevitability" (Parker 2015, p. 11).

Seeking clear-cut ethical tools as an end in themselves misses the core of both AI, ethics, and human relations, that "clarity (and conversely, ambiguity) is not an attribute of messages; it is a relational variable" (Eisenberg 1984, p. 5), and recognising this is crucial for understanding responsibility and thus ethical actions in practice. Setting up formal ethics frameworks creates an object which people then orient themselves around, obfuscating that it is the process of creating knowledge or artefacts which will limit the "open" future (and therefore freedom) of others. Formal ethics is attempting to construct an object upon which to project values, and then work backwards to prevent this object from existing by trying to propagate the values which were set up by the envisioner, and in doing so becomes viewed as its own form of objective "end". Doing this assumes the values of others, distancing the ethicist from the process and outcomes as if the project of interest concerns just others, or not even others, but an abstract notion of values. It side-lines the concrete impacts that AI projects have by virtue of existing, subordinating the unethical treatment of gig workers annotating datasets (Gray and Suri 2019) to grand abstract causes, such as values of Privacy or Transparency. This results in a kind of paternalism which ignores the fundamental relationality of both morality and AI practice.

Modes of Deflection and Reflection

I have presented de Beauvoir's notions of what it means to be ethical - to embrace moral freedom by acknowledging the ambiguities of our contexts, ensuring the freedoms of others. Here, I employ her sketches of archetypes depicting how we might deflect our responsibility to engage in ethical reflexivity, to aid reflection on the discussions of practice and value covered in the previous chapters. De Beauvoir wrote of five types of reactions to responsibility in the midst of existential ambiguity – the Sub man, Serious man, the Nihilist, the Adventurer, and the Passionate man.

Apathy and Passivity

The Sub man is the least desirable of these archetypes, characterised by apathy. This person seeks to avoid risk, to escape from positive or proactive choices instead being passive. 'He is as afraid of engaging himself in a project as he is of being disengaged…he

is thereby led to take refuge in the ready-made values of the serious world" (de Beauvoir 1962, p. 47)[10].

This archetype of passivity is most salient in instances where practitioners don't question how their work might impact on others, assuming that it is not relevant, or that failure is inevitable: "ethics is the triumph of freedom over facticity and the sub-man feels only the facticity of his existence" (de Beauvoir 1962, p. 48). Such a person sees only the rigidity of the occurrences which shape their present existence. Here, we return to the notion of the unavoidable ambiguity of impacts. Back in Chapter 5, Lukas spoke of attending a workshop on biases in datasets, speaking of how practitioners often did not question the circumstances or relevance of the data they modelled: *"people don't really do it because they don't reflect upon the fact that they're building something that has an impact upon the real world and real people, and that the models might be unfair"*. The consequences of this lack of reflexivity can have serious ramifications. Eddie reflected on how systemic unfairness of existing systems might be inducted into AI systems if practitioners weren't careful, perpetuating discrimination, which by implication would restrict the very literal freedoms of the individuals unfairly detained, to use the example of predictive policing. Furthermore, he considered how the ambiguity of AI practice itself may lend a way of enabling the practitioner to deflect their moral freedom by blaming the machine.

Seriousness and Abstraction

Another type of deflection is seen in the depiction of the Serious man, who "loses himself in the object in order to annihilate his subjectivity" (de Beauvoir 1962, p. 49). Such a person subsumes their own freedom in a larger cause, in the belief that this will mean he has the same value which is accorded to the cause. Examples of this are men who distance themselves from their own subjectivity using a "shield of rights" (de Beauvoir 1962, p. 52). bestowed from various relations whether these be religious organisation, political party, or boss. Such people are motivated at their core by notions of whether something is "useful", viewing this assessment of use as objective, just carrying out the

[10] This is conditioned on the agency an individual must proactively make decisions.

logical response to the fact of usefulness. In denying their own subjectivity, this person ends up belittling and devaluing the freedom and subjectivity of others, which can lead to harmful outcomes. Deflecting from the inconsistencies of their own beliefs, they point out the flaws of the 'useful' objects of others.

De Beauvoir describes how outside this specialism or belief, the person behaves in the same manner as the Sub-man in confirming to unquestioned values or completely disengaging with active decision-making, viewing change and uncertainty as a threat. In the context of the wider world, the goals achieved seem pointless, but the Serious man is subsumed within this Object of focus – an object often defined by the unacknowledged subjectivity. This object is seen as useful in its own right; "dishonestly ignoring the subjectivity of his choice, he pretends that the unconditioned value of the object is being asserted through him; and by that same token he also ignores the value and subjectivity of others" (de Beauvoir 1962, p. 52). The impact of the behaviour of the Serious man also applies to his relationship with his own expertise. The Serious man is not intentional in these exclusionary behaviours, rather these arise as the result of a lack of allowance for plurality of experiences, as a rejection of ambiguity. To use an example from the practitioners I spoke with, Ariel viewed the Object as superseding need for ethical reflection, seeing her work as important therefore discarding the need to consider other perspectives.

> "I've always thought that you know the progress makes it all worth it, because if you believe in your research you believe it's going to be for the best, so you kind of get over that idea that whoever whatever animal is suffering because you know you're doing this because you really believe the research is worth it and going to change something…so I haven't really struggled with ethics issues"

Ethics can equally fall foul of this type of thinking. Introducing strict top-down ethical approaches, whether a framework or the inclusion of an ethicist as the source of moral reflection in a team, can result in stripping away of ethical knowledge, furthermore, they may ameliorate practitioners' sense of responsibility to act. This sentiment was seen in the findings of Chapter 4, where the practitioners who had worked directly with an

ethicist in their team deferred to that ethicist for moral decision-making, and ascribed to them a knowledge which was greater than is possible. Even an ethics expert can't anticipate outcomes successfully (especially when they were less acquainted with the actual AI processes), however it is assumed that they could. Stefan told me *"I don't feel like I could make useful predictions about how our technology could be used...an ethicist or someone…people who are speculating about the long-term consequence, probably they could sit down and think systematically about it"*. Here, we see two problems, one is the diminishing of agency of the practitioner/method of abstracting responsibility, and the other is setting up abstract principles as Objects disembodied from the actual contexts and desires of the people impacted by AI outputs.

Nihilism and Avoidance

Next is the Nihilist, who having experienced disappointment after actively engaging with the world in the manner of the serious man, now shares traits with the Sub-man. This person recognises their own freedom and subjectivity but views it as a negative, as a source of discomfort to be avoided. Here "the negation of aesthetic, spiritual and moral values has become an ethics; unruliness has become a rule" (de Beauvoir 1962, p. 59). In reacting in disruption to the serious, the nihilist creates their own object of value, but also, in rejecting their own existence and the value of it, they also reject other peoples. De Beauvoir sees the nihilist construal of ambiguity as mistaking the "positive existence of a lack …as a lack at the heart of existence" (de Beauvoir 1962, p. 62). Stopping at viewing the world as lacking justification but not proceeding to the realisation that it is the individual themselves who can justify existence, and in wholesale rejecting the values set up by others, misses freedom as the ultimate aim, existing beyond them. This might translate as viewing power imbalances in the AI sector as so stark that engaging in ethics is pointless. In Chapter 5 we saw how Stefan felt like any chance of real change was not possible, and thus no decisions he made would really have an impact, which discouraged him from engaging.

> "I don't know that that choice makes any difference, I feel like, there's going to

> be some long-term trend in society that either the elites will get a hold of these tools and use them to control the masses…These tools are gonna be developed anyway, the ideas behind them are gonna be well-known and the rich are going to be uniquely positioned to exploit them. And yeah, I don't think that anything I do is going to change that."

Taking such an approach serves to cut the practitioner off from the possibility of engaging with their own responsibility, from truly investigating the implications of their decision-making. In doing so, it serves as a method of deflecting this responsibility.

Self-focused and Disregarding

Another archetype is the person who enjoys life as the Adventurer, finding joy in the process of achieving an aim, even if he recognises the end as not having an intrinsic value, enjoying action for its own sake. Such a person perceives ambiguity as positive, as representing possibility, and in this way can possess the aesthetic of moral character. However, this depends on how he regards his relationship with others, on how they fit into his interests. The adventurer "remains indifferent to the… human meaning of his action, who thinks he can assert his own existence without taking into account that of others…[the] accomplice of the oppressor" (de Beauvoir 1962, p. 67). To apply this to the AI context, this is someone who sees their work as divorced from the rest of the world, therefore they feel free to follow their interests without considering their relationality or the potential of their work to impact others. We saw this in previous chapters, particularly in the account given by Lukas, who viewed his work as exempt from a need for ethical deliberation, due to his position in an academic lab doing theoretical work.

Meanwhile the Passionate man, like the Serious man, sets up an object of focus, but unlike the Serious man he acknowledges its subjectivity, and finds pride in this. The Passionate man seeks possession of his Object of focus, and because only this Object is of true importance to him, he runs the risk of devaluing others. In essence, the Passionate man acknowledges that the inherent ambiguity of their processes involves moral risks yet does little to address these risks as the Object of discovery supersedes the impacts of their decisions. In the context of AI practice, I see this as particularly a risk of being

overly motivated by curiosity. At an extreme, we see perspectives such as that of Geoffrey Hinton, who concurrently stated that AI is an existential risk to humanity, and that this risk was worth it in the pursuit of academic freedom and intellectual challenge (Vance 2021).

Relationality as the Connecting Thread

"No existence can be validly fulfilled if it is limited to itself…the separation and multiplicity of existents raises highly disturbing problems" (de Beauvoir 1962, p. 72). Technologies such as AI are often employed in service of a cause, with resultant harms disregarded in pursuit of this cause. Perhaps reframing to a focus on the present, on the ambiguities of creating technologies, might prompt a different way of reasoning about ethics practice.

> "The tasks we have set up for ourselves and which, though exceeding the limits of our lives, are ours, must find their meanings in themselves and not in a mythical Historical end. But then, if we reject the idea of a future-myth in order to retain only that of a living and finite future, one which delimits transitory forms, we have not removed the antinomy of action; the present sacrifices and failures no longer seem compensated for in any point of time" (de Beauvoir 1947, p. 59)

A key point of de Beauvoir's writing is her identification of tensions between the human will for transcendence, and the desire to be fully part of the world as both stemming from the desire for freedom. These tensions map onto the practitioners' values of problem-solving and being solution oriented, balanced with an unpredictable field/workflow which practitioners are drawn towards by a desire for intellectual stimulation/freedom. This freedom results in the constant tension of deliberately choosing a path thus creating that path, against the consignment of past actions to facticity. De Beauvoir illustrates archetypes of individuals who endeavour to avoid awareness of this tension. As Lukas described, practitioners can sometimes position AI as their Object, accorded it unconditioned value which cascades values back to the practitioners who work on it, who view other Objects as lacking value.

Countering temptations to deflect from responsibility or to cut through ambiguities with ethical Objects (such as abstracted principles) involves approaching AI

ethics as knowledge work, together with identification of the narratives which this knowledge work is being conducted to address. *"What types of meaning are we trying to make?"* "*Who do we envision to benefit from these meanings?"* Having made explicit this narrative, the next step is to then endeavour to understand the negative spaces the narrative leaves, the painful decisions which are being made. Conceptual tools such as Epistemic Injustice may come in useful here, including consideration of "being-in-the-room privilege", which calls attention to the barriers excluding people from even being present or visible in such conversations (Táíwò 2022). A key part of this is identifying whose freedoms may be curtailed by both the narrative and the mundane practices which occur in service of it. Knowledge of this can only be mapped via engagement with the communities impacted by these, to avoid acting on behalf of and therefore curtailing moral freedom. This sort of approach can be seen in justice-oriented frameworks such as Data Justice (Taylor 2017) and Design Justice (Costanza-Chock 2020), which both emphasise the importance of participation. Still, participatory approaches also require critical examination of the limitations of standpoint epistemics, of consideration of who is and isn't in the room.

Engaging with ambiguities via this critical epistemic reflection is crucial for working towards the freedom of groups affected by AI models. There is no simple answer to address the systemic issues pervading AI, from resource inequalities to data biases, however, reframing ethical perspectives can provide tools to move towards change, counterintuitively grounding ethics "in the ungroundedness of all forms of decision, all political claims, human and algorithmic" (Amoore 2020, p. 148). In recognising that "agency is not an attribute, but the ongoing reconfigurings of the world" (Barad 2003, p. 818), and scaffolding ethical deliberation regarding these reconfigurings, we also serve to build up the moral freedom of practitioners. This also helps provide a practical approach to the suggestions of Mittelstadt *et al* (2019), that AI ethics be pursued as organisational rather than professional ethics, treated as a process rather than subject to solutionism, as a "*microcosm of the political and ethical challenges faced in society*" (Mittelstadt 2019, p. 505).

Using TEA as a way of thinking through alternative approaches, I have tried to illustrate how ambiguity can provide a starting point to map contexts of practice and application, part of a continuous navigation of contingencies and values. I also employed de Beauvoir's archetypes to examine ways in which responsibility may be deflected. Finally, I considered how these concepts might inform design of approaches which explicitly recognise the fundamental uncertainties shaping AI projects, incorporating them into process-based ethical thinking rather than trying to side-line or force them into a box of possible clear-cut solutions.

CHAPTER 8: Conclusions

Engaging with ethics in Artificial Intelligence practice means navigating complex, interweaving factors from the socio-historical to the material. This thesis has investigated some of the contexts which shape AI practice, the factors which influence how practitioners engage with ethics during their work, seeking to understand practitioner approaches to ethics. I presented findings regarding the contexts which dictate and influence AI, and practical engagement with ethics, sought to understand the nature of responsibility and accountability according to practitioners, and reflected on how these findings might help in our conceptualisation of ethics in AI. In this chapter I revisit these findings, and then consider their implications.

AI Ethics-in-Practice

This thesis provides several contributions to understanding AI practice and ethics, investigating the socio-material of AI practice and AI 'ethics-on-the-ground' via a series of agential cuts (Barad 2007) examined via conceptual lenses which draw upon social epistemology and feminist philosophy. These come together to form an overarching contribution in highlighting how we, as practitioners, ethicists, and policymakers, can engage with the Ambiguous Devils we are constructing, rather than attempting to tame abstracted Imperfect Demons. That is, suggesting that rather than trying to cut through the ambiguity to shape proposed solutions with universal principles or values, we view ambiguity as the one universal which can shape our processes for better or worse. Our rudder is to prioritise the moral freedom of others, requiring us to fill in a contextual map as much as possible, mapping the needs of those who our Ambiguous Devils aim to impact, and the contexts which already shape their moral freedoms, which includes acknowledging the epistemic lacks which are inevitable. Our own moral freedom lies bound up with that of others - rather than aiming for an abstract 'good', we engage in ongoing examination of the impact of our choices in developing AI models, systems, policies etc. Given the nature of AI as data-driven knowledge construction, these choices are fundamentally intertwined with social epistemology, shaped by epistemic injustice.

Chapter 1 described the research context and aims, setting out the following guiding research questions:

1) *'What kinds of constituent factors and values shape contemporary AI practice?'*

2) *'What are the implications of these for developing a practical ethics of AI?'*

These questions shaped the direction of the research processes which are outlined in **Chapter 2**, where I set out the theory and methodology informing my research approach. To answer these research questions, I used qualitative methods, in the form of semi-structured interviews and ethnography, to learn about the perspectives, experiences and the values of AI practitioners in several roles and domains. A foray into their lifeworlds set out to map interactions between facets of experiences, to provide a rich and grounded understanding of activities in the field.

Chapter 3 mapped out a broad history of AI development and associated ethical practices, with the conceptual and socio-political influences shaping the fields. I highlighted some recent studies which investigated practices of AI, though these rarely touched on the interactions between the contexts of AI development and practitioners' values, which had implications for an ethics of AI. This outlay served as a backdrop to initiate my discussion of the findings in **Chapter 4**, where I demonstrated how practitioners' experiences indicated the processes of developing AI models as experimental, highly iterative, and difficult to predict. This was further compounded by contextual pressures such as access to resources, and cultural norms, showcasing a very complicated set of hierarchies and interruptions characteristic to the field where practitioners often worked across boundaries of applied versus theoretical domains. Practitioners employed ambiguous or ill-defined concepts of AI and machine learning, allowing for flexibility and nuance of context while simultaneously complicating perceptions of the role of ethical deliberation within these varying practices. The significant material and power imbalances in both industry and academia created obstacles to practice and shaped priorities, consequently impacting moral reflection and engagement. However, acknowledging the variations in individual positions highlighted

the differing capacities for such engagements, emphasizing the importance of practitioners' framings of responsibility for ethical deliberation.

In **Chapter 5** I examined how practitioners conceptualised the nature of their own responsibility, and how they understood the implications of the socio-technical entanglements that they created and worked within. Practitioners were keen to better understand responsibility, trying to pinpoint their own roles to play in ethical decision-making, while navigating the complex situations and sometimes competing motivations which were discussed in the previous chapters. However, this was complicated by several factors, including individual apathy, industry logics and unpredictable material practices. Our discussions of responsibility and accountability often touched on issues of accessing, processing and communicating data, and associated knowledge and information, however, such responsibilities could be framed as abstracted from practitioners. Complexity and ambiguity often obfuscated the locus of responsibility, especially given the distributed nature of AI work, including stakeholders beyond practitioners, in various permutations and arrangements. Furthermore, combinations of contexts, constraints and values mediated responses to responsibility, serving to diffuse notions of accountability.

Building upon this, in **Chapter 6** discussed how this ambiguity stretched beyond immediate conditions to the uncertain relationship between society and technology, and the implications of emerging technologies. I set out to understand the important motivators of AI practitioners, using their narratives to interrogate their values and moral positions. These values were considered to intrinsically motivate practitioners, helped explicate the meanings they made from their work, and understand how they positioned themselves as moral agents. Considering the iterative nature of their craft, parallels were drawn with ambiguities inherent in artistic processes which aided the framing of ethical responsibilities. Practitioners sought freedom of expression, facilitation of creativity, and to explore the problems which they are most interested in. While several broad level ideals such as social good or reducing harm were discussed, the ambiguities of practice and unclear impacts of outputs led to the adoption of empathetic or sympathetic ethical heuristics – cognitive tools which were useful for eliciting some recognition of practical

impacts, but nevertheless limited by reliance on practitioners' subjectivities, and reduced their scope for considering relationality by abstracting users of AI to conceptual objects.

In **Chapter 7** I employed T*he Ethics of Ambiguity* to examine my findings, suggesting that seeking clear-cut ethical tools as an end in themselves misses the core of both AI, ethics, and human relations. Recognising this is crucial for understanding responsibility and thus ethical actions in practice. I suggest a better framing for ethics of AI is a process-based, more relational model where ambiguity and uncertainty is expressly recognised, a version of ethical responsibility which can allow for a multiplicity of values by instead focusing upon protecting and extending freedom to act (which includes having access to knowledge, thus challenging epistemic injustices).

Embracing Uncertainty

"This thesis describes an empirically grounded investigation of the elements which co-shape AI practice. It contributes novel insights into the contexts and motivations through which AI ethics is negotiated and navigated, and their implications for the ethics of AI practice itself."

These are the words I felt compelled to write in discussing my contributions. Proposing a hypothesis, conducting a robust study, evaluating my outputs – these are comfortable modes of writing, of thinking. Stepping outside of this requires courage, acceptance of the limitations of my own perspective which cannot be hidden behind data-points, no matter how many templates, diagrams, and triangulations I construct.

Engaging with *The Ethics of Ambiguity* helped me map the limitations of my solutions-oriented mindset and begin side-stepping it, making tangible the shifting, permeable yet fundamental contexts of the work set out in this thesis. The fictions we create for AI serve as scaffolds for perceiving and understanding this complexity and fluidity, just as seeing tangible grains of rice helps concretise our understanding of the enormity of a billion dollars. That, certainly, is a large part of my own take-away from this research, an answer to the question *What are the implications of these (findings) for developing a practical ethics of AI?'*

Critiques of AI, and AI Ethics, have formed an emergent conversation which often parallels the findings of my thesis. However, the situation is much more complex than is often conveyed in the literature. Tackling barriers to ethical engagement within AI requires addressing numerous constraints whilst recognising the limitations of certain framings of ethics. An example of this is magical discourse around AI. While this does deflect from accountability and reflect the uncertain nature of AI practices, ambiguity is not unique to Deep Learning, for example, rather, Deep Learning just represents an edge case. Rather than providing a barrier to overcome, perhaps it reflects a broader nature of the field which should be engaged with.

By taking an approach which explicitly accounted for and even sought emergent themes, this study went beyond the narratives which often constrain AI ethics discourse, part of which is an attitude to AI which perceives its models as a series of ideas with potential impacts and harms, rather than a material practice which practitioners engage physically with. This is built upon an implicit, even unconscious, acceptance of ethics as approached from a deontological or consequentialist foundation. Perceived from this angle, the primary area of importance in ethics is to make sure the 'right' ideas are developed and put out into the material world, that such ideas can be guided to achieve a status of "ethical" as long as the correct principles are adhered to.

The fields of AI Ethics and Fairness, Accountability and Transparency have begun to tackle this deficit in scope, recognising that certain practices result in undesirable social impacts if not challenged and altered (Young *et al* 2022). This still can be subject to the same issue of not addressing the immediate impact of material constraints such as hardware or process. On the other side, casting practitioners as lacking agency, that is, as cogs in a larger socio-economic machine, risks missing the opportunities that recognising the impact of embodied interaction with design and development could introduce. Likewise, formal frameworks can be counterproductive by diffusing responsibility and not accounting for the value tensions which individual practitioners must navigate (for example, resource scarcity leading to increased focus on efficiency in model and dataset design).

This results in a gap between the technological solutions proposed to narrowly-constructed problems (e.g., technical fairness), the high-level principles intended to guide design and development of AI models (e.g., privacy), and the impact which socio-historical, geo-political analyses of power have upon AI ecosystems, exemplified by perpetuated epistemic injustices.

The Ethics of Ambiguous Socio-technical Assemblages

The research presented in this thesis indicates that this gap is highly shaped by the nature of AI practice as a patchwork of contingencies. An adequate ethics of AI must therefore account for the ongoing ambiguities formed by the constraints and boundaries of designing and developing models, and the role of epistemic and hermeneutic injustice in infrastructuring these. This thesis demonstrates that employing conceptual tools from *The Ethics of Ambiguity* facilitates a framing of AI ethics which is well-situated to address these fundamental ambiguities. By doing so, it proposes a process-based approach to ethical thinking which avoids the pitfalls of trying to overcome or force ambiguity into a box. Furthermore, it facilitates critical reflection on, and new imaginings of, the role and development of AI.

I hope to see this work built upon and extended in several ways in future research. One important direction is mapping a wider and more heterogeneous pool of stakeholders working with AI, to understand their material practices and motivations, including conducting ethnographies of their work. Additionally, it would be beneficial to conduct a close analysis of the intricacies of practice itself, taking a close look at the phenomenology of data work, modelling and so on in their own right. Another important line of enquiry building upon this work could be empirical work such as case studies to demonstrate usefulness and applicability of such tools in real-world contexts. This ideally should involve challenging the typical contexts in which (and for which) current practices have been designed, especially given the existing socio-political and epistemic injustices which are baked into the status quo.

The analyses presented in this thesis, culminating in a novel conceptual framework of reflexive epistemic practice which centres ambiguity while addressing the

weaknesses of reflexivity such as practitioner blind spots, provides a starting point for those looking to develop an alternate, more relational, ethics of AI practice. Engaging with ambiguities via critical epistemic reflection is crucial. Building upon a close examination of the nuanced moral agency and responsibility of AI practitioners, it facilitates critical engagement with infrastructuring processes shaping the moral freedom of groups impacted by AI models. There is no simple answer to the complex issues intrinsic to AI, from resource inequities to data biases; however, the account given in this thesis contributes to a reframing of ethical perspectives, drawing together empirical insights and conceptual tools to move towards social and technological change. This work needs to go beyond ethical band-aids towards accounting for AI socio-materialities and developing new narratives, in doing so moving towards equitable AI futures.

References

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M. and Robinson, D.G., 2020, January. Roles for computing in social change. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 252-260).

Ackerman, M.S. and Cranor, L., 1999, May. Privacy critics: UI components to safeguard users' privacy. In *CHI'99 Extended Abstracts on Human Factors in Computing Systems* (pp. 258-259).

Agar, J.O.N., 2020. What is science for? The Lighthill report on artificial intelligence reinterpreted. *The British Journal for the History of Science*, *53*(3), pp.289-310.

Agarwal, Y. 2019. Laplace's Demon: A Unique Perspective of Machine Learning Models. Available at https://www.linkedin.com/pulse/laplaces-demon-unique-perspective-machine-learning-models-agrawal (Accessed 02.05.2020)

Alsheikh, T., Rode, J.A. and Lindley, S.E., 2011, March. (Whose) Value-Sensitive design: a study of long-distance relationships in an Arabic cultural context. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 75-84).

Altenried, M., 2020. The platform as factory: Crowdwork and the hidden labour behind artificial intelligence. Capital & Class, 44(2), pp.145-158.

Amatriain, X., 2013. Mining large streams of user data for personalized recommendations. ACM SIGKDD Explorations Newsletter, 14(2), pp.37-48.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B. and Zimmermann, T., 2019, May. Software engineering for machine learning: A case study. In 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP) (pp. 291-300). IEEE.

Amoore, L., 2020. *Cloud ethics: Algorithms and the attributes of ourselves and others*. Duke University Press.

Arblaster, A., 1982. Human factors in the design and use of computing languages. International Journal of Man-Machine Studies.

Armbrust, M., Ghodsi, A., Xin, R. and Zaharia, M., 2021, January. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In Proceedings of CIDR (Vol. 8).

Bajarin, T., 2023.Why Apple Is Not Rushing To Create A More Public Representation Of Its AI Capabilities. Accessed at https://www.forbes.com/sites/timbajarin/2023/02/21/why-apple-is-not-rushing-to-create-a-more-public-representation-of-its-ai-capabilities/# (Accessed 20.03.2023)

Barad, K., 2003. Posthumanist performativity: Toward an understanding of how matter comes to matter. *Signs: Journal of women in culture and society*, *28*(3), pp.801-831.

Barad, K., 2007. *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.

Barocas, S. and Selbst, A.D., 2016. Big data's disparate impact. Calif. L. Rev., 104, p.671.

Beauchamp T.L., Childress J.F., 2001. Principles of biomedical ethics, 5th ed. New York City, NY: Oxford University Press.

Bechmann, A. and Bowker, G.C., 2019. Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society*, *6*(1), p.2053951718819569.

Beer, D., 2022. The problem of researching a recursive society: Algorithms, data coils and the looping of the social. Big Data & Society, 9(2), p.20539517221104997.

Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S., 2021, March. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

Benjamin, G., 2020, December. "Put It In the Bin": Mapping AI as a framework of refusal. In *Resistance AI Workshop at NeurIPS2020*.

Benjamin, R., 2019. Race after technology: Abolitionist tools for the new jim code. Social Forces.

Benjamin, W., 2008. *The work of art in the age of its technological reproducibility, and other writings on media*. Harvard University Press.

Bennett, C.L. and Keyes, O., 2020. What is the point of fairness? Disability, AI and the complexity of justice. *ACM SIGACCESS Accessibility and Computing*, (125), pp.1-1.

Bennett, SJ., Claisse, C., Luger, E., and Durrant, A., 2023. Unpicking Epistemic Injustices in Digital Health: On Designing Data-Driven Technologies to Support the Self-Management of Long-Term Health Conditions. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. (Forthcoming)

Berard, T.J., 2005. Rethinking practices and structures. Philosophy of the social sciences, 35(2), pp.196-230.

Berlin, L., 2017. *Troublemakers: how a generation of silicon valley upstarts invented the future*. Simon and Schuster.

Best, J., 2012. Ambiguity and uncertainty in international organizations: A history of debating IMF conditionality. International Studies Quarterly, 56(4), pp.674-688.

Bhandari, A., Ozanne, M., Bazarova, N.N. and DiFranzo, D., 2021. Do you care who flagged this post? Effects of moderator visibility on bystander behavior. *Journal of Computer-Mediated Communication*, *26*(5), pp.284-300.

Birhane, A., 2021. Algorithmic injustice: a relational ethics approach. *Patterns*, *2*(2), p.100205.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R. and Bao, M., 2022, June. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 173-184).
Blin, F., 2016. The theory of affordances. Language-learner computer interactions: theory, methodology and CALL applications, pp.41-64.

Bloomberg, 2022. $422.37+ Billion Global Artificial Intelligence (AI) Market Size Likely to Grow at 39.4% CAGR During 2022-2028 | Industry. Accessed at https://www.bloomberg.com/press-releases/2022-06-27/-422-37-billion-global-artificial-intelligence-ai-market-size-likely-to-grow-at-39-4-cagr-during-2022-2028-industry (Accessed on 15. 01. 2023)

Bogner, A., Littig, B. and Menz, W. eds., 2009. *Interviewing experts*. Springer.

Borning, A. and Muller, M., 2012, May. Next steps for value sensitive design. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1125-1134).

Bratteteig, T. and Verne, G.B., 2012. Conditions for Autonomy in the Information Society: Disentangling as a public service. *Scandinavian Journal of Information Systems*, *24*(2), p.3.

Bregnbæk, S., 2021. Questioning care: ambiguous relational ethics between a refugee child, her parents and the Danish welfare state. *International Journal of Inclusive Education*, *25*(2), pp.196-209.

Brinkmann, S. and Kvale, S., 2018. *Doing interviews* (Vol. 2). Sage.

British Computer Society (BCS: The Chartered Institute for IT) 2020. *BCS Diversity Report 2020: ONS Analysis*. Accessed at https://www.bcs.org/media/5766/diversity-report-2020-part2.pdf (Accessed 27. 08. 2021)

Brockman, J., 1998. Consciousness is a big suitcase: a talk with Marvin Minsky. *Edge. org*.

Buchanan, R., 1992. Wicked problems in design thinking. Design issues, 8(2), pp.5-21.

Buckman, J., 2021. Please Commit More Blatant Academic Fraud. Accessed at https://jacobbuckman.com/2021-05-29-please-commit-more-blatant-academic-fraud/ (Accessed on 14. 07. 2021)

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.

Byers, W., 2010. How mathematicians think. In How Mathematicians Think. Princeton University Press.

Cadwalladr, C. and Graham-Harrison, E., 2018. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The guardian*, *17*(1), p.22.

Campolo, A. and Crawford, K., 2020. Enchanted determinism: Power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society*, *6*, pp.1-19.

Cassam, Q., 2016. Vice epistemology. *The Monist*, *99*(2), pp.159-180.

Catanzariti, B., Chandhiramowuli, S., Mohamed, S., Natarajan, S., Prabhat, S., Raval, N., Taylor, A.S. and Wang, D., 2021, October. The Global Labours of AI and Data Intensive Systems. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 319-322).

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J. and Mullainathan, S., 2016. Productivity and selection of human capital with machine learning. American Economic Review, 106(5), pp.124-27.

Christin, A., 2020. The ethnographer and the algorithm: beyond the black box. *Theory and Society*, *49*(5-6), pp.897-918.

Clouser, K.D. and Gert, B., 1990. A critique of principlism. *The Journal of medicine and philosophy*, *15*(2), pp.219-236.

Cohen, S.S. and Fields, G., 1999. Social capital and capital gains in Silicon Valley. *California management review*, *41*(2), pp.108-130.

Cogin, J., 2012. Are generational differences in work values fact or fiction? Multi-country evidence and implications. *The International Journal of Human Resource Management*, *23*(11), pp.2268-2294.

Collins, H. and Evans, R., 2014. Quantifying the tacit: The imitation game and social fluency. *Sociology*, *48*(1), pp.3-19.

Conway, P. and Gawronski, B., 2013. Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *Journal of personality and social psychology*, *104*(2), p.216.

Corbett-Davies, S. and Goel, S., 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

Costanza-Chock, S., 2018. Design justice: Towards an intersectional feminist framework for design theory and practice. *Proceedings of the Design Research Society*.

Costanza-Chock, S., 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Costello, K., 2019. Gartner Survey Shows 37 Percent of Organizations Have Implemented AI in Some Form. Accessed at https://www.gartner.com/en/newsroom/press-releases/2019-01-21-gartner-survey-shows-37-percent-of-organizations-have (Accessed 10. 06. 2021)

Crawford, K., 2016. Artificial Intelligence's White Guy Problem. Accessed at https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html (Accessed 12. 06. 2020)

Dastin, J., 2018. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics* (pp. 296-299). Auerbach Publications.

Dastin, J., 2023. Microsoft-backed OpenAI starts release of powerful AI known as GPT-4. Available at https://www.reuters.com/technology/microsoft-backed-openai-starts-release-powerful-ai-known-gpt-4-2023-03-14/ (Accessed 20.03. 2023)

de Beauvoir, S., 1947. *The Ethics of Ambiguity*. Translated by B. Frechtman. Reprint, New York: Open Road Media, 2018

De Vries, T., Misra, I., Wang, C. and Van der Maaten, L., 2019. Does object recognition work for everyone?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 52-59).

Decety, J. and Cowell, J.M., 2014. The complex relation between morality and empathy. *Trends in cognitive sciences*, *18*(7), pp.337-339.

Diochon, M. and Anderson, A.R., 2011. Ambivalence and ambiguity in social enterprise; narratives about values in reconciling purpose and practices. *International Entrepreneurship and Management Journal*, *7*, pp.93-109.

Domen, R.E., 2016. The ethics of ambiguity: rethinking the role and importance of uncertainty in medical education and practice. *Academic Pathology*, *3*, p.2374289516654712.

Dourish, P., 2004. What we talk about when we talk about context. Personal and ubiquitous computing, 8(1), pp.19-30.

Dreyfus, H.L., 1965. Alchemy and artificial intelligence. RAND CORP SANTA MONICA CALIF

Dwyer, S.C. and Buckle, J.L., 2009. The space between: On being an insider-outsider in qualitative research. International journal of qualitative methods, 8(1), pp.54-63.

Dzuranin, A.C., Shortridge, R.T. and Smith, P.A., 2013. Building ethical leaders: A way to integrate and assess ethics education. Journal of business ethics, 115(1), pp.101-114.

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, pp.131-138.

Elish, M.C. and Boyd, D., 2018. Situating methods in the magic of Big Data and AI. Communication monographs, 85(1), pp.57-80.

Elliott, C., 2014. *A philosophical disease: Bioethics, culture, and identity*. Routledge.

Eisenberg, E.M., 1984. Ambiguity as strategy in organizational communication. Communication monographs, 51(3), pp.227-242.

Elliott, C., 2007. The tyranny of expertise. In *The Ethics of Bioethics: Mapping the Moral Landscape* (pp. 43-46). The Johns Hopkins University Press.

Emmerich, N., 2015. What is Bioethics?. *Medicine, Health Care and Philosophy*, *18*, pp.437-441.

Evans, J.H., 2000. A sociological account of the growth of principlism. *Hastings Center Report*, *30*(5), pp.31-39.

Feldman, M.S. and Orlikowski, W.J., 2011. Theorizing practice and practicing theory. *Organization science*, *22*(5), pp.1240-1253.

Feldstein, S., 2019. The global expansion of AI surveillance (Vol. 17). Washington, DC: Carnegie Endowment for International Peace.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F. and Schafer, B., 2018. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), pp.689-707.

Fiebrink, R., 2019. Machine learning education for artists, musicians, and other creative practitioners. ACM Transactions on Computing Education (TOCE), 19(4), pp.1-32.

Fiesler, C., Garrett, N. and Beard, N., 2020, February. What do we teach when we teach tech ethics? A syllabi analysis. In *Proceedings of the 51st ACM technical symposium on computer science education* (pp. 289-295).

Fiesler, C., Friske, M., Garrett, N., Muzny, F., Smith, J.J. and Zietz, J., 2021, March. Integrating Ethics into Introductory Programming Classes. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (pp. 1027-1033).

Finlay, L., 2002. "Outing" the researcher: The provenance, process, and practice of reflexivity. *Qualitative health research*, *12*(4), pp.531-545.

Fisher, C.K., Smith, A.M. and Walsh, J.R., 2019. Machine learning for comprehensive forecasting of Alzheimer's Disease progression. Scientific reports, 9(1), pp.1-14.

Fleischmann, K.R., Wallace, W.A. and Grimes, J.M., 2010, January. The values of computational modelers and professional codes of ethics: Results from a field study. In 2010 43rd Hawaii International Conference on System Sciences (pp. 1-10). IEEE.

Fochler, M. and Sigl, L., 2018. Anticipatory uncertainty: How academic and industry researchers in the life sciences experience and manage the uncertainties of the research process differently. Science as Culture, 27(3), pp.349-374.

Fricker, M., 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Fricker, M., 2017. Evolving concepts of epistemic injustice. In *The Routledge handbook of epistemic injustice.* Routledge, 53–60.

Friedman, B., Kahn, P. and Borning, A., 2002. Value sensitive design: Theory and methods. *University of Washington technical report*, *2*, p.12.

Friedman, B. and Nissenbaum, H., 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), pp.330-347.

Friedman, B. and Kahn Jr, P.H., 2007. Human values, ethics, and design. In *The human-computer interaction handbook* (pp. 1267-1292). CRC press.

Gaggiotti, H. ed., 2016. Organizational ethnography: an experiential and practical guide. *qualitative research*, 20(2), pp.194-212.

Galison, P., 1995. Theory bound and unbound: Superstrings and experiment. *Laws of nature: essays on the philosophical, scientific and historical dimensions*, pp.369-408.

Galison, P., 1995. *Context and constraints. Scientific practice: Theories and stories of doing physics,* pp.13-41.

Galison, P.L. and D'Agostino, S., 1987. *How experiments end*(Vol. 88). Chicago: University of Chicago Press.

Garg, N., Schiebinger, L., Jurafsky, D. and Zou, J., 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), pp.E3635-E3644.

Gergen, K.J., 2005. Narrative, moral identity, and historical consciousness. *Narration, identity, and historical consciousness*, *3*, p.99.

Gibney, E., 2020. The battle for ethical AI at the world's biggest machine-learning conference. Nature, 577(7792), pp.609-610.

Gleiberman, M., 2023. *Effective Altruism and the strategic ambiguity of 'doing good'*. IOB Discussion Paper, 2023.01). Institute of Development Policy (IOB), University of Antwerp.

Gleichgerrcht, E. and Young, L., 2013. Low levels of empathic concern predict utilitarian moral judgment. *PloS one*, *8*(4), p.e60418.

Glikson, E. and Woolley, A.W., 2020. Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), pp.627-660

Goldkuhl, G., 2012. Pragmatism vs interpretivism in qualitative information systems research. European journal of information systems, 21(2), pp.135-146.

Gonsalves, T., 2019. The summers and winters of artificial intelligence. In *Advanced methodologies and technologies in artificial intelligence, computer simulation, and human-computer interaction* (pp. 168-179). IGI Global.

Gorichanaz, T., 2020, April. Engaging with public art: An exploration of the design space. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

Gray, M.L. and Suri, S., 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

Grinnell, F., 2013. Research integrity and everyday practice of science. Science and Engineering Ethics, 19, pp.685-701.

Grove, W., Goldin, J.A., Breytenbach, J. and Suransky, C., 2021. Taking togetherness apart: From digital footprints to geno-digital spores. *Human Geography*, p.19427786211024264.

Güss, C.D., Burger, M.L. and Dörner, D., 2017. The role of motivation in complex problem solving. *Frontiers in psychology*, *8*.

Hampton, L.M., 2021. Black feminist musings on algorithmic oppression. *arXiv preprint arXiv:2101.09869*.

Haraway, D., 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies*, *14*(3), pp.575-599.

Heimer, C.A., 2013. 'Wicked'ethics: Compliance work and the practice of ethics in HIV research. Social science & medicine, 98, pp.371-378.

Himma, K.E., 2009. Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?. *Ethics and Information Technology*, *11*(1), pp.19-29.

Hinton, G., 2018. Deep learning—a technology with the potential to transform health care. Jama, 320(11), pp.1101-1102.

Hoffman, M.L., 2001. *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.

Hoffman, S.F. and Friedman, H.H., 2018. Machine learning and meaningful careers: increasing the number of women in STEM. *J. Res. Gender Stud.*, *8*, p.11.

Hoffman, S.G., 2017. Managing ambiguities at the edge of knowledge: Research strategy and artificial intelligence labs in an era of academic capitalism. *Science, technology, & human values*, *42*(4), pp.703-740.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M. and Wallach, H., 2019, May. Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-16).

Horowitz, M.C., 2016. The ethics & morality of robotic warfare: Assessing the debate over autonomous weapons. Daedalus, 145(4), pp.25-36.

Huntingford, C., Jeffers, E.S., Bonsall, M.B., Christensen, H.M., Lees, T. and Yang, H., 2019. Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, *14*(12), p.124007.

Ihde, D., 2008. Introduction: postphenomenological research. Human Studies, 31, pp.1-9.

Introna, L., 2005. Phenomenological approaches to ethics and information technology.

Jacobs, A.Z., 2021. Measurement as governance in and for responsible AI. *arXiv preprint arXiv:2109.05658*.

Jakesch, M., Buçinca, Z., Amershi, S. and Olteanu, A., 2022, June. How different groups prioritize ethical values for responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 310-323).

Jaton, F., 2021. Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. Big Data & Society, 8(1), p.20539517211013569.

Jo, E.S. and Gebru, T., 2020, January. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 306-316).

Jobin, A., Ienca, M. and Vayena, E., 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), pp.389-399.

Johnson, D.G., 1978. Computer Ethics: New Study Area for Engineering Science Students. Professional Engineer, 48(8), pp.32-4.

Kang, E.B., 2023. Ground truth tracings (GTT): On the epistemic limits of machine learning. Big Data & Society, 10(1), p.20539517221146122.

Kaplan, A. and Haenlein, M., 2019. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), pp.15-25.

Karasti, H., 2001. Bridging work practice and system design: integrating systemic analysis, appreciative intervention and practitioner participation. *Computer supported cooperative work (CSCW)*, *10*, pp.211-246.

Kidd, D. and Castano, E., 2019. Reading literary fiction and theory of mind: Three preregistered replications and extensions of Kidd and Castano (2013). *Social Psychological and Personality Science*, *10*(4), pp.522-531.

Kieran, M., 1996. Art, imagination, and the cultivation of morals. *The Journal of Aesthetics and Art Criticism*, *54*(4), pp.337-351.

Kim, M., Zimmermann, T., DeLine, R. and Begel, A., 2016, May. The emerging role of data scientists on software development teams. In 2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE) (pp. 96-107). IEEE.

King, N., 2012. Doing template analysis. Qualitative organizational research: Core methods and current challenges, 426, pp.77-101.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S., 2018. Human decisions and machine predictions. The quarterly journal of economics, 133(1), pp.237-293.

Klumbytė, G., Draude, C. and Taylor, A.S., 2022, June. Critical tools for machine learning: Working with intersectional critical concepts in machine learning systems design. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1528-1541).

Kolenik, T. and Gams, M., 2021. Persuasive technology for mental health: One step closer to (Mental health care) equality?. *IEEE Technology and Society Magazine*, *40*(1), pp.80-86.

Ladhari, R. and Tchetgna, N.M., 2017. Values, socially conscious behaviour and consumption emotions as predictors of Canadians' intent to buy fair trade products. International Journal of Consumer Studies, 41(6), pp.696-705

Lareau, A., 2018. *Journeys through ethnography: Realistic accounts of fieldwork*. Routledge.
Latour, B. and Woolgar, S., 2013. *Laboratory life: The construction of scientific facts*. Princeton University Press.

Lee, K.F., 2018. *AI superpowers: China, Silicon Valley, and the new world order*. Houghton Mifflin.

Leith, P., 2016. The rise and fall of the legal expert system. *International Review of Law, Computers & Technology*, *30*(3), pp.94-106.

Leonelli, S., 2016. Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), p.20160122.

Leurs, K., 2017. Feminist data studies: Using digital methods for ethical, reflexive and situated socio-cultural research. *Feminist Review*, *115*(1), pp.130-154.

Littman, M.L., 2021. Collusion rings threaten the integrity of computer science research. Communications of the ACM, 64(6), pp.43-44.

Loescher, K.J., Hughes, R.W., Cavico, F., Mirabella, J. and Pellet, P.F., 2005. The impact of an "Ethics across the curriculum" initiative on the cognitive moral development of business school undergraduates. Teaching Ethics, 5(2), pp.31-72.

Lounsbury, J.W., Foster, N., Patel, H., Carmody, P., Gibson, L.W. and Stairs, D.R., 2012. An investigation of the personality traits of scientists versus nonscientists and their relationship with career satisfaction. *R&D Management*, *42*(1), pp.47-59.

Luan, H. and Tsai, C.C., 2021. A review of using machine learning approaches for precision education. Educational Technology & Society, 24(1), pp.250-266.

Luther, V.P. and Crandall, S.J., 2011. Commentary: ambiguity and uncertainty: neglected elements of medical education curricula?. Academic Medicine, 86(7), pp.799-800.

Lynch, M., 1988. The externalized retina: Selection and mathematization in the visual documentation of objects in the life sciences. Human studies, 11(2-3), pp.201-234.

Mackenzie, A., 2017. Machine learners: Archaeology of a data practice. MIT Press.

Mackworth-Young, C.R., Schneiders, M.L., Wringe, A., Simwinga, M. and Bond, V., 2019. Navigating 'ethics in practice': An ethnographic case study with young women living with HIV in Zambia. *Global public health*, *14*(12), pp.1689-1702.

Madaio, M.A., Stark, L., Wortman Vaughan, J. and Wallach, H., 2020, April. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

Manders-Huits, N., 2011. What values in design? The challenge of incorporating moral values into design. *Science and engineering ethics*, *17*(2), pp.271-287.

marquis de Laplace, P.S., 1902. A philosophical essay on probabilities. Wiley.

Mason, R.O., 1986. Four ethical issues of the information age. MIS quarterly, pp.5-12.

McDermott, D., 1976. Artificial intelligence meets natural stupidity. Acm Sigart Bulletin, (57), pp.4-9.

McLennan, S., Lee, M.M., Fiske, A. and Celi, L.A., 2020. AI ethics is not a panacea. The American Journal of Bioethics, 20(11), pp.20-22.

McStay, A., 2020. Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. Big Data & Society, 7(1), p.2053951720904386.

Mendon-Plasek, A., 2021. Mechanized significance and machine learning: why it became thinkable and preferable to teach machines to judge the world. *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*, pp.31-78.

Mencl, J. and May, D.R., 2009. The effects of proximity and empathy on ethical decision-making: An exploratory investigation. *Journal of Business Ethics*, 85, pp.201-226.

Mhlambi, S. and Tiribelli, S., 2023. Decolonizing AI ethics: relational autonomy as a means to counter AI harms. Topoi, pp.1-14.

Miceli, M., Posada, J. and Yang, T., 2022. Studying up machine learning data: Why talk about bias when we mean power?. *Proceedings of the ACM on Human-Computer Interaction*, *6*(GROUP), pp.1-14.

Miller, B., 2021. Is Technology Value-Neutral?. Science, Technology, & Human Values, 46(1), pp.53-80.

Millett, L.I., Friedman, B. and Felten, E., 2001, March. Cookies and web browser design: Toward realizing informed consent online. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 46-52).

Mittelstadt, B., 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), pp.501-507.

Moardi, M., Salehi, M. and Marandi, Z., 2016. The role of tolerance of ambiguity on ethical decision-making students: A comparative study between accounting and management students. *Humanomics*.

Monett, D. and Lewis, C.W., 2018. Getting clarity by defining artificial intelligence—A survey. In Philosophy and theory of artificial intelligence 2017 (pp. 212-214). Springer International Publishing.

Moor, J.H., 1985. What is computer ethics?. *Metaphilosophy*, *16*(4), pp.266-275.

Morley, J., Floridi, L., Kinsey, L. and Elhalal, A., 2021. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In *Ethics, Governance, and Policies in Artificial Intelligence* (pp. 153-183). Springer, Cham.

Morrison, C., Cutrell, E., Grayson, M., Thieme, A., Taylor, A., Roumen, G., Longden, C., Tschiatschek, S., Faia Marques, R. and Sellen, A., 2021, May. Social Sensemaking with AI: Designing an Open-ended AI experience with a Blind Child. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-14).

Moss, E. and Metcalf, J., 2020. Ethics owners: A new model of organizational responsibility in data-driven technology companies.

Moss, E., 2022. The Objective Function: Science and Society in the Age of Machine Intelligence. *arXiv preprint arXiv:2209.10418*.

Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q.V., Dugan, C. and Erickson, T., 2019, May. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-15).

Neto, C., 2020. When imprecision is a good thing, or how imprecise concepts facilitate integration in biology. Biology & Philosophy, 35(6), pp.1-21.

Nicholas, L., 2021. Remembering Simone de Beauvoir's 'ethics of ambiguity' to challenge contemporary divides: feminism beyond both sex and gender. *Feminist Theory*, *22*(2), pp.226-247.

Niiniluoto, I., 2018. *Truth-seeking by abduction* (Vol. 400). Springer.

Noble, S.U., 2013. Google search: Hyper-visibility as a means of rendering black women and girls invisible. InVisible Culture, (19).

Nielsen, N.R., 1972. Social responsibility and computer education. *ACM SIGCSE Bulletin*, *4*(1), pp.90-96.

Nowell, L.S., Norris, J.M., White, D.E. and Moules, N.J., 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods*, 16(1), p.1609406917733847.

Nyeko, K.E. and Sing, N.K., 2015. Academic entrepreneurs and entrepreneurial academics: are they the same. International *Journal of Social Science and Humanity*, 5(12), p.1050.

O'flynn, P., 2009. The Creation of Meaning: Simone de Beauvoir's Existentialist Ethics. Minerva: An Internet Journal of Philosophy, 13.

Oganowski, K., 2013. *Centralizing Ambiguity: Simone de Beauvoir and a Twenty-First Century Ethics* (Doctoral dissertation, Syracuse University).

Orlikowski, W.J. and Baroudi, J.J., 1991. Studying information technology in organizations: Research approaches and assumptions. *Information systems research*, *2*(1), pp.1-28.

Orr, W. and Davis, J.L., 2020. Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society*, *23*(5), pp.719-735.

Orlikowski, W.J., 2007. Sociomaterial practices: Exploring technology at work. *Organization studies*, *28*(9), pp.1435-1448.

Orr, J.E., 2016. *Talking about machines: An ethnography of a modern job*. Cornell University Press.

Parker, E.A., 2015. Singularity in Beauvoir's The Ethics of Ambiguity. The Southern Journal of Philosophy, 53(1), pp.1-16.

Parnas, D.L., 1999. Software engineering programs are not computer science programs. *IEEE software*, *16*(6), pp.19-30.

Parson, L., 2019. Considering positionality: The ethics of conducting research with marginalized groups. *Research methods for social justice and equity in education*, pp.15-32.

Passi, S. and Sengers, P., 2020. Making data science systems work. Big Data & Society, 7(2), p.2053951720939605.\

Patel, K., Fogarty, J., Landay, J.A. and Harrison, B., 2008, April. Investigating statistical machine learning as a tool for software development. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 667-676).

Patton, M.Q., 2014. *Qualitative research & evaluation methods: Integrating theory and practice*. Sage publications.

Paul, J. and Strbiak, C.A., 1997. The ethics of strategic ambiguity. *The Journal of Business Communication (1973)*, *34*(2), pp.149-159.

Perez, C.C., 2019. *Invisible women: Exposing data bias in a world designed for men*. Random House.

Petrozzino, C. Who pays for ethical debt in AI?. AI Ethics 1, 205–208 (2021). https://doi.org/10.1007/s43681-020-00030-3

Phan, T., Goldenfein, J., Mann, M. and Kuch, D., 2022. Economies of virtue: the circulation of 'ethics' in Big Tech. Science as culture, 31(1), pp.121-135.

Phan, T. and Wark, S., 2021. What personalisation can do for you! Or: how to do racial discrimination without 'race'. *Culture machine*, *20*, pp.1-29.

Pisano, E.D., 2020. AI shows promise for breast cancer screening. *Nature*, *577*(7788), pp.35-36.

Pohlhaus, G., 2012. Relational knowing and epistemic injustice: Toward a theory of willful hermeneutical ignorance. *Hypatia*, *27*(4), pp.715-735.

Ponomarenko, J.V. and Bourne, P.E., 2007. Antibody-protein interactions: benchmark datasets and prediction tools evaluation. BMC structural biology, 7(1), pp.1-19.

Proctor R., and Schiebinger., L 2008. Agnotology: The Making and Unmaking of Ignorance. Stanford, California: Stanford University Press

Raji, I.D., Scheuerman, M.K. and Amironesei, R., 2021, March. You can't sit with us: exclusionary pedagogy in AI ethics education. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 515-525).

Ray, K.S., 2021. It's time for a black bioethics. *The American Journal of Bioethics*, *21*(2), pp.38-40.

Redman, T.C., 2018. If your data is bad, your machine learning tools are useless. *Harvard Business Review*, *2*.

Renn, O., Klinke, A. and Van Asselt, M., 2011. Coping with complexity, uncertainty and ambiguity in risk governance: a synthesis. Ambio, 40(2), pp.231-246.

Rennstam, J. and Ashcraft, K.L., 2014. Knowing work: Cultivating a practice-based epistemology of knowledge in organization studies. *Human Relations*, *67*(1), pp.3-25.

Resseguier, A. and Rodrigues, R., 2020. AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, 7(2), p.2053951720942541.

Ristoski, P., De Vries, G.K.D. and Paulheim, H., 2016, October. A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web. In *International Semantic Web Conference* (pp. 186-194). Springer, Cham.

Rittel, H.W. and Webber, M.M., 1973. Dilemmas in a general theory of planning. *Policy sciences*, 4(2), pp.155-169.

Robinson, O.C., 2014. Sampling in interview-based qualitative research: A theoretical and practical guide. *Qualitative research in psychology*, 11(1), pp.25-41.

Ross, C. and Swetlitz, I., 2018. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *Stat*, 25.

Roth, S., 2009. New for whom? Initial images from the social dimension of innovation. International Journal of Innovation and Sustainable Development, 4(4), pp.231-252.

Rule, A., Tabard, A. and Hollan, J.D., 2018, April. Exploration and explanation in computational notebooks. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12).

Sabie, S. and Parikh, T., 2019, May. Cultivating care through ambiguity: Lessons from a service learning course. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N. and Beard, N., 2019. Integrating ethics within machine learning courses. *ACM Transactions on Computing Education (TOCE)*, *19*(4), pp.1-26.

Sambasivan, N., 2021. Seeing like a dataset from the global south. *Interactions*, 28(4), pp.76-78.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. and Aroyo, L.M., 2021, May. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).

Sambasivan, N. and Veeraraghavan, R., 2022, April. The deskilling of domain expertise in ai development. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

Sankaran, S., Zhang, C., Gutierrez Lopez, M. and Väänänen, K., 2020, October. Respecting Human Autonomy through Human-Centered AI. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (pp. 1-3).

Schiff, D., Rakova, B., Ayesh, A., Fanti, A. and Lennon, M., 2021. Explaining the Principles to Practices Gap in AI. *IEEE Technology and Society Magazine*, *40*(2), pp.81-94.

Schmidt, F.A., 2019. Crowdsourced production of AI training data: how human workers teach self-driving cars how to see (No. 155). Working Paper Forschungsförderung.

Schoenherr, J.R., 2022. Ethical Artificial Intelligence from Popular to Cognitive Science: Trust in the Age of Entanglement. Taylor & Francis.

Schwandt, T.A., 1994. Constructivist, interpretivist approaches to human inquiry. *Handbook of Qualitative Research*, 1, pp.118-137.

Seaver, N., 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big data & society*, *4*(2), p.2053951717738104.

Seaver, N., 2018. What should an anthropology of algorithms do?. *Cultural anthropology*, *33*(3), pp.375-385.

Sennet, A., 2011. *Ambiguity*. Accessible at http://seop.illc.uva.nl/entries/ambiguity/ (Accessed 07. 05. 2022)

Shapin, S., 2018. Making Art/Discovering Science. *KNOW: A Journal on the Formation of Knowledge*, *2*(2), pp.177-205.

Shaw, N.P., Stöckel, A., Orr, R.W., Lidbetter, T.F. and Cohen, R., 2018, December. Towards provably moral AI agents in bottom-up learning frameworks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 271-277).

Shilton, K., 2013. Values levers: Building ethics into design. *Science, Technology, & Human Values*, *38*(3), pp.374-397.

Shilton, K., Koepfler, J.A. and Fleischmann, K.R., 2014, February. How to see values in social computing: methods for studying values dimensions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 426-435).

Shilton, K. and Anderson, S., 2017. Blended, not bossy: Ethics roles, responsibilities and expertise in design. *Interacting with computers*, *29*(1), pp.71-79.

Sim, A.B. and Fernando, M., 2010. Strategic ambiguity and ethical actions.

Simon, J., 2015. Distributed epistemic responsibility in a hyperconnected era. In The Onlife Manifesto (pp. 145-159). Springer, Cham.

Simonite, T., 2021. What really happened when Google ousted Timnit Gebru. WIRED magazine, pp.1-19.

Slota, S.C., Fleischmann, K.R., Greenberg, S., Verma, N., Cummings, B., Li, L. and Shenefiel, C., 2023. Locating the work of artificial intelligence ethics. Journal of the Association for Information Science and Technology, 74(3), pp.311-322.

Smits, M., Ludden, G., Peters, R., Bredie, S.J., Van Goor, H. and Verbeek, P.P., 2022. Values that matter: a new method to design and assess moral mediation of technology. *Design issues*, *38*(1), pp.39-54.

Stark, L. and Crawford, K., 2019. The work of art in the age of artificial intelligence: What artists can teach us about the ethics of data practice. *Surveillance & Society*, *17*(3/4), pp.442-455.

Strauß, S., 2021. Deep Automation Bias: How to Tackle a Wicked Problem of AI?. Big Data and Cognitive Computing, 5(2),

Strickland, E., 2019. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. IEEE Spectrum, 56(4), pp.24-31.

Stengers, I., 2000. *The invention of modern science* (Vol. 19). U of Minnesota Press.

Suchman, L. and Suchman, L.A., 2007. *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.

Suchman, L., 2011. Subject objects. *Feminist theory*, *12*(2), pp.119-145.

Suchman, L., 2012. Configuration. In *Inventive methods* (pp. 48-60). Routledge.

Suchman, L.A. and Trigg, R.H., 1993. *Artificial intelligence as craftwork* (pp. 144-78).

Taber, B.J., Hartung, P.J. and Borges, N.J., 2011. Personality and values as predictors of medical specialty choice. Journal of Vocational Behavior, 78(2), pp.202-209.

Táíwò, O., 2020. Being-in-the-room privilege: Elite capture and epistemic deference. *The Philosopher*, *108*(4), pp.61-70.

Tarvier, E., 2022. https://www.investopedia.com/articles/markets/103015/biggest-companies-silicon-valley.asp (Accessed 20.03.2023)

Taylor, L., 2017. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, *4*(2), p.2053951717736335.

Thomer, A.K., Akmon, D., York, J.J., Tyler, A.R., Polasek, F., Lafia, S., Hemphill, L. and Yakel, E., 2022. The Craft and Coordination of Data Curation: Complicating Workflow Views of Data Science. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), pp.1-29.

Toxtli, C., Suri, S. and Savage, S., 2021. Quantifying the invisible labor in crowd work. Proceedings of the ACM on human-computer interaction, 5(CSCW2), pp.1-26.

Umbrello, S., 2018. The moral psychology of value sensitive design: The methodological issues of moral intuitions for responsible innovation. *Journal of Responsible Innovation*, *5*(2), pp.186-200.
Vallor, S., 2016. *Technology and the virtues*: A philosophical guide to a future worth wanting. Oxford University Press.

Vallor, S., 2021. Mobilising the intellectual resources of the arts and humanities. Accessed at https://www.adalovelaceinstitute.org/blog/mobilising-intellectual-resources-arts-humanities/ (Accessed 22. 05. 2022)

Van den Broek, E., Sergeeva, A. and Huysman, M., 2021. When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring. MIS quarterly, 45(3).

Van Dijk, L. and Rietveld, Ex., 2017. Foregrounding sociomaterial practice in our understanding of affordances: The skilled intentionality framework. *Frontiers in psychology*, *7*, p.1969.

van Grunsven, J. and van Wynsberghe, A., 2019. A Semblance of Aliveness: How the Peculiar Embodiment of Sex Robots Will Matter. *Techné: Research in Philosophy and Technology*, *23*(3), pp.290-317.

Van Wynsberghe, A., 2013. Designing robots for care: Care centered value-sensitive design. *Science and engineering ethics*, *19*(2), pp.407-433.

Van Wynsberghe, A. and Robbins, S., 2014. Ethicist as designer: a pragmatic approach to ethics in the lab. Science and engineering ethics, 20, pp.947-961.

Vance, A., 2021. Oral History: Geoff Hinton On How AI Came To Be And What We're Supposed To Do With It. Accessed at https://ashleevance.substack.com/p/oralhistorywithgeoffhinton. (Accessed 07. 08. 2022)

Varma, P. What to know about OpenAI, the company behind ChatGPT. Accessed at https://www.washingtonpost.com/technology/2023/02/06/what-is-openai-chatgpt/. (Accessed 20. .03. 2023)

Veale, M., Van Kleek, M. and Binns, R., 2018, April. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In Proceedings of the 2018 chi conference on human factors in computing systems (pp. 1-14).

Verbeek, P.P., 2005. *What things do: Philosophical reflections on technology, agency, and design*. Penn State Press.

Wagner, B., 2018. Ethics as an escape from regulation: From ethics-washing to ethics-shopping. Being profiling. Cogitas ergo sum, pp.1-7.

Wall, S. and Schellmann, H., 2021. Linkedin's job-matching AI was biased. The company's solution? More AI.

Wang P., 2019. On Defining Artificial Intelligence. Journal of Artificial General Intelligence 10(2) 1-37, DOI: 10.2478/jagi-2019-0002

Wang, Q., Jing, S. and Goel, A.K., 2022, June. Co-Designing AI Agents to Support Social Connectedness Among Online Learners: Functionalities, Social Characteristics, and Ethical Challenges. In Designing Interactive Systems Conference (pp. 541-556).

Washington, A.L., 2018. How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. Colo. Tech. LJ, 17, p.131.

Weber, Max. 1946. "Science as a Vocation." In From Max Weber: Essays in Sociology, translated by H.H. Gerth and C.W. Mills, 129–56. New York: Oxford University Press. 1992. The Protestant Ethic and the Spirit of Capitalism, translated by Talcott Parsons. London: Routledge.

Weisbrod, E., 2009. The role of affect and tolerance of ambiguity in ethical decision making. *Advances in Accounting*, *25*(1), pp.57-63.

Weizenbaum, J., 1976. Computer power and human reason: From judgment to calculation.

Wheelwright, P.E., 1962. Metaphor & reality. In Eisenberg, E.M., 1984. Ambiguity as strategy in organizational communication. *Communication monographs*, *51*(3), pp.227-242.

Whittaker, M., Alper, M., Bennett, C.L., Hendren, S., Kaziunas, L., Mills, M., Morris, M.R., Rankin, J., Rogers, E., Salas, M. and West, S.M., 2019. Disability, bias, and AI. *AI Now Institute*, p.8.

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

Wiggins, B.J. and Christopherson, C.D., 2019. The replication crisis in psychology: An overview for theoretical and philosophical psychology. Journal of Theoretical and Philosophical Psychology, 39(4), p.202.

Winston, P.H., 2016. Marvin L. Minsky (1927–2016). *Nature*, *530*(7590), pp.282-282.

Wong, R.Y., Madaio, M.A. and Merrill, N., 2023. Seeing like a toolkit: How toolkits envision the work of AI ethics. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), pp.1-27.

Xu, W., 2019. Toward human-centered AI: a perspective from human-computer interaction. *interactions*, *26*(4), pp.42-46.

Yampolskiy, R.V., 2020. On defining differences between intelligence and artificial intelligence. Journal of Artificial General Intelligence, 11(2), pp.68-70.

Yang, Q., Steinfeld, A., Rosé, C. and Zimmerman, J., 2020, April. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In Proceedings of the 2020 chi conference on human factors in computing systems (pp. 1-13).

Young, L. and Koenigs, M., 2007. Investigating emotion in moral cognition: a review of evidence from functional neuroimaging and neuropsychology. *British medical bulletin*, *84*(1), pp.69-79.

Young, M., Katell, M. and Krafft, P.M., 2022, June. Confronting Power and Corporate Capture at the FAccT Conference. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1375-1386).

Zhao, J., He, N. and Lovrich, N.P., 1999. Value change among police officers at a time of organizational reform: a follow-up study using Rokeach

Zhou, Z.H., 2018. Machine learning challenges and impact: an interview with Thomas Dietterich. National Science Review, 5(1), pp.54-58.

Zou, J. and Schiebinger, L., 2018. AI can be sexist and racist—it's time to make it fair.