Platform mediated ethics: exploring the tools used to engineer Machine Learning systems

Thesis proposal

Glen Berman

September 2021

Abstract

In this research project, we propose to study the relations between Machine Learning practitioners, their engineering tools, and the social outcomes of Machine learning systems (ML systems). The potential for ML systems to cause social harm and exacerbate inequity is well documented by researchers in the fair machine learning (fair-ML) community, the human-computer interaction community, and investigative journalists. In response, researchers have proposed various interventions in the design and development of ML systems, including design processes and audit frameworks, intended to shape and review deployments of ML systems. In addition, actors in industry and the public sector have promulgated high-level governing principles, intended to guide ML system design and development. These research and policy making efforts, however, generally do not consider how the process of ML system development and deployment is changing due to the recent scaling of Machine Learning platforms – emerging socio-technical systems designed to automate, simplify, and expedite the engineering of ML systems. This research project thus explores how the affordances of Machine Learning platforms create new potentialities for social harm and new opportunities for intervention in ML system design and development. We argue that Machine Learning platforms reconfigure both the process of ML system development, and the enactment of ethics considerations during such development. Further, we argue that proposed interventions in ML system design and development should be evaluated in the context of the tools used to engineer ML systems.

Contents

Chapter 1

Introduction

This research project is concerned with the engineering tools used to develop Machine Learning systems [Mitchell, 1997], and their relationship to the social consequences of the deployment of such systems [Barocas et al., 2021b]. To explore this relationship we focus on Machine Learning platforms, cloud-based tools for managing the life-cycle of Machine Learning development and deployment, and the enactment of ethics considerations by Machine Learning practitioners who use such platforms. The core premise motivating this research project is that these tools are not neutral, that they reflect particular social, economic, cultural, and political values [Davis, 2020]. We argue that, by reflecting the values of the technology firms who operate them, Machine Learning platforms envision a particular future for the development and deployment of Machine Learning systems, and influence the ways in which ethics considerations are enacted by Machine Learning practitioners. However, existing research on the enactment of ethics considerations by Machine Learning practitioners does not account for the role of engineering tools in Machine Learning system development. Similarly, existing proposals for interventions in Machine Learning practitioners' enactment of ethics considerations does not consider how practitioners are constrained by the engineering tools they use. To understand and incorporate the relationship between engineering tools and the social consequences of the deployment of Machine Learning systems requires new theoretical and methodological work.

In this introductory chapter we first situate this research proposal within the fields of Machine Learning and fair-ML. We then review key machine learning and platform terminology that are relied on throughout the proposal, before introducing our research questions. We conclude this chapter by providing an overview of the remainder of this research proposal.

1.1 Situating this proposal

Machine Learning is a sub-discipline of Artificial Intelligence focused on developing computing systems which use statistical probability methods to perform prediction tasks [High-Level Independent Group on Artificial Intelligence (AI HLEG), 2019]. The Machine Learning sub-discipline has had success in developing computing sys-

tems that solve problems in a range of areas that traditional computing methods have long struggled with, from speech recognition, to computer vision, to language translation, to game playing [Russell and Norvig, 2021]. The Machine Learning subdiscipline solves these problems by conceptualising them as prediction tasks which a computing system can complete through use of an appropriately designed statistical model. In recent years, this approach has been applied to numerous decision making processes in commercial and government settings, at times very successfully [Bernardi et al., 2019] and at times with unintended consequences [Barocas et al., 2021b]. These include high stakes decisions, such as criminal sentencing [Angwin et al., 2016; Oswald et al., 2018], predictive policing [Marda and Narayan, 2020], employee recruitment and evaluation [Sánchez-Monedero et al., 2020; Wilson et al., 2021], and allocation of health services [Obermeyer et al., 2019].

The unintended consequences of the integration of Machine Learning approaches into decision making processes is the focus of fair-ML, an emerging interdisciplinary community. Fair-ML researchers aim to understand how Machine Learning approaches can produce decision making outcomes that cause social harms [Chouldechova and Roth, 2018]. Social harms, such as decision making processes that are discriminatory towards certain social groups, are understood within fair-ML to be unintended consequences of the poor deployment of Machine Learning approaches in computing systems [Barocas et al., 2021b]. As such, fair-ML researchers also aim to develop interventions to improve the process of designing, evaluating, deploying, and managing computing systems [Katell et al., 2020]. To advance these aims fair-ML researchers draw from a range of social science approaches, including political philosophy [Binns, 2018], legal studies [Green and Viljoen, 2020; Hutchinson and Mitchell, 2018], critical race theory [Hanna et al., 2020], human-computer interaction [Albert and Delano, 2021], economics [Kasy and Abebe, 2020], and Science and Technology Studies [Selbst et al., 2019]. As many of these approaches are themselves interdisciplinary, the boundaries of fair-ML remain fairly fuzzy.

The integration of Machine Learning approaches into computing systems is a complex and resource intensive engineering challenge. Machine Learning platforms (ML platforms) are computing infrastructure designed to address this challenge [Yao et al., 2017]. ML platforms achieve this by providing Machine Learning practitioners cloud access to centrally managed computing infrastructure and services, which alleviate the need for practitioners to individually develop and maintain every component of their own Machine Learning development process [Roy et al., 2019]. ML platforms are widely used to integrate Machine Learning approaches into computing systems designed to automate or improve decision making processes. Therefore, as objects of study ML platforms are clearly relevant to the fair-ML community. In the following chapters, however, we will make the case that ML platforms are understudied in the fair-ML discourse.

1.2 Machine Learning and platform terminology

Given the interdisciplinary nature of this research project, and the abundance of inconsistent terminology used by technology firms, in order to describe this research project's objectives or scope in any detail we must first explain how we use several key terms.

1.2.1 The levels of computing systems

In this research proposal we consider any system that incorporates human and technical components, and requires some form of coordination between these components, to be a *socio-technical system* [Bauer and Herder, 2009]. Human components in a socio-technical system may include the people that commission, design, engineer, test, use, decommission, or are subjected to the technical components of the system. For this proposal the virtue of the term socio-technical system lies in the way it reminds us that all technical systems are socio-technical systems. We explore the usefulness of conceptualising ML systems and ML platforms as socio-technical systems further in Section 2.5. As we will explore in Section 2.4, many of the unintended consequences of the integration of ML approaches into decision making processes can be traced back to a poor understanding of the interactions between social and technical components of socio-technical systems.

We use the term *computing system* to describe any engineered system that incorporates components that process data or perform calculations (i.e. that compute). As computing systems are engineered – meaning, they are the products of human ingenuity and labour – they are by definition socio-technical systems. Computing systems include discrete computer devices – a smartphone, a laptop, a server – as well as systems consisting of computing and non-computing components – a modern vehicle, 'smart' appliances, a hospital's patient management system. As the last example indicates, our use of the term computing system is very broad, and spans many types of socio-technical systems. *ML platforms* are computing systems which are used to derive components of computing systems that incorporate Machine Learning approaches, and to manage these computing systems. To avoid these circular references to computing systems it is helpful to adopt terminology that distinguishes between different levels of computing systems. Each of these levels can be considered socio-technical systems in their own right. In Figure 1.1 we provide a diagram of our conceptualisation of the different levels of computing systems, and their relationship to ML platforms. An additional summary of the levels of computing systems referred to in this research project is presented in Table 1.1.

1.2.1.1 Precursor levels of computing systems

We have already described Machine Learning as a discipline focused on developing computing systems which use statistical probability methods to perform prediction tasks. A *machine learning model* (*ML model*) is thus a statistical model for processing data to perform a particular predictive task. To develop computing systems

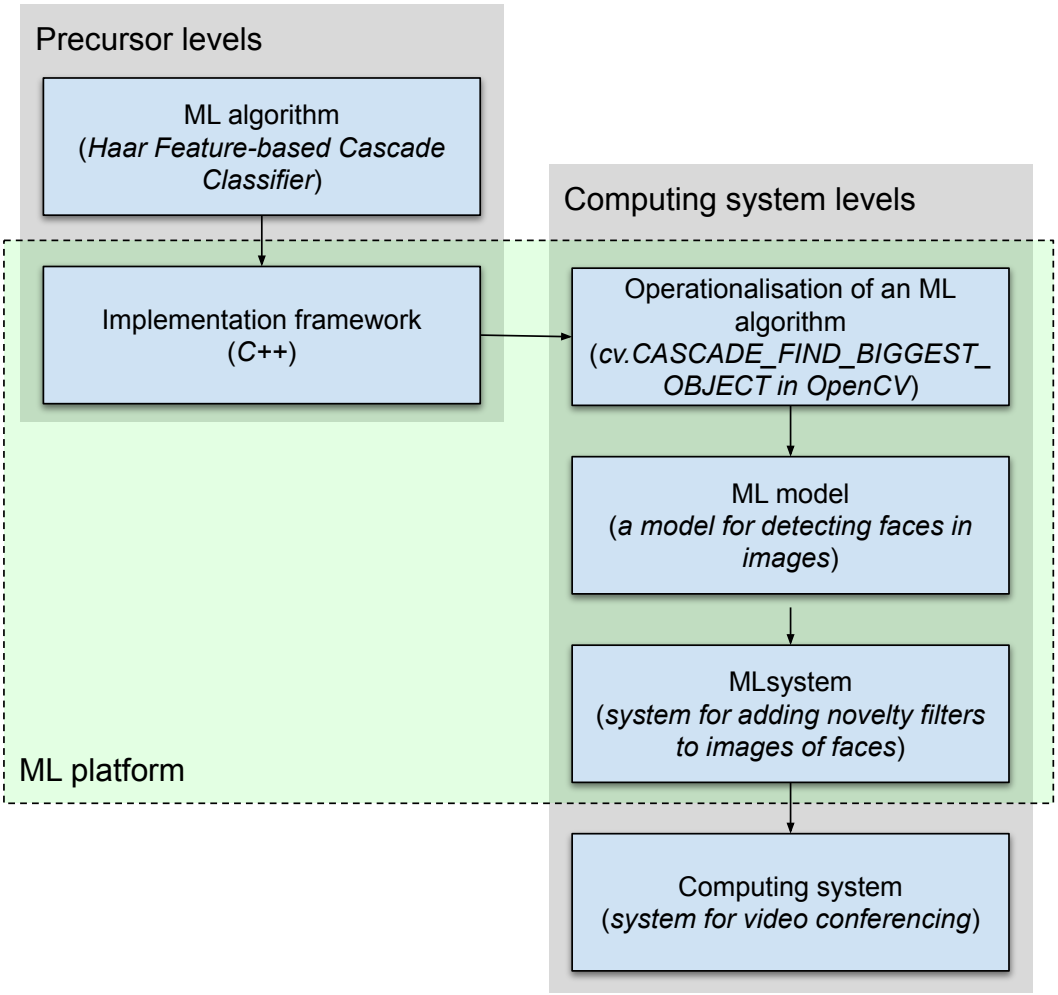

Figure 1.1: Diagram of the levels of computing systems, as we conceptualise them. In italics are an example application of these levels, based on the OpenCV library [202, 2021b]. Highlighted in light green are the levels which are incorporated into ML platforms.

that incorporate Machine Learning approaches, a *machine learning algorithm* (*ML algorithm*) and an *implementation framework* are required. For the purposes of this research project, we describe these as precursor levels of computing systems. The first precursor level, an ML algorithm, can be thought of a sequence of instructions for producing an ML model. ML algorithms are not computer code. Rather, they are mathematical and logical constructs which are *operationalised* by expressing them in computer code. This is why we consider ML algorithms to be a precursor level to computing systems. To operationalise an ML algorithm requires an *implementation framework* – a programming language designed to support engineers to write computer programs that incorporate ML algorithms and derive ML models. The implementation framework is therefore the second precursor level. In much Computer Science literature implementation frameworks are referred to as 'software platforms'; we do not follow that norm in this research proposal for risk of creating confusion between ML platforms and implementation frameworks.

1.2.1.2 Computing system levels

Within and across different implementation frameworks, practitioners may develop different approaches to operationalising a given ML algorithm, often optimised for different use cases. Additionally, some implementation frameworks may themselves include built-in operationalisations of widely used ML algorithms. In the context of this research proposal, the operationalisation of an ML algorithm on an implementation framework is the first level of a computing system. From a practical engineering perspective, the distinction between an ML algorithm and its operationalisation is often trivial, given that the practitioner is focused on applying the operationalised ML algorithm for a particular use case, not on rethinking the underlying ML algorithm itself. However, in the context of the critical perspective of this research project, this distinction is significant because the operationalisation of an ML algorithm necessitates value-laden decisions that can produce unintended consequences. For instance, because ML algorithms derive statistical models, whose outputs are expressed in probabilities, when an ML algorithm is operationalised the engineer must determine what an acceptable error rate will be for the statistical model in the context of its deployment [Barocas et al., 2021b]. If the ML model is to be used in a medical diagnosis computing system, then determining an acceptable error rate can have a direct impact on the health outcomes of those who use the system.

Once an ML algorithm has been operationalised on an implementation framework it can then be used to generate an ML model. ML models are thus the second level of computing systems relevant to this research project. An ML model is generated by inputting a training dataset into an operationalised ML algorithm. The ML model is iteratively optimised until it meets specified performance criteria – a complex training and fine tuning process, often described as both art and science by practitioners [Silipo, 2020]. At that point the ML model can be incorporated into a *machine learning system* (*ML system*), where the ML model will now be deployed. The ML system is therefore the third level of computing system relevant to this re-

search project. Distinguishing between ML algorithms, ML models, and ML systems is important because ML systems include additional components beyond ML models and operationalised ML algorithms, such as data inputting, processing, and storage components.

Machine learning systems themselves may be components in *large computing systems*, such a vehicle or a social network. These large computing systems form the fourth, and final, level of computing systems that are relevant to this research project. In practice, the distinctions between these levels are not as discrete as presented here, primarily because in most ML systems sequences of ML models are chained together, with some ML models themselves built into the implementation framework and others generated by unique operationalisations of a ML algorithm and training dataset [Lwakatare et al., 2019]. A facial recognition ML system may make use of an implementation framework's built in pre-trained ML model for detecting faces in images, and then treat the outputs of that ML model (i.e. images that are classified as containing a human face) as inputs for a custom built ML model for recognising the faces of specific individuals.

As can be seen in the diagram in Figure 1.1, we consider ML platforms to be computing systems that sit across the levels of computing systems discussed here. In Section 2.2.1 we describe in more detail how the services offered by ML platforms incorporate implementation frameworks, operationalised ML algorithms, pre-trained ML models, and systems for deploying and managing ML systems.

Finally, we note, that to avoid confusion between references to one of the levels of computing systems and the discipline of *Machine Learning*, when discussing the discipline we will use the capitalised and non-abbreviated term Machine Learning. The phrase *Machine Learning approaches* should therefore be understood to mean approaches to researching and developing computing systems adopted by researchers and practitioners in the discipline of Machine Learning. However, as the distinction between researchers and practitioners alludes to, the categories of actors relevant to the engineering of ML systems, and the terms used to describe them, can also be a source of confusion and therefore require clarification.

Category	**Level**
1. Precursor levels	(a) ML algorithm
	(b) Implementation framework
2. Computing system levels	(a) Operationalisation of an ML algorithm
	(b) ML model
	(c) ML system
	(d) Computing system

Table 1.1: Table of levels of computing systems discussed in this proposal.

1.2.2 Practitioners, stakeholders, and subjects

We use the phrase *Machine Learning practitioner* (*ML practitioner*) as a catchall term for anyone who works on the development of ML algorithms, implementation frameworks designed to support the operationalisation of ML algorithms, ML models, ML systems, or ML platforms. ML practitioners may thus be computer scientists, data scientists, software engineers, software architects, system engineers, designers, product managers, and so on. Similarly, we use the phrase *Machine Learning subjects* (*ML subjects*) as a general term for anyone who is the subject of deployed ML systems, in the sense that they are required to interact with an ML system. In most modern societies all members of the public can be considered ML subjects – certainly, all practitioners are very likely to also be ML subjects. However, in order to explore the relationships between different ML practitioners, the interactions between ML practitioners and ML platforms, and the interactions between ML subjects with ML platforms and ML systems, it is helpful to distinguish between several categories of ML practitioners and ML subjects.

As this research project is focused on how ML platforms interact with ML systems that are developed on them, and the unintended social consequences that can arise from these interactions, the most important distinction to draw within the category of ML practitioners is between those who work on developing ML platforms themselves and those who use ML platforms to develop ML systems. In this research proposal we use the term *ML platform developer* to refer to the former and *ML platform user* to refer to the latter.

The two categories of ML practitioners are complemented by two categories of stakeholders: *ML platform operators* and *ML system operators*. ML platform operators are the technology firms who run ML platforms and employ ML platform developers (as well as many other ML practitioners). ML system operators are the organisations who commission the development of ML systems, and employ ML platform users. This description of the relationship between practitioners and stakeholders is, of course, highly simplified and will be revisited and nuanced in Section 2.2.2.

Machine learning subjects also fall into two categories: *ML system users* and *ML system subjects*. ML system users are those who interact directly with the outputs of ML systems. ML system users are not ML practitioners; ML system users do not have access to the inner workings of ML systems, and may not understand their operation. Examples of ML system users include police force members who use predictive policing tools and Netflix subscribers who use Netflix's recommendation engine to select a new television series to watch. As such, ML system users is a large and heterogeneous category. The common characteristic of ML system users is that their relationship to ML systems is instrumental. ML system users use ML systems to achieve a task or make a decision. ML system users are nonetheless subjects of ML systems in the sense that they have little agency over the decision of whether or not to use an ML system to complete said task or decision. Once a police force has decided to deploy a predictive policing tool individual police force members are unlikely to be in a position to decline to use it. Similarly, a Netflix subscriber may

choose to ignore the suggestions of the Netflix recommendation engine, but they cannot avoid interacting with them.

ML system subjects are those who interact with the actions of ML system users, rather than with the ML system itself. Members of the public in a jurisdiction where the police force uses a predictive policing tool are ML system subjects of that tool. They are subjected to the outputs of the tool through the actions the police take in response to the tool's predictions of criminal activity and may not even be aware of the existence of the ML system itself. ML system subjects are one step further removed from ML systems than ML system users; ML system subjects have no access to the inner workings of ML systems, and only indirect access to the outputs of ML systems. ML systems subjects are thus also a large and heterogeneous category. The outputs of ML systems are translated by ML system users into outcomes for ML system subjects – although, as we will explore in Section 2.7 this process of translation is complex and can often be a source of unintended consequences.

Lastly, it should be noted that none of the above categories are mutually exclusive. The relevance of membership in these categories follows from the socio-technical system which is the focus of inquiry. A software engineer, for instance, may be an ML system subject of their workplace's video surveillance system and an ML system user of online search engines. During their work day they may be an ML system developer for one project and an ML platform developer for another. As such, in this research proposal, when these categories are used they will be contextually defined by the socio-technical system under discussion.

In summary:

- *ML platform operators* employ *ML platform developers* to develop ML platforms
- *ML system operators* employ *ML platform users* to develop ML systems on ML platforms
- *ML system users* interact with deployed ML systems
- *ML system subjects* interact with *ML system users*

1.2.3 Platforms, two-sided markets, infrastructure, and implementation frameworks

To the best of our knowledge, there is no accepted definition of an ML platform; the term is used primarily in the marketing materials of ML platform operators. In Section 2.2 we review the services provided by ML platforms and develop an initial description of their role in ML system development.

We note, also, that ML platforms are not the only socio-technical systems described as *platforms*. As briefly mentioned above, in Computer Science literature implementation frameworks are often referred to as 'software platforms'. Additionally, the services offered by Uber, AirBnB, Amazon, and many other large technology firms are often marketed as 'platforms', and these technology firms themselves are

sometimes referred to as 'platform businesses' [Srnicek, 2017]. Public infrastructures, such as public transit networks and the Internet, are also sometimes referred to as 'platforms' – generally when authors are seeking to highlight the instrumental function of these infrastructures in supporting social activity [Plantin and Punathambekar, 2019]. The Internet may thus be described as 'platform for retail business'. To avoid confusion, in this research proposal we use the term 'platform' only when referring to the socio-technical systems marketed by ML platform operators as ML platforms. In Section 2.3 we discuss the ways ML platforms are conceptually similar and different to other socio-technical systems which are also sometimes described as platforms.

With the terminology discussed so far we can now introduce our research questions, which are the focus of this research proposal.

1.3 Research questions

The overall focus of this research proposal is the problem of understanding how ML platform users enact ethics considerations in the development of ML systems on ML platforms. We propose to divide this problem into three phases of research, with the following research questions for each phase. In Section 4.1 we detail the research methodology we propose to adopt in answering these questions, and, in Section 4.2, we detail our rationale for dividing our research into three phases, and the research activities we propose to undertake in each phase. As we discuss further in those sections, we intend to answer these research questions for a subset of ML platforms and ML platform users, which will be defined before the activities commence.

R1 Phase one: ML platforms

- **R1a** Who are using ML platforms to develop and deploy ML systems?
- **R1b** How do ML platforms conceptualise the ML system development and deployment process and integrate ethics considerations into that process?
- **R1c** What kinds of ML systems, and ML system development, do the affordances of ML platforms support?

Our focus in phase one are ML platforms and ML platform users. In Section 2.2 we provide a preliminary analysis of how ML platforms influence ML system development and deployment, and of who may constitute ML platform users. This analysis is limited, however, by the lack of empirical data regarding ML platform services or use. As such, in phase one of our research we plan to analyse the documentation produced by ML platform operators, alongside the documentation and ML systems produced by ML platform users. Section 4.2.1 describes our research plan for phase one in detail.

R2 Phase two: enactment of ethics considerations

R2a How are ML platform users enacting ethics considerations in the development of ML systems?

R2b Are there particular fair-ML interventions that ML platform users rely on in their enactment of ethics considerations? If so, why do they rely on them?

R2c In what ways do the fair-ML interventions that ML platform users rely on, if any, shape the ML system development processes they undertake?

In phase two of our research we focus on the ways in which ML platform users enact ethics considerations during the development of ML systems. In Section 2.4 we argue that inherent in the process of ML system development are ethics considerations, which must be resolved in order for an ML system to be deployed. In Section 2.6 we introduce the concept of enactment to highlight the ways in which ML platform users, knowingly or unknowingly, resolve ethics considerations during ML system development. We review several existing studies of the enactment of ethics considerations by ML practitioners, and demonstrate that these studies do not extend to consider how use of ML platforms may influence the enactment of ethics considerations. Phase two of our research aims to address this gap through semi-structured interviews and a survey of ML platform users. Section 4.2.2 describes our research plan for this phase.

R3 Phase three: evaluation of interventions

R3a How can an fair-ML intervention for enacting ethics considerations be evaluated, in the context of ML system development and deployment on ML platforms?

In the final phase of our research we focus on how fair-ML interventions in ML system development and deployment can be evaluated. In Section 2.7 we discuss existing fair-ML interventions; these are proposals for activities or processes that ML practitioners could undertake to support their enactment of ethics considerations. We argue that fair-ML interventions are rarely evaluated in the context of ML system development in industry settings, and show that the use of ML platforms in ML system development may frustrate the effectiveness of some fair-ML interventions. Phase three of our research aims to address this through a case study of the adoption and evaluation of an fair-ML intervention in organisations where ML platform users are responsible for developing and deploying new ML systems. In Section 4.2.4 we describe our research plan for this phase, though we note that it is still in an early stage of development and the research question itself will likely need to be refined.

1.4 Thesis Proposal structure

In this chapter we briefly introduced and situated the research problem we propose to focus on in our research project. We also reviewed key terms used throughout this proposal, and provided a summary of the research questions that our project aims to address. The remainder of this proposal aims to substantiate the significance of our research problem, to provide theoretical justification for our research questions, and to detail our research design and thesis plan.

The following chapters are organised as follows: Chapter 2 presents a literature review of the fair-ML discourse; Chapter 3 introduces the theoretical frameworks that inform our research design; Chapter 4 describes our research methodology and design; and, Chapter 5 discusses how we plan to implement our research design and manage this research project.

Chapter 2

Literature Review

In the preceding introductory chapter, Chapter 1, we described the overall focus of our research proposal as being the problem of understanding how ML platform users enact ethical considerations in the development of ML systems on ML platforms. In this Literature Review we aim to demonstrate the significance of this research problem, and to show that existing discourse in the fair-ML and associated disciplines does not provide a theoretical or empirical framework for solving it. In the chapter that follows this review, Chapter 3, we introduce the conceptual framework we propose to apply to address this research problem.

2.1 Literature review structure and argument

The argument put forward in this literature review is:

I **ML platforms are multi-dimensional objects of study.**

In Section 2.2 we provide a short description of popular ML platforms, their features and use, and the aspects of ML platforms that are most relevant to this research proposal.

In Section 2.3 we consider how best to conceptualise ML platforms, and contrast ML platforms to other socio-technical systems which are also sometimes described as 'platforms' in the literature.

II **ML systems development requires navigating complex ethical issues.**

In Section 2.4 we review existing research on ethics issues associated with ML system development and deployment. We show that ML system development requires ML system operators and developers to navigate complex ethical issues, which helps explain the significance of our research focus on the enactment of ethical considerations during ML system development.

III **Conceptualising ML systems and ML platforms as socio-technical systems helps us navigate these complex ethical issues.**

In Section 2.5 we focus on the concept of *socio-technical systems*, and review its application within the fair-ML literature. We show that in this discourse the

concept of socio-technical systems has been used in analyses of the social impact of ML systems, and to support the development of interventions designed to enable ML system operators and developers to identify and mitigate ethical issues.

IV **ML practitioners navigate these complex ethical issues through processes of enactment.**

In Section 2.6 we introduce the concept of *enactment*, and review existing research on the enactment of ethical considerations during ML system development.

In Section 2.7 we review existing proposals for interventions in the research, design, and management of ML systems from the fair-ML literature, which are intended to support more substantive enactment of ethical considerations.

V **The tools used to support ML system development can influence the manner and meaning of the enactment of ethical considerations.**

In Section 2.8 we return our focus to ML platforms and consider how the key features of ML platforms described in Section 2.2 may influence proposed fair-ML interventions in the enactment of ethical considerations during ML system development.

2.2 What is an ML platform?

In this section we introduce ML platforms, as the primary object of study in this research proposal. We provide an overview of four key aspects of ML platforms. These aspects will be touch points throughout this literature review, as we consider how the use of ML platforms during ML system development may influence the ways ethics issues manifest in ML systems. In the following section, Section 2.3, we consider how other socio-technical systems, which are also sometimes termed platforms, have been conceptualised in the literature, and explore how these conceptualisations can be applied to the aspects of ML platforms introduced here.

We note that in our future research it will be necessary to develop a detailed technical description of how ML platforms operate, their key features, and points of convergence and divergence across different ML platforms. In this section, for the purposes of enabling consideration of the merits of our research proposal, we focus on only four aspects of ML platforms, which we argue are common across all the large, commercially available ML platforms, and which we see as particularly pertinent to the existing discourse on ethics associated with ML system development and deployment. The four aspects of ML platforms focused on in this section are their:

- *Centralising effect*: ML platforms centralise ML system development in a single suite of cloud-based tools and services;

- *Scaling effect*: ML platforms enable ML systems to rapidly scale and be deployed at a global level;
- *Refocusing effect*: ML platforms refocus the efforts of ML practitioners towards ML system design and deployment and away from ML pipeline development and ML system maintenance; and
- *Decentralising effect*: ML platforms decentralise the capacity to develop and deploy ML systems, and disaggregate the components of ML systems.

2.2.1 Brief description of ML platforms

Although ML platforms are computing systems, they do not fit neatly within any of the levels of computing systems introduced in Section 1.2.1. The primary purpose of an ML platform is to support the development and deployment of ML systems. All of the major technology firms offer ML platforms to practitioners. These include: Microsoft's Azure AI [Microsoft, a], Google's Vertex AI [Google, a], Amazon's AWS AI services [Amazon, b], and IBM's SPSS platform [IBM]. We note, however, that the term ML platform is largely a marketing term rather than an established term within the Machine Learning discourse, and as such it is possible that very different systems or services may be united under the label ML platform. That said, while the full breadth services offered by each ML platform listed above are different, there is still significant overlap in their core services. Core services offered by most of the ML platforms listed here include:

- automation of routine Machine Learning tasks, such as data processing;
- dataset and ML model management;
- low-code or no-code ML model training environments;
- access to implementation frameworks with state-of-the-art operationalisations of ML algorithms;
- on-demand access to servers and compute capacity for training ML models and maintaining ML systems;
- end-to-end ML system solutions addressing specific industries or business needs; and
- marketplaces for user-developed pre-trained ML models or user-curated datasets for ML model training.

To offer these services, an ML platform must incorporate an implementation framework, and operationalisations of ML algorithms and pre-trained ML models into a user interface and cloud-based development environment, together with the substantial physical computing hardware required to support ML model training

and deployment. As such, ML platforms could be considered large computing systems within the multi-level framework outlined in Section 1.2.1. At the same time, from the perspective of ML practitioners, ML platforms are more comparable to the level of implementation frameworks, as their primary purpose is to support the practitioner to develop an ML system that will be integrated into a large computing system. This is the *centralising effect* of ML platforms: ML platforms bring together, in one assemblage, the technical tools and computing hardware required to develop, deploy, and maintain ML systems.

The significance of the centralising effect of ML platforms lies in the complexity and cost associated with developing ML systems. In the Machine Learning discipline, the metaphor of a pipeline is used to describe the sequence of steps and feedback loops which an ML practitioner may follow to operationalise an ML algorithm, train an ML model, and develop an ML system. As such, machine learning pipelines (ML pipelines) run parallel to the levels of computing systems outlined in Section 1.2.1; a practitioner's ML pipeline is their approach to uniting all the levels of computing systems into a coherent engineering process [Polyzotis et al., 2017]. ML systems not only centralise the technical features of an ML pipeline, they also centralise responsibility for defining a coherent engineering process – an ML platform presents an ML practitioner with a pre-defined engineering process to follow in order to develop ML systems. Thus the centralising effect of ML platforms can be considered isomorphic, in that it contributes to the convergence of ML pipeline structure and processes across ML practitioners [Caplan and Boyd, 2018].

The *scaling effect* of ML platforms is also revealed from the above list of services. The development, deployment and ongoing management of ML systems is a complex engineering process, with requirements that differ from traditional software engineering [Lwakatare et al., 2019]. ML systems, when deployed in commercial settings, often incorporate cascading ML models that interact with each other and ongoing processes of experimentation and optimisation, and must meet strict performance requirements in terms of processing time [Lwakatare et al., 2019]. The services offered by ML platforms attempt to address these challenges, and in doing so reduce the commercial cost of deploying and taking to scale ML systems. Further, the capacity of the largest ML platforms to offer platform users near global access to server centers and on-demand access to computing power, enables ML systems to scale across geographies. As such, the scaling effect of ML platforms contributes to the increasing ubiquity of ML systems [Lwakatare et al., 2020], and may also influence the pattern of ML system spread, which can involve the refinement of an ML system in one geography before its deployment across many others (see [Li et al., 2017] for a description of this pattern of scaling in the ML systems used by the ride-sharing service Uber).

2.2.2 Brief description of ML platform users

In Section 1.2.2 we described the ML practitioners who use ML platforms to develop ML systems as ML platform users. To the best of our knowledge, there is no data

in the public domain detailing who are ML platform users. Indeed, this is a gap in the literature that our proposed research aims to help address. As such, our current understanding of who are using ML platforms to develop ML systems is largely derived from informal conversations with ML practitioners and the user groups implicit in the marketing materials of ML platform operators. In our research design, described in Section 4.1, we propose to undertake a systematic analysis of the marketing materials and technical documentation produced by ML platform operators, with the objective of more rigorously defining ML platform user groups. From our brief preliminary review, we propose that four categories of ML platform user groups may be discerned: ML practitioners in large technology firms, ML practitioners in technology consulting firms, ML practitioners in organisations seeking to develop and deploy ML systems, and ML practitioners working in startup firms or small organisations.

As discussed above, Google, Amazon, Microsoft and IBM each operate ML platforms. The marketing materials produced by these firms indicate that ML practitioners working within them are users of their ML platforms. Google's ML platform, Vertex AI, thus markets itself with the claim, "*Build with the groundbreaking ML tools that power Google*" [Google, a]. Similarly, Microsoft's ML platform is marketed with the claim, "*Use the same proven AI services that power AI capabilities in Xbox, HoloLens, and Microsoft Teams*" [Microsoft, a]. Several other large technology firms also operate ML platforms, which are used primarily by in-house ML practitioners (see [Li et al., 2017] for a technical overview of Uber's ML platform and [Hazelwood et al., 2018] for a similar overview of a Facebook ML platform).

The marketing materials of ML platform operators also indicate that ML practitioners working in technology consulting firms are significant ML platform users. Google and Amazon, for example, operate technology partner networks, enabling technology consulting firms to be accredited as 'trusted' ML platform partners [[Google, b; Amazon, d]]. Google and Amazon direct organisations interested in developing ML systems to their technology partners, who can then provide advice and technical support to enable interested organisations to use the relevant ML platform to develop and deploy the desired ML system. Whilst evidence regarding the use of technology consulting firms to develop ML systems is lacking, it appears likely that public sector organisations will often employ the services of a technology consulting firm to support their development of ML systems. Certainly, this is implicit in advice prepared for public sector organisations on procuring ML systems [Goldenfein, 2019], and is a feature of several high-profile instances of problematic deployments of ML systems [Eubanks, 2018].

The use of ML platforms by ML practitioners in large technology firms and ML practitioners in technology consulting firms highlights the *refocusing effect* of ML platforms. ML platforms shift the work of ML platform users by automating or rendering trivial certain aspects of ML system development. In doing so, ML platforms refocus the work of ML platform users towards ML system design and deployment. Uber's ML platform, for example, provides ML practitioners at Uber with access to a central repository of up-to-date datasets, a system for managing ML model iterations

and versions, and automated management of deployed ML models [Li et al., 2017]. Together, these features enable ML practitioners at Uber to focus on developing and deploying new ML systems [Li et al., 2017, p.1].

The marketing materials of ML platform operators also advertise the potential for organisations to develop and deploy ML systems without the support of a technology consulting firm. As such, we assume that ML practitioners in organisations seeking to develop and deploy ML systems are an additional category of ML platform users. ML practitioners in organisations seeking to develop and deploy ML systems may have similar expertise to those working for technology consulting firms, or in-house at ML platform operators. However, it may also be the case that some ML practitioners in organisations seeking to develop and deploy ML systems have substantially less training or expertise than those working in technology consulting firms or ML platform operators. Microsoft's ML platform, for example, offers access to several ML systems, including a customer engagement chat bot and document data extraction service, that can be deployed by organisations with little to no technical development [Microsoft, b]. These services reflect another way in which the refocusing effect operates: ML platforms expand the boundaries of who may be considered an ML practitioner by creating user interfaces and ML system development environments that are accessible to people without expert training in data science or software engineering. In this way, ML platforms refocus the scope of who might be considered an ML practitioner. Further, we note that several technology firms offer free online courses designed to train people with no experience in software engineering or data science in using their ML platforms [Amazon, c; Microsoft, c]. Whilst it is unclear whether members of any of the user groups we describe in this section engage in such training courses, the existence of such courses potentially also further enhances the refocusing effect of ML platforms.

Finally, the marketing materials of ML platform operators and our informal discussions with ML practitioners indicate that ML practitioners working in startup firms or small organisations are also likely to be ML platform users. Amazon's marketing materials provide fifty case studies of startup firms who have used their ML platform, spanning healthcare, electronics, financial services, marketing, hospitality, education, and gaming industries [Amazon, e]. The use of ML platforms by ML practitioners in startup firms and small organisations highlights the *decentralising effect* of ML platforms. By radically reducing the resources required to develop and deploy ML systems, ML platforms decentralise the capacity do so, even as they centralise the underlying infrastructure. One of the primary ways in which ML platforms reduce the resources required to undertake ML system development is by providing ML platform users access to centrally managed data storage and computing capacity, thereby reducing the need for ML platform users to obtain specialist hardware. Central management of data storage and compute capacity, however, contributes to the decentralising effect by enabling the components of ML systems to be disaggregated across multiple sites, organisations, and geographies (see [Joler and Crawford, 2018] for an example of an effort to trace all of the disaggregated components of a computing system that incorporates ML systems).

The user groups outlined above are preliminary, and serve only to indicate that ML platform users are likely an heterogeneous amalgam of ML practitioners. Most relevantly, for this research proposal, the four user groups outlined above reflect different social contexts in which ML systems are developed: startups may have different funding arrangements to established organisations; large technology firms may be under far greater public scrutiny than technology consulting firms. Social contexts may also vary considerably within the user groups: public sector organisations may be under different regulatory obligations to private sector organisations; large technology firms operating in Europe may similarly be under different regulatory obligations to their peers operating in the United States. These different social contexts, we assume, are likely to influence how ML platforms are used by ML practitioners in each user group.

In this section we introduced four aspects of ML platforms. By describing the services offered by ML platforms we introduced the centralising effect and scaling effect of ML platforms. By examining who use ML platforms to develop ML systems we introduced their refocusing effect and decentralising effect. In the following section, Section 2.3 we consider how existing literature conceptualises socio-technical systems that share some of these aspects.

2.3 Platforms as two-sided markets, infrastructure, and implementation frameworks

In the previous section, Section 2.2, we outlined the services offered by ML platforms, and highlighted four key aspects of ML platforms, which we will revisit throughout this review. In this section we expand our analysis of platforms by considering how other socio-technical systems which share the platform label have been conceptualised in the literature. We highlight three conceptualisations of platforms – as two-sided markets, infrastructure, and implementation frameworks – and reflect on how each conceptualisation can be applied to our understanding of ML platforms. We do not aim to provide comprehensive overviews of the literature on two-sided markets, infrastructure, and implementation frameworks. Rather, for each conceptualisation we introduce a theoretical framework from the literature and apply this framework to our understanding of ML platforms. We note that these conceptualisations are not mutually exclusive [Plantin et al., 2018], and argue that each of these conceptualisations are relevant, but imperfect, descriptions of ML platforms. The literature focused on each conceptualisation is similarly relevant, but insufficient, for our research focus on the enactment of ethics considerations by ML practitioners using ML platforms.

2.3.1 Two-sided markets

Two-sided markets are, in general terms, markets in which 'platforms' compete to facilitate interactions between two different user groups [Rochet and Tirole, 2004]. As such, two-sided markets are a subset of the broader category of many-sided markets, in which the interactions between more than two user groups are mediated by the platform [Rochet and Tirole, 2004]. Whilst two-sided markets are primarily set up as sites of economic transaction, in their operationalisation two-sided markets can influence and mediate the economic and social behaviour of market actors [Islind et al., 2021]. For an economic model of competition between 'platforms' in two-sided markets see [Rochet and Tirole, 2003]. In Rochet and Triole's work, and the subsequent economics literature on two-sided markets, 'platform' is used as a shorthand label for firms who operate socio-technical systems which enable distinct user groups to transact [Evans et al., 2011]. See Sanchez-Cartas and León for a survey of this literature [2021]. In fair-ML literature, these firms themselves are sometimes referred to as 'two-sided markets' (e.g. see [Patro et al., 2020]). As we discuss below, this commingling of terms reflects one of the fundamental characteristics of two-sided markets, which is that network effects can enable firms operating platforms in two-sided markets to reach very large scale, such that the firm and the market may become synonymous, as in the case of Uber and the two-sided market for private transport. In the terminology we introduced in Section 1.2.2, such firms correspond to platform operators.

Two-sided markets have long been an object of study in economics (e.g. see [Demange and Gale, 1985]), although they gained particular prominence in the early 2000s in the context of a series of antitrust cases against the credit card networks of Visa and MasterCard [Roson, 2009]. Computer game systems, television networks, and card payment systems have all been conceptualised as two-sided markets, in which platform operators compete to be the dominant facilitator of interactions between user groups [Rochet and Tirole, 2004]. More recently, web-based businesses, software platforms (implementation frameworks in our terminology), mobile phone devices, and social networks have also been studied as two-sided markets [Evans et al., 2011].

ML platforms can be thought of as two-sided markets, which facilitate interactions between different groups of ML practitioners engaged in the production of ML systems. The firm Hugging Face operates a web-based ML platform (hugginface.co) where ML system developers can upload trained ML models and training datasets, which can then be downloaded and integrated into ML systems by other ML system developers [HuggingFace]. Whilst the ML platforms operated by Microsoft, Amazon, and Google do not (currently) allow independent ML system developers to upload trained ML models, they can nonetheless still be conceptualised as two-sided markets, where ML system developers working *within* the formal boundaries of these technology firms interact with ML system developers working outside them.

Nick Srnicek argues that contemporary corporate strategies and organisational structures of technology firms have coalesced into 'platform capitalism', whereby

technology firms seek to develop their offerings and services as platforms in two, or many-sided, markets [2017]. Srnick argues that the emergence of platform capitalism is a response to the transition to a computer-mediated economy, where low-cost and ubiquitous data generation is possible. Traditional business models and organisational structures were ill equipped to extract value from this transition, and so under the pressure of competition 'platform capitalism' emerged as a dominant corporate and technology strategy [Srnicek, 2017, p.42]. Srnicek defines platforms as " *digital infrastructures that enable two or more groups to interact*" [2017, p.43], reflecting a conceptualisation of platforms as two-sided markets. Organisations that incorporate two-sided markets include the major technology firms (Google, Facebook, Amazon), technology startups that have reached near global scale (Uber, AirBnB), and large corporations in a variety of industries (GE, Siemens, John Deere, Monsanto) [Srnicek, 2017]. Srnicek's argument is an illustration of the actors and relations that conceptualising platforms as two-sided markets helps bring to the fore.

Srnicek highlights four characteristics of two-sided markets. Each of these characteristics present ethics issues. First, compared to traditional organisational structures, digitally-mediated two-sided markets are far better placed for enabling technology firms – platform operators – to collect and extract value from data. Digitally-mediated two-sided markets are both intermediaries between users and a site of user activity. The Uber application connects drivers and passengers, processes their economic transaction, and coordinates the execution of that transaction (i.e. provides route directions to enable the driver to transport the passenger to their destination) – and converts all of this activity into data [Srnicek, 2017]. The politics and logic of value extraction from the data accumulated by platform operators has been described as 'surveillance capitalism' [Zuboff, 2020], and is closely associated with privacy and agency issues. The centralising effect of ML platforms enables ML platform operators to collect enormous volumes of data regarding ML system use cases and every stage of the ML system development pipeline. It is unclear whether or how ML platform operators are extracting value from this data, although given the overlap between technology firms who have adopted platform capitalism strategies and who operate ML platforms, it seems likely that extraction strategies and issues are relevant to ML platforms.

Second, two-sided markets are subject to network effects: the more drivers and passengers using Uber, the more useful Uber's two-sided market is for both groups [Srnicek, 2017]. Srnicek argues that two-sided markets therefore have monopolistic tendencies, as users will gravitate towards the most popular markets. Additionally, as digitally-mediated two-sided markets have limited physical infrastructure, outside of servers, they are capable of scaling extremely quickly, both in terms of the size of their markets and in terms of the products offered on their market. It should be noted, however, that the relation between network effects and monopolistic dynamics is subject to ongoing debate. Network effects may give incumbent platform operators a significant advantage over new entrants, but there are few true monopolies [Roson, 2009]. Uber, for instance, competes with Lyft and Didi. In the context of ML platforms, network effects may augment the scaling effect we discussed in

Section 2.2.

Third, to maintain positive network effects (i.e. to keep attracting users), organisations operating two-sided markets enact complex cross-subsidisation strategies [Srnicek, 2017; Rochet and Tirole, 2004]. Supply and demand on both sides of the market need to be continually balanced by subsidising some activities, whilst increasing the price of others. New users may receive heavy subsidies to induce them to enter the two-sided market, while existing users who are unlikely to shift to a competitor's market due to network effects may see their activities priced more highly. Further, operators of two-sided markets will often offer users many different services, even access to many different two-sided markets, in an effort to further cross-subsidise their operations and maximise network effects. Uber, for example, operates a transport market and a food delivery market. It is beyond the scope of this research proposal to consider the economic models underpinning ML platforms. However, we note that, on the surface at least, the pricing structures of ML platforms do appear to reflect complex cross-subsidisation strategies. Several ML platforms offer new users free access to basic services [Microsoft, b; HuggingFace], and, for paid users, bundle services together in ways that may be consistent with cross-subsidisation. Additionally, ML platforms must balance supply of, and demand for, compute capacity [Li et al., 2017; Hazelwood et al., 2018].

Fourth, whilst two-sided markets may present themselves as neutral intermediaries for economic activity, they enact a politics [Srnicek, 2017]. The organisations operating two-sided markets set the rules of the market, and govern the market in the interests of their shareholders – potentially displacing, in part, the role of state regulators [Kenney et al., 2019]. The rise of AirBnB, for example, as a two-sided market for accommodation, has significantly reduced the scope of regulations requiring hotels to offer accessible accommodation. Previously such regulations covered a large portion of the accommodation industry. Today, they apply to a far smaller portion because so much of the industry has been subsumed by AirBnB's two-sided market, which is not covered by hotel regulations [Boxall et al., 2018]. The market rules are often structured so as to support users to develop new forms of interaction or new services that both make use of the data generated by the two-sided market and, by expanding the scope of the market to new activities, generate more data for the organisation operating the market. Uber's API, for instance, enables third-party access driver data, which can be used to develop new services for drivers, such as loyalty programs [Uber, 2021]. This aspect of two-sided markets has prompted the emergence of Platform Governance as a new discipline, focused on understanding and intervening in the politics enacted by the operators of two-sided markets [Gorwa, 2019]. The centralising and scaling effects of ML platforms position ML platform operators in a similar way to the operators of two-sided markets; ML platform operators effectively set the rules for how ML platform users can use their platforms.

Whilst the four characteristics of two-sided markets highlighted by Srnicek are relevant to ML platforms, they do not account for the refocusing and decentralising effects of ML platforms we introduced in Section 2.2. This may reflect a limitation in the conceptualisation of two-sided markets in the literature: conceptualisations

of two-sided markets often treat market actors as homogeneous categories of users, failing to account for how users' social context may mediate their engagement with a two-sided market [Anable, 2018].

2.3.2 Infrastructure

Whilst Srnicek defines platforms as "*digital infrastructures that enable two or more groups to interact*", his primary concern is with the second half of that definition: the way two-sided markets act as intermediaries between different groups [2017, p.43]. Other authors, who are concerned with the way 'platforms' become essential and widely participated in socio-technical systems, explore in more detail the way 'platforms' can be conceptualised as 'digital infrastructure' [Plantin et al., 2018].

Infrastructure is itself a slippery concept. Infrastructures are often thought of as technological systems, but, as Brian Larkin observes, they may also be financial instruments, or management structures, or social networks, or embodied norms of behaviour [2013, p.338]. Susan Leigh Star argues that the heterogeneity of systems that may be considered infrastructure reflects the fundamentally relational nature of infrastructure [1999]. What one considers infrastructure depends on one's local context and relation to the system in question. Star offers the example of a staircase, which one person might understand as a benign piece of the urban infrastructure and another might experience as an obstacle [1999, p.380]. In his review of anthropological practices for studying infrastructure, Larkin identifies four aspects to thinking about socio-technical systems as infrastructure [2013], each of which can be applied to ML platforms. Star, who has also written extensively about the practice of studying infrastructural systems, identifies nine properties of infrastructure [1999]. Reading Larkin and Star side by side helps reveal the breadth of possibilities that conceptualising ML platforms as infrastructure opens up.

As material objects, Larkin considers infrastructures to be "*built networks that facilitate the flow of goods, people, or ideas and allow for their exchange over space*" [2013, p.328]. At the same time, infrastructures are also systems that support the functioning of other objects. This duality of infrastructure, as both material object and system for supporting other objects, is the first aspect of thinking about infrastructures described by Larkin. ML platforms, as computing systems themselves, and systems for supporting the development and deployment of ML systems, clearly embody this duality. In our daily lives when we interact with infrastructures the material objects of the infrastructure itself may recede into the background. As Larkin observes, we experience light, not electricity; hot water, not plumbing [2013, p.329]. Star describes this characteristic of infrastructure as 'transparency' [1999]. The material objects of infrastructure appear transparent to users in the sense that for users, the tasks associated with infrastructure seem easy and straightforward. It should be noted that in Star's understanding of infrastructure as relational, transparency is a characteristic that users experience when their relationship to a system becomes infrastructural – it is not a characteristic inherent in technical components of a system, but rather a product of the relationship between users (i.e. social components) and technical

components of infrastructural systems.

When we focus on the material objects of infrastructure, we realise that infrastructures exist in circular webs of relations with other infrastructures [Larkin, 2013]. We can think of computers as relying on the infrastructure of the electricity grid to function, and can also think of modern electricity grids as relying on the infrastructure of computing for their functioning [Larkin, 2013]. Similarly, Star describes infrastructures as embedded within other infrastructures, social arrangements, and technologies [1999]. Star additionally describes infrastructures as "*built on an installed base*", highlighting how infrastructures often follow paths laid down by preceding infrastructures [1999, p.382]. The poles and wires of the electricity grid, for example, often follow the structure of the road network. ML platforms, similarly, have followed the path laid down by the preceding infrastructure of cloud-computing. For researchers of infrastructure a critical theoretical and methodological challenge is therefore the question of where to draw the boundaries around an infrastructural system. Researchers, argues Larkin, must recognise that infrastructures can exist simultaneously at many different levels, and that the act of defining an infrastructure has political and epistemological implications [2013, p.330]. Star, similarly, observes that infrastructures often have reach beyond a single event, point in time, or physical site [1999]. The decentralising effect of ML platforms appears to be consistent with this understanding of infrastructure, as the technological components of ML platforms are geographically dispersed, and dislocated from social components, which themselves may also be dispersed. Temporally, too, ML platforms operate across many different scales, from data storage services that may retain training datasets for many years, through to near instantaneous feedback loops between input data and live ML models. Drawing the boundaries of ML platforms is a significant research challenge, which we will need to carefully consider in future work.

The second aspect of thinking about infrastructures highlighted by Larkin is their close connection to our understanding of modernity [2013]. Among the objects that infrastructures move and support are people and ideas. In doing so, infrastructures create opportunities for social progress or change, and are often closely tied to notions of modernity and the future. Modern infrastructures – the Internet, railways, electricity, sewage – come to symbolise modernity itself [Larkin, 2013, p.332]. This symbolic function is no accident; those who are invested in infrastructures have an interest in positioning infrastructures as enablers of modernity. Researchers of infrastructure, therefore, must study the technical functions of infrastructure, the futures that these technical functions enable and disable, and the societal dynamics that infrastructures embody [Larkin, 2013]. Star nuances Larkin's description of the connection between modernity and infrastructure. Star describes infrastructures as "*"learned as part of membership"* [1999, p.381]. Members of a group learn how to use the infrastructure of the group as part of their membership. Shared use of infrastructure reinforces group cohesion. Non-members, meanwhile, will encounter the infrastructure as something they need to learn to use in order to integrate into the group. As such, whilst modern infrastructures may symbolise modernity, the more significant relation is that one's *use* of modern infrastructures indicates membership

in the project of modernity. This conceptualisation of the relationship between infrastructures and group membership appears to align closely to the relationship between ML practitioners and ML platforms. As the refocusing effect of ML platforms transforms the role of ML practitioners, fluency in using ML platforms to develop ML systems may become part of what defines membership in the ML practitioner group. 'Modern' ML system development may become synonymous with the use of ML platforms, implying that understanding the ways in which ML platforms imagine and support ML system development may help us project future directions for the field of Machine Learning more generally.

The third aspect of thinking about infrastructures highlighted by Larkin is their 'poetic' function [2013]. Larkin draws on linguist Roman Jakobson's concept of poetics (see [1960]), which holds that in some speech acts the palpable qualities of speech (roughly, sound patterns) have primary importance over representational or substantive qualities (meaning). Infrastructures, argues Larkin, can have a poetic function, which is not reflected in the declared intention of those building infrastructural systems or the technical function of such systems [2013]. Some ML platforms, for example, incorporate tools that enable ML platform users to assess the performance of ML models against formal definitions of fairness (e.g. [Bird et al., 2020; Amazon, a]). The declared intention behind these tools is to support platform users to mitigate potential social harms. It may be, however, that these tools serve a poetic function: they enable ML platform operators to claim they are fulfilling their social responsibilities. Infrastructures enact the politics of the institutions who develop them, and can come to represent or reflect key characteristics of their associated institutions. Researchers of infrastructure, then, must take seriously the aesthetic aspects of infrastructure. Larkin's description of the poetics of infrastructure mirrors Jenna Burrell's critique of blithe descriptions of algorithms as 'opaque', which ignore the ways the appearance of opaqueness in ML systems can reflect the politics of the institutions who operate them [2016]. In our research, we intend to consider the poetics of ML platforms by critically evaluating how ML platform operators document and market their platform functions, and contrasting this with the actual use of ML platforms by practitioners. ML platforms, we assume, may reflect the politics of ML platform operators and through the centralising, scaling, and refocusing effects discussed earlier enforce these politics on ML platform users and, by extension, on ML systems.

For Larkin, the aesthetic aspects of infrastructure include the way infrastructures may at times appear invisible [2013]. Larkin takes issue here with Star's description of infrastructures as 'invisible'. To Larkin, the various aspects of infrastructure have a wide range of visibility – the claim that infrastructures are invisible can only ever be partially valid. Star's perspective, based on her understanding of infrastructures as relational, is that for the users of an infrastructure, when that infrastructure is working normally, the infrastructure itself will appear invisible to users. Star describes this characteristic of infrastructure as "*becoming visible upon breakdown*" [1999, p.382]. The invisibility of infrastructure, in Star's analysis, is a product of the way infrastructures are transparent to use. Infrastructural systems standardise interactions

between the system, users, and other infrastructures, enabling these interactions to become transparent, which then causes the system itself to appear invisible to users, at least when it is functioning properly. Our everyday experience of telecommunications infrastructure reflects this. When telecommunications networks are working as intended they make the process of connecting and communicating with another person transparent and the infrastructure supporting the connection largely unnoticeable; but, when the network experiences an outage, our connections suddenly become difficult to establish or troubleshoot, and the network infrastructure becomes unavoidably noticeable. Perhaps missing in Star's analysis of invisibility, however, is an analysis of the relationship between infrastructure breakdowns and infrastructure use. In practice, whilst telecommunications networks may work as intended most of the time, most regular users of the telecommunications network (at least in Australia) will have many experiences of the network failing. As such, even when the network is functioning, it is unlikely to be experienced as invisible to users.

As infrastructural systems, ML platforms explicitly standardise interactions between ML platform users, ML system components, and ML system users. The HuggingFace platform, for example, standardises how ML models and training datasets are shared between platform users – this standardisation is one of the attractions of using the platform [HuggingFace]. Larkin and Star's debate on invisibility is thus relevant to our understanding of how ML platforms may influence the ethics issues associated with ML systems: by standardising interactions, ML platforms may enable some ethics issues to become transparent (i.e. invisible) to ML platform users. For instance, the ease with which ML platforms enable ML platform users to deploy pre-trained ML models may render transparent ethics issues associated with the translation of an ML model from one social context to another. At the same time, ML platforms may also heighten the visibility of some ethics issues to ML platform users. For example, the tools for measuring the fairness of ML models which we discussed early, could help ML platform users more easily identify and act on issues of bias in their training datasets [Bird et al., 2020]. As such, in our research we intend to explore how the design of standardised interactions by ML platform operators may influence the enactment of ethical considerations by ML platform users.

The final aspect of thinking about infrastructures highlighted by Larkin is their relationship to our everyday embodied experience [2013]. Infrastructures produce the ambient conditions that inform our embodied experiences. In doing so, infrastructures create our (literal) sense of modernity. See [Shove, 2003] for a description of how the infrastructures of air-conditioning and modern bathrooms shape our embodied understanding of temperature and cleanliness. As Shove does, Larkin argues that researchers of infrastructure need to study the material and routine interactions between infrastructures and people, and the way these interactions shape societal values and shared concepts. Star similarly describes how infrastructures relate to *conventions of practice* [1999, p.381]. Infrastructure is shaped by the conventions of those who use it, while the infrastructure also shapes those conventions. This understanding of the co-evolution of infrastructures and conventions of practice helps nuance our description of the four aspects of ML platforms. Whilst we have de-

scribed centralising, scaling, refocusing, and decentralising as effects of ML platform use that originate with the design, structure and services of ML platforms, in our research it will be important for us to consider whether it may be more appropriate to conceptualise these effects as the product of the co-evolution of conventions of ML system development by ML platforms and ML platform users.

Star notes one additional characteristic of infrastructures that does not fit neatly into Larkin's four aspects of thinking about infrastructure. Infrastructures are fixed in modular increments [1999, p.382]. Whilst infrastructures may be presented as coherent, hierarchical structures, they are rarely built or managed in this way. This characteristic highlights a limitation with Larkin's description of infrastructures: Larkin largely focuses on describing aspects of thinking about socio-technical systems that have become infrastructure, which reflects his interest in studying extant infrastructures. This leaves open, however, the question of how to think about the process by which socio-technical systems become infrastructures. This question is particularly salient in the context of ML platforms, as these systems are not fixed. ML platform operators regularly make new services and features available to ML platform users. At the same time, new features must integrate with, and build on top of, existing features – in this sense, the development of ML platforms follows a pattern of path dependence, whereby the future growth and development of a platform is constrained by past decisions regarding its design [Bauer and Herder, 2009]. In future work, we intend to draw on the research of Jean-Christophe Plantin and his co-authors, who have explored the question of how 'platforms' become infrastructure [Plantin et al., 2018; Plantin and Punathambekar, 2019; Plantin and de Seta, 2019]. K. Sabeel Rahman has also sought to understand the regulatory implications of privately owned 'platforms' scaling into widely relied on infrastructure [2018], and Anne Helmond has studied the process by which Facebook has incrementally transformed into critical media infrastructure [2019; 2015].

Larkin does highlight how systems thinkers have conceptualised the growth and scaling of infrastructures, drawing on the work of T.P. Hughes [2013, p.330]. Hughes describes how 'large-scale technical systems', massive networks of infrastructure, manifest [1987]. Infrastructures begin as independent technologies with varying standards (i.e. capacity to interact). Technologies become infrastructures when one technological system becomes dominant, or when independent systems become a network with shared standards. In this context, Hughes understands 'technologies' broadly – a financial management system, a design process, as well as physical objects, are all technologies. A key process, then, is one of adaptation and translation: a technical system starts in one social context, and as it grows moves into others, with different conditions, and so must be adapted and translated for that new context. Translation may be technical or social, and although it may appear neutral, is also often political. In our future work, we intend to consider whether Hughes' understanding of how infrastructures scale may be relevant to the emergence of ML platforms as infrastructure for ML system development, and the scaling effect of ML platforms on ML systems.

2.3.3 Implementation frameworks

In Section 1.2.1 we introduced the concept of *implementation frameworks*, which we described as programming languages designed to support engineers to write computer programs. We noted that ML algorithms are operationalised *on* implementation frameworks. Implementation frameworks, then, can also be conceptualised as platforms. Reflecting this conceptualisation, Marc Andreessen, founder of Netscape, in a blog post analysing Facebook's Application Programming Interfaces (APIs), popularised a definition of a platform as a "*system that can be reprogrammed and therefore customized by outside developers – users – and in that way, adapted to countless needs and niches that the platform's original developers could not have possibly contemplated, much less had time to accommodate*" [2007]. Conceptualisation platforms as implementation frameworks enables authors to focus on the relations between technical components of platforms and overall platform dynamics. The field of Platform Studies, in particular, focuses on how implementation frameworks designed to support creative production are connected to creative works [Bogost and Montfort, 2007]. Computer game systems have been a primary object of study in Platform Studies (e.g. see [Montfort and Bogost, 2009]).

ML platforms can be conceptualised as implementation frameworks that support the production of ML systems – indeed, this is often how ML platforms are presented in computer science literature (e.g. see [Baylor et al., 2017; Li et al., 2017]). Adrian Mackenzie reveals the analytical value of this conceptualisation in his study of the Facebook application programming interface (API) [2019]. Mackenzie argues that Facebook is an implementation framework in the process of becoming infrastructure. Relevantly, he considers how the technical components of Facebook, in particular its API, are contributing to this transformation. Through analysis of Facebook's API documentation and its public code repository, Mackenzie explores how the programmability – the ability to be customised and expanded – of Facebook's implementation framework has changed over time, shifting towards enabling programming activities focused on predicting the behaviour of Facebook users. Mackenzie considers how this shift towards prediction at the level of Facebook's implementation framework configures the form of infrastructure Facebook is becoming. In future work we intend to consider whether Mackenzie's method of analysis can be applied to the technical components of ML platforms, and whether doing so may reveal how ML platforms seek to configure the form of ML systems developed on them. In other words, we intend to consider whether conceptualising ML platforms as implementation frameworks can help reveal the mechanisms by which the four effects of ML platforms we have introduced operate.

In this section we have argued that our study of ML platforms may be enriched by conceptualising them as two-sided markets, infrastructure, and implementation frameworks. At the same time, we have sought to demonstrate that, by themselves, each of these conceptualisations can offer only an incomplete understanding of the role of ML platforms in supporting ML system development. In the following sec-

tion, Section 2.4, we turn our attention to ML systems themselves, and introduce three ethics issues associated with ML system development.

2.4 Ethics issues associated with ML systems

In the previous two sections, Section 2.2 and Section 2.3, we described ML platforms in some detail. This literature review is primarily concerned, however, not with ML platforms as standalone systems, but rather with their relation to the social outcomes of ML system development and deployment. Our concern is informed by the well documented ethics issues associated with ML systems – from our perspective it is the societal significance and pervasiveness of these issues that merits a close scrutiny of the tools ML practitioners use to develop ML systems. As such, in this section we introduce some of the known ethics issues associated with ML systems.

We do not aim in this section for a comprehensive review of the literature on ML ethics issues (see [Mehrabi et al., 2019] and [Barocas et al., 2021b] for these), rather we aim only to demonstrate that ML systems are associated with significant ethics issues. We focus, in particular, on issues where we believe ML platforms, through their use to develop ML systems, are likely to play a significant role in the manifestation of the issue. In Section 2.5, which follows this section, we use the context provided in this section to argue for the importance of conceptualising ML platforms as socio-technical systems. In Section 2.6 we then consider how growing awareness amongst ML practitioners of ethical issues associated with ML systems is influencing the way in which practitioners enact ethical considerations during ML system development. And, in Section 2.7 we review existing fair-ML proposals for interventions in ML system development, designed to address to the ethics issues introduced in this section.

The scholarship that has brought the ethics issues we describe below to light is largely the product of fair-ML researchers, working in North America and Europe since 2016. In 2016 ProPublica, a United States (US) based publishing site for investigative journalism, published a detailed account of racial biases against Black people in an ML system used by many US jurisdictions to predict the risk of recidivism [Angwin et al., 2016; Larson et al., 2016]. The system was incorporated into COMPAS, a commercial software product developed by Northpointe to support judges in criminal justice cases to make sentencing decisions. ProPublica's reporting on COMPAS gained widespread attention in the news media and prompted public policy debate regarding the use of predictive ML systems in the criminal justice system (e.g. see [Smith, 2016]). ProPublica's analysis of COMPAS was critiqued by Northpointe, who disputed ProPublica's evidence that the ML system was producing racially biased outputs [Dieterich et al., 2016]. The controversy regarding COMPAS, along with other high profile analyses implicating ML systems in biased treatment of minority populations (e.g. in healthcare see [Obermeyer et al., 2019]), prompted a concerted effort amongst researchers and investigative journalists concerned with the social impact of computing to investigate how ML systems were being put to use,

particularly in the US. From this effort, the fair-ML community has grown [2021a].

Four key texts have been published documenting negative impacts of ML approaches on individuals and communities, mostly in the US. These texts provide critical empirical evidence for academic discourse on the ethical issues raised by ML systems' integration into contemporary life. In 2016 Cathy O'Neil published *Weapons of Math Destruction*, which documents the social harm caused by automated algorithms, many incorporating ML models, deployed in high-stakes decision making [2016]. In 2018 Virginia Eubanks published *Automating inequality*, which documents how ML systems ossify inequitable decision making and resource allocation processes in the private and public sectors [2018]. Also in 2018, Safiya Umoja Noble published *Algorithms of oppression*, which documents how the algorithmic systems embedded in search engines reinforce racism [2018]. Finally, in 2019, Ruha Benjamin published *Race after technology*, which documents more broadly the relationship between emerging technologies and white supremacy [2019]. The following section draws on these four texts.

To the best of our knowledge, for the three ethics issues detailed below there is no published research on their relation to ML platforms. As such, for each issue, our description of its relevance to ML platforms is speculative, and based largely on abductive reasoning informed by publicly accessible information on the features of popular ML platforms. We hope this speculation, however, is sufficient to demonstrate that, at least for the ethics issues we describe, their association with ML systems implies also an association with the ML platforms used to develop such systems. We also note that the ethics issues associated with ML systems extend well beyond the three issues we describe below. Among the issues we do not discuss, but that may be of relevance to ML platforms are: the environmental impact and energy demands of ML systems [Joler and Crawford, 2018; Bender et al., 2020; Henderson et al., 2020]; the labor conditions of workers involved in the collection and pre-processing of ML training datasets [Joler and Crawford, 2018; Kshetri, 2021; Parekh and Natarajan, 2021]; and, the relationship between ML systems and new forms of commercial practices [Zuboff, 2020; Srnicek, 2017].

2.4.1 Algorithmic opacity

The empirical evidence documented in the above texts is critical due to the research challenge of accessing ML systems and particularly the ML algorithms, models and training data they rely on. This phenomenon is described as 'algorithmic opacity' in the literature [Burrell, 2016; Christin, 2020]. Jenna Burrell describes three sources of opacity in ML systems: opacity due to intentional corporate or state secrecy; opacity due to technical illiteracy; and, opacity due to the way ML systems operate at scale [2016].

Corporations or states managing ML systems (*ML system operators* in the terminology we introduced in Section 1.2) may wish to ensure the details of their ML systems, particularly the technical details of how the ML models within their systems operate, are secret to avoid enabling ML system users to game the system or

for security reasons [Burrell, 2016]. Additionally, corporate or state secrecy may be motivated by the desire to avoid regulatory oversight or to retain competitive advantage (see [Pasquale, 2015] for a detailed exploration of corporate motivations and structures for enforcing algorithmic opacity).

Technical illiteracy amongst ML system users can render the operations of ML systems opaque to them [Burrell, 2016]. Even where ML system users are trained in machine learning approaches, technical illiteracy may still be an issue due to poor code documentation, or to complex ML system structures which make it practically challenging for even ML practitioners to reverse engineer how an ML system is functioning [Seaver, 2017]. These challenges can be particularly acute where an ML system's place of development and place of deployment are spatially or culturally distanced [Sambasivan et al., 2021].

ML systems when operating at scale incorporate heterogeneous and high dimensional data, with ML models producing complex predictions, which can be opaque to ML system users and ML practitioners [Burrell, 2016]. In theory, argues Burrell, the complexity of predictions is not an inherent feature of ML systems, but in practice ML systems are only useful when complex predictions are required [2016, p.5]. Complex predictions are a source of algorithmic opacity because the interplay between large training datasets, input data, and the operationalised ML algorithm can be very challenging for humans to comprehend. Fundamentally, the purpose of an ML model is to learn patterns in data which human minds cannot otherwise express [Dourish, 2016]. Paul Dourish offers a useful example of this: the engineers at Twitter can explain how the ML system used to determine trending topics operates, but cannot explain why specific events are or are not classified as 'trending' [2016]. The challenge of engineering computing systems that are not opaque in this way is the focus of the Explainable Artificial Intelligence sub-field [Miller, 2019].

Where ML systems are deployed to support the automation of high-stakes decisions, algorithmic opacity can become an ethical issue in two ways: algorithmic opacity can negate the agency of ML system users; and, algorithmic opacity can frustrate efforts to hold ML system operators and ML system developers accountable for the outcomes of their systems. Oswald et al. explore these issues in the context of the Harm Assessment Risk Tool (HART), an ML system developed for a regional police force in the UK to predict the risk of future offending by arrestees and to inform decisions regarding enrolling arrestees in a prison diversion program [2018]. The authors note that the HART system incorporates over 4.2 million interdependent decision points, and as such is opaque due to its operation at scale [2018]. In theory, the decision points could be made public, so as to enable members of the public to review how the HART system was generating its predictions, but such a review would take an enormous amount of time and effort. As such, the agency of arrestees who are subject to predictions of the HART system is diminished, as their lack of ability to understand how predictions are generated diminishes their capacity to question or challenge the system's predictions. The accountability of decision-makers in the police force is also challenged by the HART system's opacity: decision-makers who rely on the HART system's predictions are not expert ML practitioners, and, simi-

larly to arrestees, are unable to explain or defend the system's predictions [Oswald et al., 2018].

ML platforms can be thought of as tools that moderate the opacity of ML systems. An ML platform that enables platform users to easily deploy poorly documented pre-trained ML models may contribute to the profusion of ML systems that are opaque, because it may be very challenging for an ML system subject or ML practitioner to determine how such a model is generating its predictions. Conversely, an ML platform that offers platform users tools for visualising how ML models are interpreting and reacting to inputs might contribute to the development of more explainable and less opaque ML systems – although, user research on such 'interpretability' tools challenges this assumption [Kaur et al., 2020]. Additionally, Burrell's third source of algorithmic opacity, the operation of ML systems at scale, is, in many instances, the raison d'etre for the existence of ML platforms; ML platforms are fundamentally designed to enable easier development and operation of ML systems at scale. As such, it seems possible that ML systems which system subjects and users experience as opaque, like the HART system, are reliant on an ML platform for their development and operation.

2.4.2 Biased or discriminatory outcomes

Much of the existing data regarding ethics issues associated with ML systems is drawn from studies of ML system bias or discrimination. In the field of Machine Learning, bias has a specific meaning and so in discussing these works it is helpful to distinguish between *statistical bias* and *demographic disparities* [Barocas et al., 2021b]. In Machine Learning, statistical bias occurs when an ML model produces estimates that are systematically different to the true value that the ML model is attempting to estimate [Barocas et al., 2021b; Mitchell et al., 2021]. In some contexts, statistical biases in an ML model can cause an ML system to performs differently for different social groups in ways that contravene societal expectations. In popular discourse this might be described as biased or discriminatory performance, but following Barocas et al., to avoid confusion we prefer the term *demographic disparities* for describing such ML system outcomes [2021b].

To illustrate the relationship between statistical bias and demographic disparities, consider an ML system designed to recognise objects in images. DeVries et al. found that the object-recognition systems offered by Microsoft, Google, Amazon, IBM and Clarifai's ML platforms perform less accurately for household items found in the houses of people in countries with low average household income than for household items found in the houses of people in countries with high average household income [2019]. To collect datasets of images with labelled items to train an object-recognition system, ML practitioners tend to start with an English-language list of nouns, which they then search for in online image depositories, such as Flickr [DeVries et al., 2019]. An object-recognition systems trained with an image dataset collected from English-language nouns is likely, however, to be statistically biased towards items that are well covered by English-language nouns, or to images that

reflect English-language understandings of such nouns. Some items may simply not have an English-language noun associated with them, and so will be completely unrepresented in the training dataset and therefore not be accurately recognised by the object-recognition system. Other items may have language-dependent meanings, and so will only be partially represented in the training dataset, and only sometimes recognised by the object-recognition system. When this statistical bias towards items associated with English-language nouns is contextualised within the global distribution of the English-language it becomes clear how such object-recognition systems can have demographically disparate performance. English is more likely to be spoken in countries with high household income than those with low household income, and therefore the image datasets based on English-language nouns are more likely to include images of objects found in countries with high household income [DeVries et al., 2019]. Were English spoken universally, then whilst the object-recognition systems might still be said to be statistically biased towards items well covered by the English-language, such a bias may not produce the demographic disparity.

Several authors have compiled lists of the various types of statistical bias which can manifest in ML systems (see [Mehrabi et al., 2019; Silva and Kenney, 2018; Suresh and Guttag, 2019]). One form of statistical bias particularly relevant to this research project is *representation bias*, also sometimes referred to as *sampling bias* (e.g. in [Mitchell et al., 2021]). Representation bias occurs when the dataset used to train an ML model does not adequately represent (i.e. sample) a population that will be ML system subjects [Mehrabi et al., 2019; Mitchell et al., 2021; Suresh and Guttag, 2019]. Such a model is likely to perform less well for populations under-represented in the training dataset compared to those who are adequately represented [Suresh and Guttag, 2019]. Suresh and Guttag identify three sources of representation bias [2019] – these sources help explain the relationship between representation bias and disparate demographic performance by ML systems. The first source, as alluded to above, is when the training dataset does not reflect the target population for the ML system. A facial recognition model trained on a dataset that is representative of the United States is likely to have a variety of representation biases if it is deployed in an ML system to be used in Australia. For population groups in Australia, which are not in the United States, the facial recognition model is likely to under perform. The second source is when the target population itself contains under-represented groups. A facial recognition model trained on a dataset that is an accurate representation of the Australian population is likely to have representation biases, because there are sub-populations within Australia who are minorities, such as Aboriginal and Torres Strait Islander peoples (3.3% of the Australian population, [ABS, 2016]). For these sub-populations, the facial recognition model will likely under perform because very few examples of these sub-populations will have been included in the training dataset. Finally, representation bias can arise if the sampling method used to develop a training dataset is limited in some way. If the training dataset for a facial recognition system to be deployed in Australia is drawn from images posted on social media by accounts located in Australia, then the dataset will be a skewed subset of the Australian population. Each of these sources of representation bias present

ethics issues, depending on the context in which an ML system is deployed.

Sub-populations that are the subject of representation biases in datasets may also be subject to discrimination in their interactions with social institutions, such as police or schools. When such institutions become ML system operators there is a risk that representation biases in their training datasets may result in them deploying ML systems that under-perform for sub-populations that are already experiencing institutional discrimination. In other words, ML systems may exacerbate the pre-existing disparate demographic performance of social institutions (see [Perkowitz, 2021] for a description of this process in the context of policing in the United States and [Andrejevic and Selwyn, 2020] for a similar description in the context of schooling, also in the United States). Representation bias is particularly relevant to our exploration of ML platforms because ML platforms are designed to increase the interoperability of ML models between different ML systems. ML platforms enable ML practitioners to make use of pre-trained ML models, and to make their ML models available for use by other ML practitioners. In doing so, ML platforms may increase the risk of representation biases occurring in ML system development, as ML models become abstracted away from their training datasets and deployed in target populations that may be significantly different from population represented in the training dataset.

Whilst some authors begin their study of bias and discrimination in ML systems from the perspective of statistical bias, seeking to follow issues of statistical bias in ML models through to their manifestation as disparate demographic performance in ML systems, others begin by seeking to understand the impact of disparate demographic performance of ML systems on ML system subjects. Disparate demographic performance based on race [Obermeyer et al., 2019], gender [Buolamwini and Gebru, 2018; Albert and Delano, 2021], disability [Whittaker et al., 2019], and other features (e.g. accents [Lima et al., 2019]) has been explored in the fair-ML literature. Algorithmic opacity can exacerbate issues of disparate demographic performance by making it challenging for those who are impacted to collect evidence establishing their claims, particularly in terms of demonstrating discriminatory intent (a requirement for those seeking legal intervention in many jurisdictions) [Cofone, 2019]. The cloud-based nature of ML platforms may further compound this issue by spreading ML system components across servers in multiple jurisdictions, and increasing the physical and legal distance between ML system subjects and ML system developers (see [Sambasivan et al., 2021] for a discussion of this in the context of India).

2.4.3 Normative judgments on contested terms

ML system development often requires ML practitioners to attempt to translate complex social concepts, such as fairness, into mathematical formalisms [Selbst et al., 2019; Green and Hu, 2018]. Additionally ML system development often asks ML practitioners to infer unobservable theoretical constructs, such as teacher effectiveness, from proxy data [Jacobs and Wallach, 2021]. In both of these cases, ML systems encode the normative judgments of ML practitioners into their operation, which can create ethics issues when the judgments of ML practitioners differ from societal

expectations or when societal expectations change over time, whilst the ML system's encoding of normative judgments remains static [Jacobs and Wallach, 2021]. Additionally, researchers have highlighted the lack of gender and racial diversity within ML teams in the technology industry and argued that the normative judgments made by ML practitioners often reflect the interests of cis-gendered white male ML practitioners, rather than the populations who will be ML system subjects [West et al., 2019]. Os Keyes, for example, conducted a comprehensive survey of human-computer interaction and computer science articles on automatic gender-detection ML systems and found that these systems overwhelmingly encode a binary physiological model of gender, which erase transgender people from their conceptualisation of gender [2018]. The impact of this erasure can be the development of ML systems which either exclude transgender people, or require transgender people to conform to the false gender binary in order to interact with the system (see [Albert and Delano, 2021] for an exploration of how this exclusion manifests in a smart scale consumer device).

The issue of normative judgments on contested terms may be of relevance to ML platforms due to the role ML platforms can play in normalising ML system development practices and in increasing access to pre-trained ML models, which may increase the reach of the normative judgments encoded in such models. ML platforms may enable the normative judgments of ML practitioners to scale in unexpected ways. Additionally, by surfacing pre-trained ML models that may have been designed for one socio-political context to ML practitioners in other socio-political contexts, ML platforms may increase the potential for the normative judgments encoded in ML models to be inconsistent with the socio-political context in which the model is deployed. Alternatively, ML platforms, as centralised repositories for pre-trained ML models and ML system development practices, could also encode limitations on normative judgments or processes for supporting ML practitioners to be more reflexive in the judgments they encode in the ML models they develop. To this end, tools such as Microsoft's Fairlearn, which is designed to enable ML practitioners to test their models against multiple definitions of fairness [Bird et al., 2020], could be integrated into ML platforms. Meaningful integration, however, will need to grapple with the challenge of training ML platform users in applying the tool [Kaur et al., 2020].

In this section we introduced three ethics issues prominently associated with ML systems. In the following section, Section 2.5, we introduce the concept of socio-technical systems, and consider how conceptualising ML systems and ML platforms as socio-technical systems can help reveal the relationship between the ethics issues discussed here and their manifestation in ML systems developed on ML platforms.

2.5 Socio-technical systems

In the previous section we introduced three ethical issues associated with ML systems – algorithmic opacity, biased or discriminatory outcomes, and normative judgments on contested terms – and argued that these issues are also relevant to ML platforms. In this section we introduce the concept of socio-technical systems, as a way of conceptualising ML systems that helps reveal the relations between ML system components, ML practitioners, and ethics issues associated with ML systems. We also apply the concept of socio-technical systems to ML platforms, using this way of conceptualising ML platforms to reveal the complex web of relations between ML systems, ML platforms, ML system users and subjects, and ML platform users.

Our argument in this section draws on a tradition of sociological critique of how those engaged in the practice of computing research conceptualise the products of their efforts. This tradition has unfolded in parallel to the last seven decades of computer science research and development. At the core of the critique is the observation that the practice of computing research often treats computing as a positivist technical enterprise in which technical components of computing systems are considered within the boundaries of computing systems, and the legitimate focus of computing research, whilst social components are considered outside the boundaries of computing systems, and therefore outside the focus of computing research. We start this section by reviewing two examples of this critique, applied to sub-fields of artificial intelligence that predate the current machine learning field. We then turn to examples of this critique from the fair-ML literature. We interpret this literature as a discourse focused on expanding how ML practitioners understand the social and technical boundaries of ML systems. We argue that the aim of this literature is to ensure that the boundaries of ML systems are conceptualised in such a way as to position ML practitioners as responsible for the ethics issues associated with ML systems. We conclude by outlining how we conceptualise the system boundaries of ML platforms, and reviewing the social and technical components that sit within them.

2.5.1 Socio-technical systems in literature preceding fair-ML

An early example of the sociological critique of computing research can be found in the writings of Norbert Wiener, a mathematician and AI researcher. Wiener, in his 1948 treatise on cybernetics, described the potential for intelligent and autonomous computing systems to be constructed by modelling the neural structure and processes of the human brain – and argued that such systems were of comparable social significance to the atomic bomb [Wiener, 1961]. Like the atomic bomb, Wiener claimed that intelligent computing systems could be put to use by humans in both ethical and unethical pursuits [Turilli, 2008, p.3]. Wiener additionally predicted that computer technology would reshape society, causing an industrial revolution that would lead to sweeping changes in human work, organisation, and thought [Bynum, 2001]. Wiener's argument implicitly treats computing systems as socio-technical systems – computing systems are intertwined in their social contexts, with the social

context shaping how computing systems are used and the use of computing systems similarly shaping the social context itself. In recognition of this, Wiener argued that scientists, in practicing science, have an ethical responsibility to consider how the outputs of their scientific activities are likely to be put to use [1947]. In other words, scientists needed to treat their scientific outputs as socio-technical systems, rather than as purely technical constructs.

More recently, in the 1990s, Diana Forsythe critiqued the way computer scientists developing expert systems conceptualised knowledge in an anthropological study of computer science laboratories in the United States [1993]. Expert systems were computing systems designed to act as consultants to human decision-makers in specialist domains by emulating the reasoning of specialists [Duda and Shortliffe, 1983]. Computer scientists developing expert systems attempted to formalise and codify the reasoning processes of specialists, such as medical doctors when diagnosing a patient, through a three stage process of information extraction from specialists, ordering that information into machine-readable procedures, and developing computer programs to use these procedures to make classification or other decisions based on new inputs [Forsythe, 1993]. Forsythe drew on Susan Leigh Star and Anslem Strauss's sociology of work to argue that computer scientists developing expert systems were attempting to delete the social and cultural context from their construction of expert knowledge. The computer scientists Forsythe observed treated the process of information extraction from specialists and ordering of that information as straightforward, rendering their own role in determining what was codified in the expert system invisible [1993, p.463]. Further, the computer scientists treated knowledge itself as a straightforward phenomenon, in which universal formal rules are consciously applied by experts and where experts' actions perfectly reflect their logical reasoning (see [Flyvbjerg, 2001] for a similar argument regarding how science more generally conceptualises knowledge). Forsythe connected this naive treatment of knowledge to the project of developing expert systems. If computer scientists were to accept the idea that knowledge is socially contingent, an "*outcome of ongoing processes of construction and interpretation*", then their method of developing rules based expert systems would immediately appear doomed to fail because it was reliant on the positivist assumption that knowledge could be abstracted into a static and coherent set of logical procedures [Forsythe, 1993, p.466]. To improve the functioning of expert systems Forsythe recommended that computer scientists rethink their conception of knowledge, attend more to the processes of knowledge elicitation and data-gathering, and reflect on their own power as the definers of 'expertise' in the development of expert systems [Forsythe, 1993]. In Forsythe's argument, like Wiener's, computing systems are implicitly conceived of as socio-technical systems. In the language of socio-technical systems, Forsythe's argument is that expert systems should be conceived of as socio-technical systems, whose social components include the computer scientists developing them, the specialists who knowledge the systems are attempting to emulate, and the processes of information extraction and information ordering.

2.5.2 Socio-technical systems in the fair-ML literature

In the context of Machine Learning, Forsythe's argument is echoed in the fair-ML literature, where researchers have conceptualised ML systems as socio-technical systems, and sought to problematise the construction of 'fairness' by ML practitioners.

Andrew Selbst et al. describe early fair-ML literature as focused on "*trying to engineer fairer and more just machine learning algorithms and models by using fairness itself as a property of the (black box) system*" [2019, p.59]. In this mode of thinking, fairness is treated as a property of the system in the sense that the objective of the researcher becomes to develop a definition of fairness that can be expressed as a set of metrics which can be used to evaluate or constrain the performance of the system. Selbst et al. argue, however, that in these works the system is bounded too narrowly: only the machine learning algorithm and model, the inputs, and the outputs are considered, and any social context surrounding the system is abstracted away. The problem is that the social context surrounding the system includes information needed to meaningfully conceptualise fairness and to substantively evaluate ML system outcomes. Issues like fairness, therefore, cannot be usefully discussed without conceptualising ML systems as socio-technical systems, in which social actors, institutions, and interactions are included within the system boundaries. To treat ML systems as purely technical systems, in Selbst et al.'s argument, is to make an "*abstraction error*" [2019]. Selbst et al. conclude that fair-ML research should focus on "*the process of determining where and how to apply technical solutions*", such as Machine Learning approaches, to social problems [2019, p.66].

Ben Green and Salome Viljoen also put forward an argument in favour of conceptualising ML systems as socio-technical systems, although they argue specifically for the inclusion of computer scientists' ways of thinking within the conceptual boundaries of ML systems [2020]. The authors focus on the relationship between 'algorithmic thinking' – the way algorithms are taught and understood – and 'algorithmic interventions' – how algorithms are deployed. Green and Viljoen describe algorithmic thinking as shaping the boundaries of what computer scientists consider relevant to the development of algorithmic interventions. Green and Viljoen are critical of 'algorithmic formalism', which they describe as the dominant mode of algorithmic thinking. Algorithmic formalism is characterised by notions of objectivity and neutrality, internalism, and universalism. Formalism as a method, they note, helps produce analytical clarity, and can help identify solutions to complex problems [Green and Viljoen, 2020, p.21]. But, failure to recognise the limitations of formalism can be dangerous. The fundamental limitation is that formalism entails a narrowing of vision through formal expression of complex concepts which means that formal approaches must inevitably leave out some elements of the socio-technical context. This narrowing of vision involves boundary decisions about what is and is not important or legitimate to the problem and solution, which are inherently value-laden [Midgley et al., 1998]. The creation of formal solutions is therefore a political act, although the narrowing of vision itself often makes this hard to see as political considerations themselves are generally considered difficult to formalise. In response to the limita-

tions of algorithmic formalism Green and Viljoen argue for developing 'algorithmic realism' as a complimentary mode of algorithmic thinking for computer scientists to deploy when considering algorithmic interventions. Algorithmic realism eschews a focus on the formal attributes of algorithms for a focus on the ways algorithms materially manifest in everyday life. In doing so, algorithmic realism focuses on how social context shapes algorithmic interventions, expanding the boundaries of computer systems to include social components, particularly computer scientists and their ways of thinking, alongside technical components.

2.5.3 ML platforms as socio-technical systems

Green and Viljoen, and Selbst et al. advocate for expanded conceptualisations of ML systems as socio-technical systems. Their arguments apply also to ML platforms. Indeed, thinking about ML platforms as socio-technical systems can elucidate limitations in how Green and Viljoen, and Selbst et al. conceptualise ML systems. Applying Selbst et al.'s argument, a purely technical description of ML platforms abstracts away the social context surrounding ML platforms. The social context, however, is needed to explain where and how ML platforms contribute to the development and management of ML systems, and therefore how ML platforms relate to the ethics and governance issues associated with ML systems. Applying Green and Viljoen's concepts of algorithmic thinking, ML platforms can be seen to reflect the core notions of algorithmic formalism. ML platforms are designed to be universal, to present ML platform users with objective and neutral metrics on the performance of ML systems, to act essentially as technical systems. However, the perspective of algorithmic realism is needed to understand how ML platforms fit into and reconfigure the role of ML practitioners, and introduce new social components into the process of developing ML systems. By recognising that ML platforms are socio-technical systems that support the development and creation of another form of socio-technical system, that is ML systems, we draw attention to the process of ML system development. In addition to Selbst et al.'s call for fair-ML to focus on the process of determining where and how to apply technical solutions, we can add that fair-ML should treat technical solutions as ongoing processes of creation, and focus on the tools and actors engaged in this process. And, when the tools and actors engaged in the process of developing ML systems are brought into focus it becomes clear that Green and Viljoen's focus on computer scientists needs nuancing. Whilst computer scientists (ML practitioners in the nomenclature we introduced in Section 1.2) may be the primary social actors who interact with ML platforms, the features of ML platforms themselves (e.g. systems for collaborative work) make it clear that ML system development is a social process involving numerous actors. It is not clear that all of these actors should be described flatly as computer scientists, or whether algorithmic formalism is the primary mode of thinking for all these actors.

By conceptualising ML platforms as socio-technical systems, and placing them alongside ML systems in our analysis, we hope to bring into sharp relief the processes by which ML systems are developed and the contexts in which the ethical

issues described in Section 2.4 manifest. We sweep into our conceptualisation of ML platforms as socio-technical systems the ML practitioners who design and maintain the platforms (ML platform developers) and the ML practitioners who are users of the platforms (ML platform users). Thinking of ML platforms as socio-technical systems therefore helps reveal the ways in which their boundaries are socially contingent; the social context we are studying will determine who ML platform users are in that context. Additionally, we can deepen our understanding of the social components of ML platforms by paying close attention to the breadth of people who may be swept into our definitions of ML platform developers and ML platform users. Clearly, amongst both ML platform developers and ML platform users are computer scientists, software engineers, and data scientists. But, so too are the business analysts, product managers, compliance managers, user experience designers, and developer operations specialists who contribute to and shape ML platform design and maintenance, and ML system development. Finally, by tracing the connections between these social components and the technical features of ML platforms we can explore the politics of these technical features – we can consider how ML platforms embed the social values and structures of those who manufacture and maintain them in the broader system of ML development, and are embedded in these social values and structures [Jasanoff, 2004, p.3].

In this section we reviewed how ML systems have been conceptualised as socio-technical systems, and argued that this conceptualisation is appropriate for ML platforms too. In the following section, Section 2.6, we discuss how the social components of ML platforms – particularly ML platform users – enact ethics considerations when developing ML systems.

2.6 Enactment

In this section we consider how knowledge of the ethics issues discussed in Section 2.4 may influence the ways in which ML practitioners develop ML systems. We draw on the conceptualisation of ML systems as socio-technical systems discussed in the previous section, Section 2.5, to reflect on how ethics considerations are enacted by ML practitioners and the tools they use to develop ML systems, in particular ML platforms. This section is in an early stage of development, and we expect to update and expand it considerably as our research unfolds. In particular, we hope to collect empirical evidence of the ways in which ethics issues are considered by ML practitioners, which may validate or invalidate our conceptualisations of this consideration as a form of enactment. Our data collection plans and research design are described in Section 4.1.

The etymology of *enact* speaks to the multiple meanings of enactment. To enact, historically, has referred to the process of entering a decree into the public record and also to the performance of a play or role within a play [Hoad, 2003]. In modern language, to enact may refer to making something law, putting into practice an idea

or concept, or performing a role [Stevenson, 2010a]. Enactment thus refers to the process by which something is made law, a concept is put into practice, or a role is performed [Stevenson, 2010b]. In the context of this research proposal, we use the term *enactment* to refer to the process by which ethics considerations are acted out and made manifest during ML system development, deployment, and management. Enactment of ethics considerations may include the creation of 'law', in the form of documents that state ethical principles by which ML practitioners intend to abide, the establishment of processes or methodologies for supporting ML practitioners to consider specific ethical issues during ML system development or put into practice documented ethical principles, and the manner in which ML practitioners perform their roles and duties, both formal and informal. In our research, the performative aspect of enactment is also relevant, as it points to the potential for competing and divergent motivations to inform the ways in which ethics considerations are enacted in technology firms, and it enables us to consider how some enactments of ethics considerations may be primarily performative or instrumental [Bietti, 2020]. Importantly, it should be noted that the enactment of ethics considerations is an implicit and inherent part of ML system development [Turilli, 2008]. In other words, we assume that all ML system development involves the enactment of ethics considerations, and seek to understand how enactments differ across different social contexts and how different enactments may result in different outcomes for ML system subjects.

Whilst, to the best of our knowledge, the concept of enactment has not been explicitly applied in fair-ML discourse to date, it is reflected in existing fair-ML research on how ML practitioners develop ML systems and engage (or, sometimes, fail to engage) in ethics considerations. Four recent studies have explored the engineering practices of ML practitioners, largely through a mix of interviews (N between 11 and 35) and surveys (N between 197 and 267) [Kaur et al., 2020; Holstein et al., 2019; Orr and Davis, 2020; Veale et al., 2018].

Kaur et al. explored how ML practitioners at a large technology firm use interpretability tools to support ethics considerations during ML system development [2020]. Reflecting our conceptualisation of enactment, Kaur et al. describe how ML practitioners use interpretability tools in their day-to-day work to surface potential ethical issues, such as bias or under-representation in datasets, and to inform their considerations for how to mitigate such issues [2020]. Kaur et al. find, however, that interpretability tools can be misused by ML practitioners; at times ML practitioners may rely on interpretability tools to enact ethics considerations in ways that align with their own interests, particularly their interest in deploying ML models they have developed, rather than in ways that address the substantive ethical issue [2020]. We discuss this study further in our research design section, Section 4.1, as the contextual inquiry method adopted by Kaur et al. may be applicable to our research project.

Holstein et al. conducted a systematic investigation of the challenges faced by ML practitioners working in product teams in large technology firms to develop 'fairer' ML systems [2019]. Holstein et al. note that the development of ML systems raises a host of ethics issues, which ML practitioners need to navigate in the course of

their day-to-day work [2019]. Holstein et al.'s research identifies a number of challenges facing ML practitioners attempting to enact ethics considerations in a robust manner, including: support for fairness-aware data collection and curation practices, processes for conducting audits of ML systems, and frameworks for deciding how to mitigate or address ethics issues once they are discovered [2019]. Relevantly, Holstein et al. conclude that future fair-ML research should focus on developing tools and interventions that respond to these challenges, and can be applied by ML practitioners in their day-to-day work to support their enactment of ethics considerations [2019].

Orr and Davis explored how ML practitioners in a range of organisational settings in Australia attribute responsibility for ethics considerations during ML system development [2020]. Orr and Davis's conceptualisation of the practice of ethics aligns closely with our description of enactment. They study "*ethics as part of practitioners' in-situ processes*", noting that practitioners' working definitions of ethical norms "*reveal how ethics manifest into procedures, decisions, and material goods*" [Orr and Davis, 2020, p.2]. Orr and Davis found that the ML practitioners they interviewed all recognised ethics considerations as a significant part of their work, and were well aware of instances of ethical failure within ML system development [2020]. Orr and Davis recognise, however, that ML practitioners enact ethics considerations within a complex web of relations with other actors and interpret responsibility for ethics considerations to be dispersed across this web [2020]. Orr and Davis highlight, in particular, the relationship between actors with the power to set the parameters for ethics considerations, such as legislators, managers, and clients, and actors with the expertise to enact ethics considerations during ML system development, ML practitioners themselves [2020]. Orr and Davis's research highlights that the enactment of ethics considerations is a social process that occurs through processes of negotiation between ML practitioners, ML system operators, ML system users and their broader social context. For our research, this highlights the need for tools and interventions intended to support the enactment of ethics considerations to be designed in a way that attends to this complex web of relations.

Finally, Veale et al. investigated how ML practitioners working in public sector organisations grapple with ethics issues during the development of ML systems intended to support high-stakes decision making [2018]. Veale et al. highlight that ML practitioners in public sector organisations often have a high degree of awareness regarding ethical issues associated with ML system development. However, these practitioners sometimes struggle to enact ethics considerations in ways that engage with other actors within their organisations, such as senior management or front-line workers [2018]. Veale et al. conclude that tools and interventions intended to improve the fairness and accountability of ML systems need to consider "*[d]omain-specific, organisational and contextual factors*" [2018, p.10]. This conclusion is salient to our research as it highlights an inherent challenge in our focus on ML platforms as sites at which ethics considerations are enacted. ML platforms are effectively domain, organisation and context agnostic – they are designed to support ML practitioners to develop ML systems in any domain. As such, a focus in our research will need to

be on accounting for potential mismatches between assumptions made at the platform level regarding how ML system development occurs and ethics considerations are enacted by ML platform users, and the reality of ML system development at the level of ML platform users, working in product teams, to develop specific ML systems.

In this section we reviewed four studies of how ML practitioners enact ethics considerations in ML system development. These studies have informed proposals by fair-ML researchers for new tools and interventions designed to support ML practitioners and other actors in their enactment of ethics considerations. In the following section we consider these proposals.

2.7 Proposals for interventions in the enactment of ethics considerations

In the previous section, Section 2.6, we conceptualised fair-ML research on ML system development as the study of the enactment of ethical considerations by ML practitioners. The fair-ML community, however, is also highly interventionist and a substantial portion of fair-ML discourse focuses on proposals for intervening in the research, design, and management of ML systems, with the aim of supporting more substantive enactments of ethical considerations so as to avoid the proliferation of ML systems that contribute to social harms or inequities. In this section we introduce these proposals. We aim to provide an overview of the breadth of proposed interventions, rather than a detailed or systematic review of them (see [Morley et al., 2019a; Schiff et al., 2020] for a more detailed review of proposed interventions). In the following section, Section 2.8 we consider how these proposed interventions in the enactment of ethical considerations account for the tools, particularly ML platforms, used to support ML system development.

We note that much fair-ML research has focused on establishing formal definitions of fairness, which can be used to evaluate the performance of an ML model (see [Corbett-Davies and Goel, 2018; Mitchell et al., 2021] for overviews of formal definitions of fairness, and [Kasy and Abebe, 2020; Jacobs and Wallach, 2021] for critiques of formal definitions). The work of establishing a formal definition of fairness can be thought of as an intervention in the enactment of ethical considerations, however it is not the primary focus of this section or our literature review. This is because our research is focused on the enactment of ethical considerations in the context of ML system development, largely in industry settings, such as technology firms. As such, we focus primarily on interventions that have been designed to be implemented by actors in these settings. As we discuss below, these interventions, such as audits and technical tools for assessing fairness, tend to support ML practitioners to apply formal definitions of fairness in the course of their enactment of ethical considerations, rather than to develop new definitions as an output of their enactment. In other words, we consider formal definitions of fairness to be one form of input into

interventions designed to support the enactment of ethical consideration.

2.7.1 Governance principles

One of the primary modes of enacting ethical considerations in the context of Machine Learning is the creation of high-level governance principles (often referred to as AI principles or ethical principles) [Jobin et al., 2019]. Governance principles have been created and championed by various interested entities, from technology firms [Pichai, 2018; **?**], to public sector agencies [High-Level Independent Group on Artificial Intelligence (AI HLEG), 2019], to non-governmental organisation [IEEE, 2019]. Jobin et al. identified and reviewed 84 documents containing governance principles, produced by organisations primarily in Europe and the United States [2019]. For an overview of the entities engaged in developing governance principles see [Schiff et al., 2020]. For an analysis of the contents of the principles contained within these documents see [Jobin et al., 2019].

Governance principles are normative statements intended to reflect societal expectations and values, and designed to guide the development of AI technologies, including ML systems. As such, the documentation of governance principles have been compared to a Constitution; their role is to provide a foundation for the development of practical ethical guidance [Morley et al., 2021]. Key themes reflected in the existing body of published governance principles are transparency, justice and fairness, non-maleficence, responsibility, and privacy [Jobin et al., 2019]. These themes reveal a close connection between the content of governing principles and the ethics issues associated with ML systems, discussed in Section 2.4. Concerns regarding algorithmic opacity, for instance, may result in calls for increased transparency [Felzmann et al., 2019]. Similarly, concerns regarding bias and discrimination may result in calls for clear commitments to justice and fairness [Kasy and Abebe, 2020]. As enactments of ethical considerations, however, it should be noted that governance principles are also artifacts that reflect the interests and context of those who create them. As such, the translation of concern regarding bias and discrimination into governance principles focused on fairness has been critiqued as too narrow (see [Kasy and Abebe, 2020] for a critical examination of the limitations of 'fairness' as a governance principle in ML system development).

The fair-ML discourse has highlighted several limitations with the enactment of ethical considerations through the development of governance principles. First, there can be considerable challenges in translating high-level principles into practice, and debate focused on the creation of high-level principles can sometimes obfuscate more prosaic efforts to establish accountability mechanisms within organisations developing ML systems [Morley et al., 2021; Baxter et al., 2020]. Second, high-level governance principles may describe best practice for ML system development in contexts where no minimum standards exist, further complicating the translation of principles into practice [Floridi, 2018]. Third, the creation of governance principles may facilitate some ethics discussions whilst forestalling others [Greene et al., 2019]. Where governance principles are created by an organisation they are likely to reflect the

fundamental assumptions and business logic of that organisation, even when those assumptions or logic themselves may be the source of ethics issues [Green, 2021]. Finally, governance principles may be positioned by the organisations creating them as universal statements, ignoring the potential for societal expectations or values to diverge significantly from place to place [Sambasivan et al., 2021]. Given these critiques, in this research project, whilst we recognise that the creation of governing principles is an almost ubiquitous form of enactment of ethical considerations by technology firms, we intend to focus our research efforts on how governing principles are operationalised in the context of ML system development, rather than how the principles themselves are established.

2.7.2 Auditing ML systems

To evaluate the alignment of ML system development practices and ML system performance with high-level governance principles, and to hold ML practitioners accountable for the performance of their systems, various audit processes have been proposed by fair-ML researchers. In general use, an 'audit' is usually understood to be "*an official examination, often conducted by third parties, with mutually agreed-upon actors, responsibilities, and expectations* [Barocas et al., 2021a, p.2]. Fair-ML researchers, however, tend to use the term more broadly to refer to processes that assess an algorithm's "*impact on the rights and interests of stakeholders, with a corresponding identification of situations and/or features of the algorithm that give rise to these negative impacts*" [Brown et al., 2021, p.2]. Audit processes have largely been conceptualised in fair-ML research as either internal, meaning that the process is designed to be run within the organisation operating an ML system, or external, meaning that the process is designed to be run by a third party, with or without the cooperation of the ML system operator [Raji et al., 2020; Krafft et al., 2021]. Some audit process proposals have, however, resisted this dichotomy, and sought to establish processes that are conducted by a mix of internal and external actors [Morley et al., 2021; Wilson et al., 2021].

As the definition above indicates, fair-ML audit processes have generally focused on evaluating algorithms, or more specifically, potential bias or discrimination in the dataset used to train an ML model and the ML model itself. To do so, fair-ML audit processes must grapple with issues of algorithmic opacity [Cobbe et al., 2021]. Both the training dataset and the ML model may be commercially sensitive, and the training dataset may contain personal data of ML system subjects. In these contexts, internal audit processes have been proposed as a workable solution, enabling ML system operators to provide some level of public assurance regarding the performance of their ML system, without compromising privacy or commercial interests [Raji et al., 2020]. However, it is unclear what the utility of an internal audit is, as a tool for supporting the enactment of ethical considerations, in contexts where there are no external requirements for the audit process or audit reporting [Rességuier and Rodrigues, 2020]. For this reason, fair-ML researchers note that internal audit processes should be thought of as only one part of a broader suite of tools for supporting, and

enforcing, substantive enactment of ethical considerations [Raji et al., 2020].

External audit processes have been proposed as a response to algorithmic opacity in contexts where those conducting the audit do not have privileged access to the training dataset or the ML model, and so must attempt to reverse engineer the ML model in order to understand potential biases or discrimination in the system's performance [Wilson et al., 2021]. External audit processes systematically probe the outputs of an ML model through inputs designed to reflect social groups who may be subjects of the ML model – doing so may reveal that the ML model is biased in favour or against people from particular social groups [Abebe et al., 2020]. External audit processes have been used widely in the fair-ML community to analyse the performance of ML systems deployed in high-stakes domains, such as healthcare [Obermeyer et al., 2019] or the criminal justice system [Angwin et al., 2016] (see [Bandy, 2021] for a recent literature review of the biases and discrimination documented by external audit processes). External audit processes have also been proposed as tools to empower ML system subjects to participate in the enactment of ethical considerations regarding ML systems [Krafft et al., 2021]. It is unclear, however, how external audit processes relate to the enactment of ethical considerations by actors within an ML system operator [Raji et al., 2020]. Indeed, a fundamental challenge for external audit processes is that they are generally designed to be conducted after an ML system has been deployed, and therefore after social harms have potentially accrued [Morley et al., 2021].

As tools to support the enactment of ethical considerations, four limitations of existing proposals for audit processes have been identified in the fair-ML literature. First, audit processes are single point-in-time snapshots of the performance of ML systems, but ML systems may be continuously developed and updated, and audit processes are ill-equipped to make predictions about future performance [Morley et al., 2021; Raji et al., 2020]. Second, audit processes need to balance ML system subjects' right to privacy, which may be established by consumer data protection laws, and ML system operator's right to protect trade secrets, with the integrity of the audit [Morley et al., 2021]. Third, the audit process may rely on narrow technical definitions of fairness, bias or transparency that are unable to account for the full breadth of potential harms caused by an ML system [Brown et al., 2021]. Fourth, audit processes that focus on evaluating the performance of training datasets and ML models, in other words the algorithmic components of an ML system, may assume that ML model outputs are implemented deterministically, and may therefore fail to consider the role of social components in an ML system, such as ML system users who inherit and respond to ML model outputs and mediate their impact on ML system subjects [Brown et al., 2021]. Taken together, these limitations may produce ML system audits that serve as performative rather than substantive enactments of ethical considerations (see [Perrow, 2011] for a similar critique of risk assessment methodologies). Nonetheless, audit processes are a significant category of intervention in the enactment of ethics considerations due to their adoption by several large technology firms (e.g. see [Raji et al., 2020] for a description of Google's internal audit framework) and incorporation into formal guidelines for public sector deployment

of ML systems [Morley et al., 2021].

2.7.3 Improving design and development processes

Audit processes generally support the enactment of ethics considerations after the ML system development process has concluded [Brown et al., 2021]. Fair-ML researchers have also developed several proposals for interventions in the enactment of ethics considerations during ML system design and development. Morley et al. have developed a typology of ethics tools [2019b] that maps proposed interventions to stages of a classic ML system development framework. A current version of the typology is maintained at https://tinyurl.com/appliedaiethics. The stages of ML system development addressed by the typology are: business and use-case development, system requirements development, data set procurement, building the system, testing the system, deployment, and monitoring [Morley et al., 2019b]. Whilst the scope of some interventions extends to the entire design and development process (e.g. see [IEEE, 2019; Richardson et al., 2021; Morley et al., 2021], in this section we focus on more narrow interventions in specific stages of ML system design and development, as the actual implementation of these in industry has been better documented in the fair-ML literature.

Two of the most widely cited interventions in the design and development process are Datasheets [Gebru et al., 2018] and Model Cards [Mitchell et al., 2019]. These interventions were developed by fair-ML researchers at Google, and have been implemented by several technology firms, including in the HuggingFace ML platform [HuggingFace] and the OpenAI API (accessible at https://beta.openai.com/ with an accredited account). Datasheets and Model Cards are templates for documenting key properties of ML training datasets and ML models. Datasheets document the motivations informing the creation of a ML training dataset, the contents, structure and limitations of a dataset, the process by which data was collected and curated, any pre-processing of the data that has occured, the intended use of the dataset, the planned distribution of the dataset, and any planned maintenance of the dataset [Gebru et al., 2018]. Similarly, Model Cards document the origins of an ML model, the intended use of the model, factors used to evaluate and design the model, metrics used to measure model performance, the dataset used to evaluate the model, the dataset used to train the model, any analysis of the model's biases, ethical considerations relevant to deploying the model, and caveats on such deployment [Mitchell et al., 2019]. As such, Datasheets and Model Cards are intended to support ML practitioners to enact ethical considerations during the data set procurement, building the system, testing the system, and deployment of the system stages of ML system development. In particular, the interventions are intended to help ML practitioners developing datasets or ML models to identify and document boundaries in which the use of a given dataset or ML model is most likely to align with societal values or expectations – and to support ML practitioners who may become consumers of such datasets or ML models to make informed decisions about their application [Gebru et al., 2018; Mitchell et al., 2019]. In this way, Datasheets and Model Cards can

be thought of as interventions designed to address issues of algorithmic opacity, by encouraging better documentation during the ML system design and development process, and normative judgments on contested terms, by encouraging boundary setting for potential uses of a dataset or ML model.

To the best of our knowledge, whilst the proposals for Datasheets and Model Cards have been evaluated from the perspective of ease of implementation they have not been evaluated from the perspective of impact on the outcomes of ML system design and development. From our perspective, it seems possible that, as tools to support the enactment of ethical considerations, Datasheets and Model Cards may be limited in two ways. First, to put it crudely, the purpose of a design process is to do *something* – design processes in software engineering are rarely focused on identifying reasons not to proceed with the development of a software solution [Turilli, 2008], despite the fact that there may be instances where ethical considerations indicate that a software solution is inappropriate [Abebe et al., 2020; Moor, 1979]. From this perspective, Datasheets and Model Cards, as interventions that operate within the design and development process, may enable only a narrow range of ethical considerations to be enacted. Secondly, Datasheets and Model Cards seem to implicitly assume that the cause of poor or ineffectual enactments of ethical considerations during ML system design and development is lack of information or support for ML practitioners. To this end, Gebru et al. describe the value of transparency as "*necessary for dataset consumers to be sufficiently well informed that they can select appropriate datasets for their tasks and avoid unintentional misuse*" [2018, p.2]. This assumption positions ML practitioners as benevolent actors with complete agency over ML system development. In practice, however, ML practitioners are likely to be influenced by the social context in which they work, and this social context may include pressures on ML practitioners to make decisions that amount to intentional misuse. For example, ML practitioners may be pressured by their supervisors to use a pre-trained ML model rather than develop their own model to save development time and costs. In the Safety Science literature this pressure is described as production pressure, and is considered a critical challenge to the enactment of safety protocols in organisational settings [Hinze and Parker, 1978; Wright, 1986; Vaughan, 1989]. From this perspective, the effectiveness of Datasheets and Model Cards, as interventions that focus narrowly on the role of ML practitioners, may be constrained by the organisational dynamics within which ML practitioners operate. These limitations inform the emphasis in our research design on exploring the enactment of ethical considerations by ML practitioners in situ, and on processes for evaluating the effectiveness of proposed interventions in the enactment of ethical considerations.

2.7.4 Software packages for measuring fairness

In recent years several software packages have been developed to support ML practitioners to enact ethical considerations. These software packages are tools for improving design and development processes, but merit their own description due to their relationship to ML platforms. In particular, software packages for assessing

the fairness of an ML model and biases in datasets have been incorporated into the ML platform operated by Amazon [Simon, 2020] and made open-source by Microsoft [Bird et al., 2020]. These software packages have also been evaluated by fair-ML researchers, who find that ML practitioners need training support to robustly implement them [Holstein et al., 2019; Kaur et al., 2020].

In future work we intend to closely review existing software packages for enacting ethical considerations. The incorporation of these software packages into ML platforms highlights the potential significance of ML platforms as sites for the enactment of ethics considerations during ML system development. ML platforms could serve as central distribution points for interventions in the enactment of ethics considerations. More ambitiously, ML platforms could mandate the use of certain interventions in sensitive use-cases. To this end, the Microsoft Azure platform requires ML platform users to declare that they have reviewed and will abide by Microsoft's Responsible AI guidelines in order to access certain ML systems [Microsoft, b].

In this section we introduced four categories of fair-ML interventions in the enactment of ethics considerations. In the following section, Section 2.8 we consider how these interventions may be reconciled with the four aspects of ML platforms we introduced in the opening section of this review, Section 2.2.

2.8 Applying fair-ML ethics interventions on ML platforms

In the previous section, Section 2.7, we reviewed four categories of proposed interventions in the enactment of ethical considerations during ML system development. These categories were: governing principles; audit processes; interventions in design and development processes; and, software packages for enacting ethical ethical considerations. In this section we consider how each of these categories of proposed fair-ML interventions account for the role of ML platforms in ML system development. Our analysis in this section is speculative in nature. To the best of our knowledge, there is no existing fair-ML literature that considers the relationship between ML platforms and the enactment of ethics considerations by ML practitioners. We structure our analysis around the four aspects of ML platforms described in Section 2.2: their centralising, scaling, refocusing, and decentralising effects. Our aim in this section is to demonstrate that ML platforms may complicate the effectiveness of fair-ML interventions in ML system development – a claim we propose to substantively examine through our future research.

2.8.1 Governing principles and ML platforms

The scope of governing principles adopted by technology firms shifts when those firms operate large ML platforms. Should the governing principles of a technology firm extend to constraining the way its ML platform can be used to develop ML systems?

The centralising effect of ML platforms empowers ML platform operators to act in a governance capacity; by extending governing principles to cover ML platform users, ML platform operators can potentially significantly increase the potential for governing principles to support widespread enactments of ethics interventions in ML system development. Meanwhile, the scaling, refocusing, and decentralising effects of ML platforms may complicate the operationalisation of governing principles. Governing principles reflect values and interests of the technology firms who create them. The scaling effect of ML platforms contributes to the global reach of the technology firms who operate them. How can governing principles account for the global scale of ML platforms, and for differences in societal values and expectations for ML system development from place to place? The refocusing and decentralising effects, by expanding who can act as an ML practitioner and enabling organisations of all sizes to develop and deploy ML systems potentially reduces the relevance of governing principles. Governing principles that only constrain the behaviour of ML practitioners working at large technology firms may be of little relevance in contexts where the ML platforms these firm operate enable organisations to 'shop' globally for ML practitioners to engage. In such contexts, it may be that the governing principles of organisations commissioning the development of ML systems are more significant for the enactment of ethics considerations than those of organisations providing the tools to enable ML system development.

2.8.2 Audit processes and ML platforms

The requirements for, and obstacles to, implementing an audit of an ML system likely change when that system is developed and maintained on an ML platform.

The centralising effect of ML platforms may make it possible for ML platform operators to audit ML systems developed on their platforms. Similarly, the centralising effect may create new opportunities for external audits by enabling third-parties to use ML platform APIs to systematically probe ML systems developed on the platform. Indeed, an emerging category of external audit is the cross-platform audit, where the performance of ML models built into ML platforms is systematically tested and compared [Isdahl and Gundersen, 2019; Roy et al., 2019; Kyriakou et al., 2019]. On the other hand, by disaggregating ML system components, the decentralising effect of ML platforms may make it more challenging for ML system developers to audit their own systems, and may present new obstacles for third-parties seeking to access data or technical components needed to complete an external audit process. Meanwhile, the refocusing effect of ML platforms may increase the importance of audit processes; ML practitioners who assemble ML systems through combinations of pre-trained ML models, rather than training their own models, may not be aware of biases or limitations in training datasets, or may rely on audit reports furnished by the providers of pre-trained ML models to determine whether their use case is appropriate. At the same time, such audit reports on pre-trained ML models may not be able to account for the potential for assembled systems, as a whole, to behave in unexpected ways. Similarly, the scaling effect of ML platforms also may increase

the importance of audit processes; ML systems that scale globally will need to grapple with the limitations of how representative the training or evaluation datasets are of a global population of ML system subjects. Internal audit processes may support this sort of reflection and enactment of ethical considerations.

2.8.3 Interventions in design and development processes and ML platforms

As we discussed in Section 2.2, each ML platform envisions the ML system design and development process occurring in a particular way. When ML practitioners choose an ML platform they are, in effect, committing to the system design and development process around which that platform has been conceived. As such, fair-ML interventions in design and development processes may be impacted by the degree to which the overarching design and development process they imagine aligns with the design and development process envisioned by ML platforms.

The centralising effect ML platforms have on ML system design and development can facilitate the adoption of fair-ML interventions. The two interventions in the design and development process we discussed in Section 2.7.3, Datasheets [Gebru et al., 2018] and Model Cards [Mitchell et al., 2019], for instance have been embedded in the HuggingFace platform [HuggingFace]. However, the centralising effect of ML platforms can also diminish the effectiveness of fair-ML interventions. As an illustrative example, the Model Cards template embedded in the HuggingFace platform is less detailed than the template described by the intervention's designers. As such, it may be the case that the way in which HuggingFace has implemented Model Cards significantly reduces their effectiveness as tools for supporting ML platform users to enact ethics considerations.

The centralising effect of ML platforms can further enhance the adoption of fair-ML interventions where those interventions are incorporated in the training materials produced by ML platform operators. Microsoft's online learning tools, for instance, include modules on responsible use and deployment of ML systems and promote Microsoft's design guidelines [Microsoft, c]. Meanwhile, the decentralising effect of ML platforms may increase the risk that fair-ML interventions in design and development processes are misused, or are used only to support limited enactments of ethics considerations, as ML practitioners may feel that their services are easily replaceable, and therefore may be wary of alerting organisations who commission ML systems to potential ethics issues. Similarly, the scaling effect of ML platforms may reveal previously unrecognised normative assumptions embodied in fair-ML interventions in the design and development process, which may limit the contexts in which proposed interventions can usefully be applied.

2.8.4 Software packages and ML platforms

As we discussed in Section 2.7.4, software packages for assessing the fairness of an ML model and biases in datasets have been incorporated into some ML platforms –

an example of leveraging the centralising effect of ML platforms to embed fair-ML interventions in ML system development. Embedding fair-ML software packages in an ML platform, however, may not result in uptake of the package by ML platform users. Software packages that were originally designed to meet the needs of ML platform users working within large technology firms may not meet the needs or be technically accessible to other ML platform user groups. In this regard, the scaling, refocusing, and decentralising effects of ML platforms may impact software packages for assessing fairness in the same way they impact fair-ML interventions in the ML system design and development process; the breadth of ML platform users and spread of social contexts in which ML platforms are used to develop and deploy ML systems may reveal previously unrecognised limitations in the effectiveness of a software package for assessing fairness, particularly in the definitions of fairness that such a package operationalises.

In this section we considered how fair-ML interventions designed to support the enactment of ethics considerations may be applied in the context of ML platforms. We hypothesized that ML platforms may act as sites that support or enhance the adoption of certain fair-ML interventions, whilst also frustrating or diminishing the adoption of other interventions. We highlighted, however, a gap in the fair-ML literature: the evaluation of proposed fair-ML interventions in the context of the tools used to support ML system development. Our proposed research questions, outlined in Section 1.3, aim to address this gap. In questions **R1a**, **R1b**, and **R1c** we explore how ML platforms are used by practitioners to support ML system development. In questions **R2a**, **R2b**, and **R2c** we then explore how practitioners who use ML platforms enact ethics considerations during ML system development. Finally, in question **R3a** we plan to explore how proposed fair-ML interventions can be evaluated in the context practitioners' use of ML platforms during ML system development.

2.9 Conclusion

Our Literature Review has now come full circle. We began this review by introducing ML platforms. We argued, in Section 2.2, that ML platforms are distinguished by the centralising, scaling, refocusing, and decentralising effects they have on ML system development. ML platforms centralise the locus of the process of ML system development in a small number of ML platform operators. The services provided by ML platforms enable ML systems to rapidly reach global scale, refocus the role of ML practitioners away from ML system maintenance and towards ML system design, and decentralise the social and technical components of ML systems. We noted in this section that, to the best of our knowledge, there have been no public analyses to date of ML platform users. In Section 2.3 we looked to literature on other technical systems that are also sometimes described as 'platforms' – two-sided markets, infrastructures, and implementation frameworks – to help us conceptualise the role of ML platforms in supporting and shaping ML system development. We argued

that these discourses could inform our study of ML platforms, but that, conceptually, ML platforms did not fit well into their existing treatments of 'platforms'. We thus concluded that the study of ML platforms required new theoretical work.

We then turned our attention to ML systems, and, in particular, to the ethics issues that have been associated with ML system development and deployment. In Section 2.4 we reviewed three categories of ethics issues associated with ML systems – algorithmic opacity, biased or discriminatory outcomes, and normative judgments on contested terms – and argued that the ongoing challenges of addressing these issues warranted our research focus on the processes and tools used in ML system development and deployment. In Section 2.5 we argued that ML systems have been conceptualised as socio-technical systems in the fair-ML literature, and applied this conceptualisation to ML platforms too. We concluded that understanding the relationship between ML platforms, ML platform users, and the ML systems that they develop may help reveal how the ethics issues associated with ML systems manifest.

We next shifted our attention to the way ML systems are developed. In particular, in Section 2.6, we introduced the concept of enactment to describe how ML practitioners engage, intentionally or unintentionally, in ethical considerations during the development of ML systems. We noted that the concept of enactment is not currently used in fair-ML discourse, although it is consistent with how the discourse has studied the practice of ML system development. In Section 2.7 we reviewed four categories of interventions in ML system development, which fair-ML researchers have proposed to support ML practitioners in their enactment of ethics considerations. For each of these categories of intervention – governing principles, auditing ML systems, improving design and development processes, and software packages for measuring fairness – we noted that an ongoing research challenge is evaluating their effectiveness.

Finally, we returned our focus to ML platforms. In Section 2.8 we considered how fair-ML proposed interventions account for the role ML platforms play in supporting ML system development. We noted that, to the best of our knowledge, the fair-ML discourse has not yet explored the relationship between the enactment of ethical considerations by ML practitioners and their use of ML platforms. Nonetheless, we reasoned that ML platforms could frustrate or limit the effectiveness of proposed interventions in ML system development, and we argued that such interventions ought be evaluated in the context of ML system development occurring on ML platforms.

In the following Chapter, Chapter 3, we introduce the conceptual frameworks we intend to use to explore how ML platforms relate to the enactment of ethics considerations by ML practitioners. In Chapter 4 we then detail the research questions we intend to focus on, and the research design we propose to undertake to address them.

Chapter 3

Conceptual frameworks

In the preceding chapter, Chapter 2, we developed the argument that ML practitioners enact ethics considerations during their development of ML systems. We highlighted that the development of ML systems involves the use of complex tools, particularly ML platforms, and we argued that these tools may influence the ways in which ethics considerations are enacted. Our focus on the enactment of ethics considerations by ML practitioners thus spans three levels of abstraction, ranging from a relatively high level to a relatively low level: the enactment of ethics generally; the enactment of ethics considerations in the context of ML system development; and, the enactment of ethics considerations in the context of ML system development occurring on ML platforms.

In this chapter, we outline the conceptual frameworks which inform our understanding of each level of abstraction. At the most abstract level of analysis, we draw on the conceptual framework of discourse ethics to explain our focus on enactment. At the intermediary level of abstraction, we draw on existing typologies for tools used to support ML practitioners to enact ethics considerations and a conceptual distinction between hard ethics and soft ethics. At the lowest level of abstraction, we draw on the theory of affordance to interpret the relationship between ML platforms and the enactment of ethics considerations by ML practitioners.

In the chapter that follows this, Chapter 4, we detail our research methodology and design, which apply the conceptual frameworks detailed here to address the research problem of understanding how ML platform users enact ethics considerations in the development of ML systems on ML platforms.

3.1 Discourse ethics

At the most abstract level, our research focus is the *enactment* of ethics. In Section 2.6 of our Literature Review we provided a working definition of enactment, and reviewed fair-ML literature that has studied the enactment of ethics considerations by ML practitioners. Our focus on the enactment of ethics is informed by the conceptual framework of discourse ethics, most closely associated with Jürgen Habermas [Fultner, 2014]. Discourse ethics is a mode of moral reasoning in which norms are justified through dialogue between all those who may be impacted by the

adoption of a norm [Ingram, 2010, p.125]. Whilst a full review of discourse ethics is beyond the scope of this research proposal (and will need to be undertaken in future work), two aspects of discourse ethics are particularly relevant.

Firstly, discourse ethics is procedural; discourse ethics does not seek to establish specific normative claims, but rather conceives of the practice of ethics as a process by which normative claims can be evaluated [Metselaar and Widdershoven, 2016]. This conception of ethics explains our focus on the enactment of ethics considerations by ML practitioners. We hold that the processes by which ML practitioners reflect, discuss, debate, and reach consensus on ethical dilemmas is of vital importance to the moral resolution of such dilemmas. We note, however, as the studies of enactment reviewed in Section 2.6 demonstrate, that in practice these processes may be informal and partial, or may unfold without ML practitioners ever recognising them as conscious enactments of ethics considerations. In other words, we assume that ethical dilemmas are unavoidable in ML system development, and as such ML system development necessitates their resolution, whether that resolution occurs through formal processes designed to support ethics considerations, or implicitly through the day-to-day processes of ML system development. Accordingly, we must study both the artefacts created to document ethics considerations – statements of AI principles, audit reports, checklists, notes in computer code, and so on – and the process of ethics consideration too. We must also avoid *a priori* assumptions regarding where within the process of developing ML systems do ethics consideration occur, or in what artefacts such processes may be documented.

Secondly, discourse ethics accepts that possibility of universal norms and provides procedural guidance for their ethical establishment [Fultner, 2014]. Habermas's universalization principle states that a norm is valid when " *the foreseeable consequences and side-effects of its general observance for the interests and value-orientations of each individual could be jointly accepted by all those affected without coercion*" [Fultner, 2014, p.120]. Given the defining feature of an universal norm is the breadth of its application, and the universalization principle's requirement that all those affected are able to participate in debate, discourse ethics proposes a very high procedural standard for justifying claims to universality [Mingers, 2011]. This conception of universal norms helps explain our focus on ML platforms as important sites for the enactment of ethics considerations by ML practitioners. As we discuss in Section 2.3, ML platforms are critical infrastructure for ML system development, and have a centralising effect on that development. As such, decisions made regarding the design of ML platforms have an universal aspect to them – they directly impact all ML platform users and indirectly impact all ML system subjects. Where such decisions are normative in nature they therefore merit the sort of close scrutiny envisioned by discourse ethics.

3.2 Frameworks for conceptualising ethics interventions

At a lower level of abstraction, our research focus is on the enactment of ethics in the context of ML system development. In this context, the enactment of ethics considerations is often facilitated by various interventions (see Section 2.7 for a review of some of these). At this level of analysis, we draw on Morley et al.'s typology of ethics interventions [2019b] and Luciano Floridi's framework for distinguishing between soft ethics and hard ethics [2018].

Morley et al.'s typology maps existing fair-ML proposals for interventions in ML system development to stages of a classic ML system development framework, and to different categories of ethics considerations. A current version of the typology is maintained at https://tinyurl.com/appliedaiethics. The typology is intended to help ML practitioners identify specific tools and methodologies they can use to support their enactment of ethics considerations and to help researchers to identify opportunities for the development of new ethics interventions [Morley et al., 2019b]. To this end, Morley et al. describe the interventions indexed by their typology as "*a pragmatic version of Habermas's discourse ethics*" [2019b, p.2152]. The use of a given ethics intervention does not guarantee a particular 'ethical' outcome, but rather can facilitate dialogue aimed towards establishing just norms for the development of a given technology. We intend to use Morley et al.'s typology to conceptualise the space of potential ethics interventions, and to inform the development of our own prototype tool.

Floridi's conceptualisation of soft ethics and hard ethics enables us to expand our conceptualisation of the enactment of ethics considerations beyond the application of ethics interventions. Applied ethics, argues Floridi, must grapple with the establishment of minimum standards of conduct and with best practice [2018]. Hard ethics is the realm of minimum standards, and is focused on the development of regulations and laws that specify and enforce minimum standards. Soft ethics is the realm of best practice, and is focused on the development of standards of conduct that go above and beyond those required by law. Floridi notes that the effective and just application of soft ethics depends on an appropriate foundation of hard ethics [2018]. Floridi's conceptualisation of soft and hard ethics also sits comfortably within the frame of discourse ethics; the regulations and laws of hard ethics can be compared to universal norms, and as such the procedural guidance offered by discourse ethics for justifying universal norms can be used to inform the practice of hard ethics. We intend to use Floridi's conceptualisation of soft and hard ethics to inform our exploration of the effectiveness of different ethics interventions at supporting meaningful enactment of ethics considerations. In particular, we will consider how to account for different soft ethics and hard ethics contexts in the evaluation of an ethics tool.

3.3 Affordance

Our specific object of study are ML platforms, and our subject of study are, in the first instance, the ML practitioners that use them to develop ML systems. At this

level of analysis we use the concept of affordance to interpret the relationship between ML platforms and ML practitioners. The affordances of an object are the link between the object's features or properties, and the outcomes of the object's use by subjects [Davis, 2020]. These links are conceptualised as "*relational processes among users, designers, environments, and things*" [Davis and Chouinard, 2016, p.2]. By recognising that both the materiality of objects and the human agency of subjects influence how affordances manifest in social contexts, the concept of affordance enables us to describe the way that an object enables and constraints its subjects, without assuming a deterministic relationship between the two [Davis and Chouinard, 2016; Davis, 2020]. Importantly, affordances are understood to be variable, rather than binary; affordances make certain outcomes easier or harder for subjects to achieve through use of the object, but do not necessitate a particular outcome [Davis and Chouinard, 2016]. In the context of ML platforms, we intend to explore how the particular design and features of an ML platform, in relationship with social, political, cultural, and economic context in which the ML platform is used by ML practitioners, facilitate easier and more substantive enactment of some ethics considerations, whilst obfuscating or frustrating the enactment of other ethics considerations. For example, as we discussed in Section 2.2, several ML platforms offer ML practitioners access to pre-trained ML models, which afford practitioners new modes of ML system development. The use of pre-trained ML models, however, may make it more challenging for practitioners to develop ML systems that are transparent to their system subjects, thereby also affording new forms of algorithmic opacity.

To provide theoretical and conceptual support for our exploration we intend to draw on Jenny Davis's work on the operationalisation of affordances [2020]. Davis describes mechanisms for how affordances link the features of technological objects to the outcomes of their use by subjects, and conditions by which affordances vary across different subjects and contexts. We note that other authors have conceptualised affordances differently (e.g. see [Robert et al., 2020], who apply a more narrow understanding of affordances as the ways an individual perceives and interacts with a system) and intend to explore these different perspectives in the course of our research. Davis proposes six mechanisms that contribute to the operationalisation of affordances: *request*, *demand*, *encourage*, *discourage*, *refuse*, and *allow* [2016; 2020]. These mechanisms are intended to serve as a tool to support critical analysis, rather than as a taxonomy of the features of technological objects – in other words, technical features do not correspond to just one mechanism, but rather each mechanism can reveal different aspects of a technical feature [Davis, 2020].

Requests and *demands* are mechanisms by which technological objects seek to direct how user-subjects interact with them. Technological objects *request* certain courses of action when they exhibit a preference for some actions over others; they *demand* certain courses of action when those courses are presented as the only possible actions for user-subjects to take. The HuggingFace ML platform, for instance, does not *demand* that users create a Model Card (a form of documentation), but does *request* this course of action by including prompts to complete a Model Card in the user interface for uploading a new ML model (see Section 2.7.3 for a description of

Model Cards). It should be noted, however, that whilst at the point of interaction between technological object and user-subject, *requests* and *demands* originate with the technological object, they are still socially constructed. *Requests* and *demands* are the end result of decisions made by system designers, and express their assumptions, values, and politics. Davis's conceptualisation of affordances is thus consistent with our description of ML platforms as socio-technical systems (see Section 2.5). *Encourage*, *discourage*, and *refuse* are mechanisms by which technological objects respond to the actions of user-subjects. Technological objects *encourage* certain actions by making them easy, or obvious, for user-subjects to complete and *discourage* others by making them hard or non-obvious to complete. Additionally, technological objects *refuse* certain actions when those actions are impossible or implausible for user-subjects to complete. *Encourage*, *discourage*, and *refuse* cover similar conceptual ground to *request* and *demand*, although shift the starting point of analysis from the materiality of the technical object to the action of the user-subject. Returning to the HuggingFace ML platform, should a practitioner who is uploading a trained ML model onto the platform wish to limit how other practitioners can use the model, perhaps to prevent the ML model being deployed in certain use cases, then they will find that the platform *discourages* them from doing so. The HuggingFace ML platform *discourages* practitioners from attempting to restrict how their uploaded models can be used by providing them with only two sharing options when they upload a new model: public or private. Thus practitioners can only choose between making their ML model available to every HuggingFace ML platform user, or none. It should also be noted, that technological objects are dynamic; actions that were once *refused* may become *encouraged*, actions that were once *encouraged* may become *discouraged*, and so on. The final mechanism, *allow*, refers to actions that technological objects neither *encourage* or *discourage*, *request* or *demand*. The HuggingFace ML platform *allows* platform users to access any dataset offered by the platform. Datasets are presented neutrally: the HuggingFace ML platform does not rate or review datasets, does not *encourage* platform users to download particular datasets or *discourage* platform users from downloading others. Davis notes that setting boundaries for the actions that technical objects *allow* users to take is often one of the most politically charged and challenging decisions for system designers [2020].

Davis proposes three factors that influence the conditions by which affordances vary across different subjects and contexts: *perception*, *dexterity*, and *cultural and institutional legitimacy* [2016; 2020]. Each factor reflects a different aspect of the agency of subjects, and shapes how the affordances of technical objects manifest differently across different social contexts. *Perception* refers to "*the extent to which a subject is aware of an object's functionality*" [Davis, 2020, p.91]. In effect, *perception* is required for an affordance to manifest in a particular social context; without perception, the material functionality of an object is useless to the subject. *Perception*, alone, is insufficient. The subject must also possess *dexterity*, the capacity to "*enact the functions of an object*" [Davis, 2020, p.94]. *Dexterity* influences which mechanisms of affordance a subject experiences when they interact with a technical object. An expert user, with a deep understanding of an object's technical functionality, may experience a

technical object as discouraging certain actions, whilst a novice user might experience the object as refusing the same actions. As this example makes clear, *dexterity*, and *perception*, are not fixed or innate characteristics of subjects, but rather reflect the nature of subject-object relations at a point in time. Lastly, *cultural and institutional legitimacy* refers to the "*way one's location within the larger social structure and the related norms, values, rules, and laws of a social system inform human-technology relations*" [Davis, 2020, p.97]. A subject's *cultural and institutional legitimacy* can expand or narrow the scope of what is allowed within a particular subject-object relation, and reconfigure whether actions are encouraged, discouraged, or refused. Thus, while a technical object may appear to treat users as a single homogeneous group, the affordances of a technical object are very rarely consistent across all users [Davis, 2020].

For the purposes of our research, Davis's conceptualisation of the mechanisms and conditions of affordance is crucial in enabling us to critically examine the relationship between ML platforms and ML platform users without assuming that ML platform users are an homogeneous group. The conditions of affordance provide a theoretical framework for exploring how ML platform users across different social contexts may relate to technical features of ML platforms differently, and thereby enact ethics considerations differently. The mechanisms of affordance, meanwhile, provide a theoretical framework for structuring our analysis of ML platform features themselves. In other words, the mechanisms of affordance provide a framework for expanding our analysis of the services offered by ML platforms, which we introduced in Section 2.2.1, whilst the conditions of affordance provide a conceptual and operational framework for deepening the analysis of ML platform user groups we undertook in Section 2.2.2.

In this Chapter we reviewed the three levels of analysis that are implied by our research problem – the enactment of ethics considerations in ML system development on ML platforms. For each level of analysis we introduced a conceptual framework that we intend to use to inform our analysis and structure our research. In the following chapter, Chapter 4, we discuss our methodological approach and introduce our proposed research design. When discussing our research design we will highlight, in particular, where the conceptual frameworks introduced in this Chapter have shaped our research decisions.

Chapter 4

Research methodology

In this chapter, in response to the research questions outlined in Section 1.3, we propose a three-phase sequential study [Teddlie and Tashakkori, 2009]. In Section 4.1 we discuss the theoretical and philosophical underpinnings of this research design. In Section 4.2 we describe our proposed research questions and the three phases of our research design in detail. In Section 4.3 we discuss issues of validity. In Section 4.4 we describe the ethics considerations that have informed our research design, and in Section 4.5 we describe the limitations of our proposed research design. In the chapter that follows this, Chapter 5 we detail our work plan for undertaking the research activities we describe in Section 4.2.

In the following sections we follow Teddlie and Tashakkori's distinctions between research *paradigms*, research *methodology*, and research *methods* [2009]. A research *paradigm* is a "*worldview, complete with the assumptions that are associated with that view*" [Mertens, 2003, p.139]. A research *paradigm* informs how researchers select research questions and the processes they use to answer them [Teddlie and Tashakkori, 2009]. The worldview offered by a research *paradigm* extends to philosophical questions (e.g. the nature of reality) and sociopolitical issues (e.g. whose interests should the research inquiry serve) associated with research [Teddlie and Tashakkori, 2009]. A research *methodology* is informed by a research *paradigm*, but is not the same thing; a research *methodology* is a "*broad approach to scientific inquiry specifying how research questions should be asked and answered*" [Teddlie and Tashakkori, 2009, p.27]. A research *methodology* therefore reflects a researcher's worldview and philosophical positions, and extends these to apply to issues associated with the design and conduct of research, such as criteria for selecting research methods and evaluating research quality [Teddlie and Tashakkori, 2009]. Research *methods* are "*specific strategies and procedures for implementing research design*", and reflect the research *methodology* held by the researcher [Teddlie and Tashakkori, 2009, p.27].

4.1 The methodology of mixed methods

Mixed methods is a methodology whereby the "*investigator collects and analyzes data, integrates the findings, and draws inferences using both qualitative and quantitative approaches or methods in a single study or a program of inquiry*" [Tashakkori and Creswell,

2007, p.4]. As such, mixed methods researchers aim to use qualitative (QUAL) and quantitative (QUAN) methods to reveal a deeper understanding of the object of study than may be possible with either approach alone [Creswell, 1999; Edmonds and Kennedy, 2020; Schoonenboom and Johnson, 2017]. Whilst the research design we introduce in Section 4.2 primarily adopts QUAL methods, with QUAN methods suggested as alternative or complementary data collection and analysis strategies, we locate the design within the context of mixed methods research due to the fact that mixed methods methodology has informed our research design choices.

Mixed methods researchers often adopt a pragmatist research paradigm, eschewing steadfast commitments to particular methodological approaches in favour of a commitment to using the methodological approaches that work best for the research question under investigation [Teddlie and Tashakkori, 2009]. Alternatively, some mixed methods researchers adopt a transformative paradigm [Mertens, 2007], in which methodological approaches are selected on the basis of their capacity to support those who have been historically marginalised to participate in the research [Jackson et al., 2018]. In this research proposal we adopt the pragmatist research paradigm. In this section we briefly describe and defend our adoption of the pragmatist research paradigm.

Pragmatism, as a research paradigm, "*offers a practical and outcome-orientated method of inquiry that is based on action and leads, iteratively, to further action and the elimination of doubt*" [Johnson and Onwuegbuzie, 2004, p.17]. The pragmatist paradigm distinguishes itself by focusing on the relative and contextual strengths of QUAN and QUAL research methodologies, and on how these strengths can be synthesised in response to the research questions under study [Teddlie and Tashakkori, 2009; Edmonds and Kennedy, 2020]. The practical, outcome-orientated, and iterative aspects of the pragmatist paradigm align closely with the software engineering practices of ML platform users. Software engineering is an iterative and outcome-orientated process, with software engineers trained to adopt a practical approach to problem solving and systems development, using both inductive and deductive reasoning to support systems development [Kruchten et al., 2019]. Further, ML platforms themselves are dynamic objects of study; the technical features and interfaces of ML platforms are rapidly evolving, as to are the social and infrastructural import of ML platforms. ML systems, similarly, are far from static; best practice in ML system design is fast changing, both as new ML technologies emerge and new use cases are identified. The pragmatist paradigm's acceptance of QUAL and QUAN methods, and its recognition of the social nature of research, enables the development of a research design that can respond to these ongoing changes in the objects of study. Additionally, as we hope through this research to improve the fair-ML community's conceptual and empirical understanding of ML platform users' engineering practices, and to influence the practices themselves, we can enhance the potential impact of our research by ensuring that the paradigm we adopt in designing our research can be reconciled with the paradigm implicitly adopted by ML platform users.

The pragmatist paradigm is particularly reflected in fair-ML literature focused on establishing the need for an inter-disciplinary approach to studying ML systems.

Abigail Jacobs and Hanna Wallach advocate for incorporating the concept of measurement modelling, which originates in the quantitative social sciences, into the field of computer science as a means of supporting fair-ML researchers to analyse fairness in computational systems [2021]. The structure of Jacobs and Wallach's argument reflects the pragmatist paradigm. They begin by observing a shortcoming in existing fair-ML research: the failure to recognise the potential for a mismatch between a theoretical understanding of an unobservable construct, such as teacher effectiveness, and its operationalisation in a measurement model based on observable proxy, such as student performance on exams [Jacobs and Wallach, 2021]. After illustrating the impact of this shortcoming, Jacobs and Wallach then introduce measurement modelling as a new conceptual tool that has proven to 'work' to address the shortcoming in other academic disciplines [2021]. Ben Green and Salomé Viljoen draw on two contrasting strands of legal scholarship, legal formalism and legal realism, to critique dominant modes of thinking about algorithms in computer science [2020]. The algorithmic realism that Green and Viljoen advocate for echoes the pragmatist paradigm: formal notions of objectivity and neutrality are supplanted with " *reflexive political consciousness*", and claims to universalism are supplanted with " *contextualism*" and recognition of the "*complexity and fluidity of the social world*" [2020, p.20]. Finally, the pragmatist paradigm is also implicit in some fair-ML studies where only a QUAL or QUAN method has been used to explore a propose intervention in ML system design. In these studies, in their analysis of the limitations of their work, researchers' have noted the value of a complementary study (i.e. a QUAN follow up to a QUAL study, or vice versa) [Harrison et al., 2020; Richardson et al., 2021]. We therefore conclude that, from the perspective of the fair-ML research community, it is reasonable for us to adopt the pragmatist paradigm in this research project.

Adopting the pragmatist paradigm, however, does not directly necessitate a mixed methods research design. Rather, the pragmatist paradigm necessitates reflection on the research questions to be focused on in a study, and creates the possibility that a researcher may conclude that their research questions will best be answered through a mixed methods research design [Teddlie and Tashakkori, 2009]. Reflecting the pragmatist paradigm's focus on developing research methods that are responsive to the research questions to be focused on in the study, Teddlie and Tashakkori [2009] and DeCuir-Gunby and Schutz [2017] both list several aspects of research questions which may indicate that a mixed methods research design is appropriate. Relevantly, to the research questions focused on in this proposal, these aspects include:

- Where research questions are both confirmatory and exploratory in nature, both QUAN and QUAL research methods can enable a researcher to develop a more holistic response to the research questions [Teddlie and Tashakkori, 2009, p.37]. QUAN methods enable a greater degree of generalisability, and as such are more effective at responding to confirmatory questions than QUAL methods [Teddlie and Tashakkori, 2009; DeCuir-Gunby and Schutz, 2017]. QUAL methods can enable participant voice, and in-depth analysis of participant actions, and as such are more effective at responding to exploratory ques-

tions [DeCuir-Gunby and Schutz, 2017]. The research questions focused on in our proposal are both confirmatory and exploratory in nature. We seek to establish that ML platform users are enacting ethics considerations as they develop ML systems on ML platforms, and seek to understand how ML platform users, their enactment of ethics considerations, and the development of ML systems are interrelated and impact each other.

- Where research questions are focused on examining a complex problem or social phenomenon, the use of QUAN and QUAL research methods can enable a researcher to incorporate more perspectives into their analysis of the phenomenon [DeCuir-Gunby and Schutz, 2017]. QUAN methods can be used to collect data on how a complex problem manifests at a population level, whilst QUAL methods can be used to explore how the problem relates to particular sub-populations [Teddlie and Tashakkori, 2009]. The research questions proposed for this study reflect an interest in how ML platform users generally interact with ML platforms, and also in how particular ML platform user groups enact ethical considerations during their interactions with ML platforms.

- Additionally, where the social phenomenon to be studied are potentially contested, the use of QUAN and QUAL research methods can create opportunity for greater divergence of views to be reflected in data collected [Teddlie and Tashakkori, 2009; Schoonenboom and Johnson, 2017]. QUAL research methods may reveal limitations in QUAN data collection, for instance, highlighting the potential for alternate interpretations of the phenomenon [DeCuir-Gunby and Schutz, 2017]. The research questions proposed for this study probe ML platform users' enactment of ethics considerations, in a social context where there may be significant personal consequences for platform users who fail to address ethical issues associated with the ML systems they develop. As such, it is possible that data collected through either QUAN or QUAL methods may not reflect ML platform users' true (and potentially limited) enactment of ethical considerations. Using both QUAN and QUAL, however, increases the possibility for divergent views to be collected and therefore for more nuanced analysis of the research questions.

Given our pragmatist perspective, and the nature of our research questions, we therefore begin our research design from the assumption that a mixed methods study may be most appropriate. In the following section we describe our research design, and the reasoning justifying it.

4.2 Research design

In response to the research questions we introduced in Section 1.3, and reflecting the pragmatist paradigm we outlined in Section 4.1, we propose to undertake a three-phase sequential study [Teddlie and Tashakkori, 2009]. In this section we describe

each phase of our proposed research design. As there is significant uncertainty regarding the impact of COVID-19 restrictions on our capacity to undertake certain data collection methods, and, additionally, as some of the data collection methods proposed may require support of ML platform operators, for phases one and two we describe preferred and alternative data collection methods.

Numerous typologies of mixed methods research designs have been proposed (e.g. see [Creswell et al., 2006; Teddlie and Tashakkori, 2006]). As discussed in Section 4.1, a core tenet of the pragmatic paradigm is that research designs should follow research questions. Given this, an exhaustive typology of mixed methods research is not possible – as creating one would require, first, an exhaustive typology of research questions [Teddlie and Tashakkori, 2009, p.125]. Accordingly, the role of a typology in the research design process is to support the researcher to reflect on the most appropriate research design for their research questions, rather than to constrain the researcher to selecting from a pre-determined list of design structures [Edmonds and Kennedy, 2020]. To inform the research design used in this proposal we draw on the typology provided in [Teddlie and Tashakkori, 2009] as it is relatively recent and intended to inform research design choices. We note this typology extends to cover QUAL, QUAN and mixed methods research designs, although because its primary focus is on enabling researchers to develop mixed methods designs it is referred to as a mixed methods typology.

Design Type	***Monostrand Designs***	***Multistrand Designs***
Monomethod designs	Monomethod monostrand designs	Monomethod multistrand designs
Mixed methods designs	Quasi-mixed monostrand designs	Mixed methods multistrand designs

Table 4.1: Reproduction of mixed methods typology in [Teddlie and Tashakkori, 2009, p.130]

Teddlie and Tashakkori's mixed methods typology conceptualise different mixed methods research designs in a two-dimensional matrix, with four decision points for researchers [2009]. The matrix is reproduced in Table 4.1. The horizontal axis in the matrix distinguishes between single phase and multi-phase research designs (labelled 'monostrand' and 'multistrand' in the matrix). The vertical axis in the matrix distinguishes between single method and mixed methods research designs (labelled 'monomethod' and 'mixed methods' in the matrix). The initial two decisions for researchers using Teddlie and Tashakkori's typology, then, are to determine the number of methodological approaches and number of phases to be employed in the research design – these decisions determine which cell within the matrix is most relevant [2009].

The overall focus of this research proposal is the problem of understanding how ML platform users enact ethics considerations in the development of ML systems on ML platforms. Our proposed research questions conceptualise this problem as

having three components:

- ML platforms: who are ML platform users? how does the use of ML platforms reconfigure the practice of ML system development?
- The enactment of ethics considerations: how do ML platform users enact ethics considerations? what form do ethics considerations take? and, how does enactment of ethics considerations impact ML system development?
- The evaluation of interventions supporting the enactment of ethics considerations: how can proposed fair-ML interventions designed for ML platform users to deploy in their enactment of ethics considerations be evaluated?

These three components imply a three-phase study, with phase one focused on ML platforms and ML platform users, phase two focused on the enactment of ethics considerations, and phase three focused on the evaluation of interventions in the enactment of ethics considerations. Below we describe the specific research questions associated with each phase of the study, and for each question discuss whether a QUAL or QUAN method is appropriate. We conclude that in each phase a QUAL method is most likely to be appropriate. We note, however, that depending on the success of our participant recruitment efforts, and decisions regarding the breadth of ML platforms to study in phase one, QUAN methods may also need to be considered. Accordingly, our proposed research design might be described, in Teddlie and Tashakkori's typology, as a quasi-mixed multistrand design [2009].

Each cell in the mixed method typology matrix contains within it a wide range of different research designs. To help distinguish between these different designs, Teddlie and Tashakkori use two additional criteria, which offer researchers two additional decision points [2009]. The additional criteria are: the type of implementation process to be used; and, the point of integration between QUAL and QUAN approaches.

The implementation process refers to the process by which the research phases will be conducted. Teddle and Tashakkori identify four different types of implementation process: parallel, sequential, conversion, and multilevel [2009, p.131]. Multilevel processes relate to processes designed to support collection of QUAN and QUAL data from multiple levels within an hierarchical social structure, such as a large organisation or social institution [Teddlie and Tashakkori, 2009]. Multilevel processes are not relevant to this research project, as we are focused on a broad category of ML practitioners, rather than a single organisational setting. Conversion processes use a research method whereby QUAN data is converted into QUAL data for analysis (qualitizing) or QUAL data is converted into QUAN data for analysis (quantitizing). Data conversion is generally contingent on the scale and depth of data collected, and is a complex process [Teddlie and Tashakkori, 2009]. As such, whilst conversion processes may be suitable for this research design, we feel it is sensible to start our research design from the assumption that we will not be able to collect sufficient data to enable conversion. In parallel processes QUAN and QUAL data collection and analysis phases occur in parallel, with the results merged for purposes of

comparison [Hesse-Biber and Johnson, 2015]. Parallel processes are indicated when a single research question is proposed to be answered using both QUAN and QUAL methods [Teddlie and Tashakkori, 2009]. Finally, sequential processes are those in which phases of research are conducted one after the other, with results from one phase informing the design of the next [Teddlie and Tashakkori, 2009]. The three phases of our research problem build on each other; each phase will be informed by inferences drawn from the analysis of data collected in the preceding one, which suggests a sequential implementation process [Schoonenboom and Johnson, 2017]. Additionally, as a sole researcher, undertaking a sequential research phases is far more feasible than undertaking parallel research phases, due to the ability to focus research efforts on a single method at a time [Teddlie and Tashakkori, 2009].

The criteria regarding the point of integration between QUAL and QUAN approaches refers to where in the research process QUAL and QUAN data are integrated into a single analysis [Teddlie and Tashakkori, 2009]. This criteria is used to distinguish between what Teddlie and Tashakkori term quasi-mixed designs, where the research design is primarily QUAL or QUAN with the non-primary approach used only to enhance findings of the primary approach, and mixed-method designs, where the research design does not prioritise either approach [Teddlie and Tashakkori, 2009, p.132]. As noted above, our research design is primarily dependent on QUAL methods, although we may need to apply QUAN methods to augment our data analysis in some phases. As such, we consider our research design to be quasi-mixed.

We now turn to consider our proposed research questions in detail, and whether QUAL or QUAN approaches are appropriate for each phase of our study. Figure 4.1 diagrammatically presents our proposed research design, which is also described in the subsections below. The alternative data collection methods discussed for phase one and phase two are presented in Figure 4.2, although this figure should not be taken to imply that use of an alternative data collection method in one phase of the research design necessitates it in another.

4.2.1 Phase one: ML platforms

The overarching focus in phase one of our research project is documenting and interpreting the role of ML platforms in ML system development, and on selecting an ML platform user group to focus phases two and three of our research design on. As such, the three research questions we propose to address in phase one are:

- **R1a**: Who are using ML platforms to develop and deploy ML systems?
- **R1b**: How do ML platforms conceptualise the ML system development and deployment process and integrate ethics considerations into that process?
- **R1c**: What kinds of ML systems, and ML system development, do the affordances of ML platforms support?

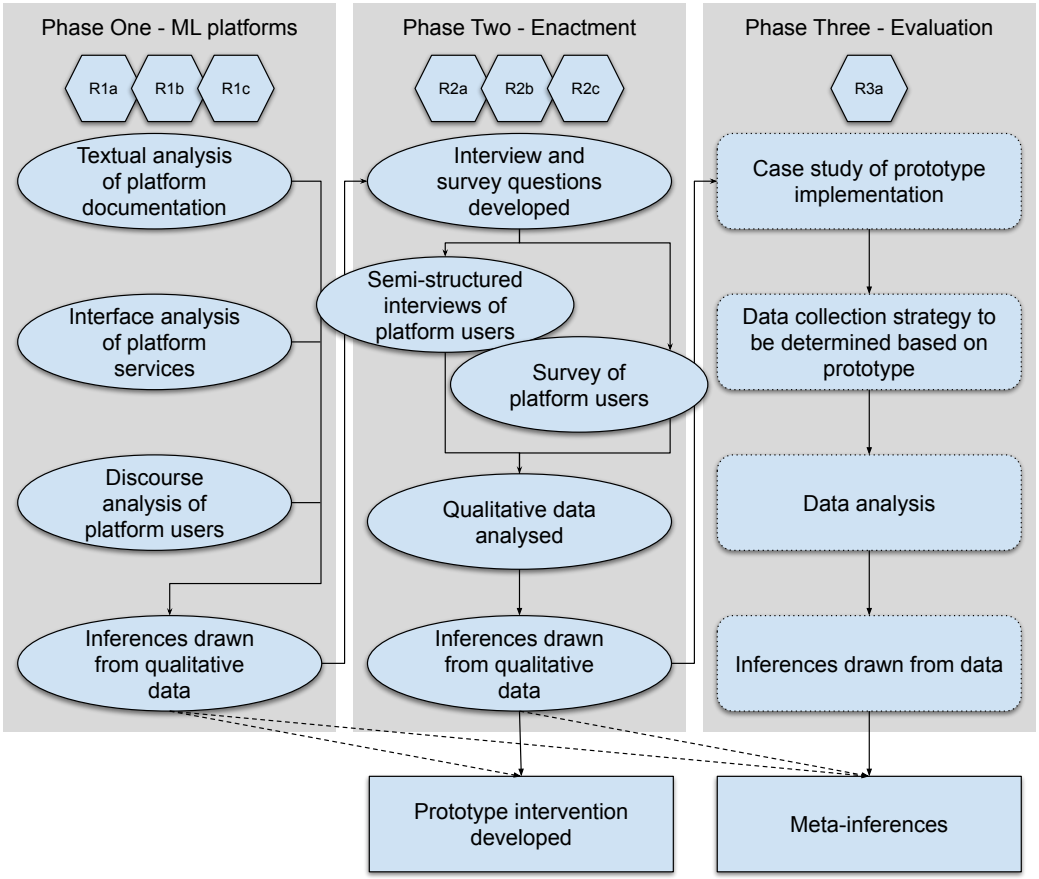

Figure 4.1: Diagram of proposed research design, and alignment with research questions.

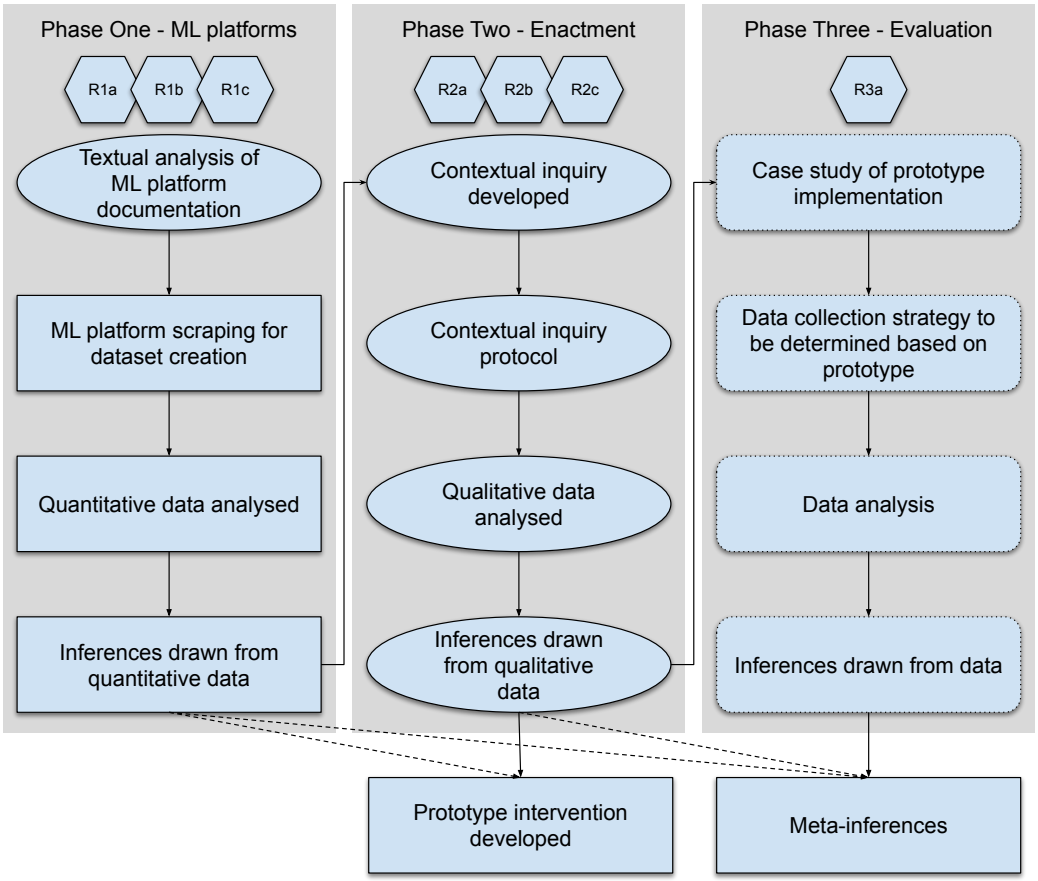

Figure 4.2: Diagram of proposed research design, featuring alternative data collection methods for phase one and phase two.

The exploratory nature of these questions indicates that they are best responded to through QUAL methods [Teddlie and Tashakkori, 2009]. From a conceptual perspective, these three research questions reflect Davis's framework for mechanisms and conditions of affordance [2020], which we described in Section 3.3. Davis notes that research methods designed to study affordances should attend to the political dynamics of technology use, to perspectives that may deviate from the 'norm' of how a technology is used, to the value-laden choices of the researcher, to the multiple meanings of technological artefacts, and to the relationship between material aspects of an artefact and its social meaning [2020, p.106-108]. Accordingly, in phase one of our research we propose to begin our study by critically examining the services and technical components of an ML platform, alongside materials produced by the ML platform operator to explain and promote the platform's use, with a view to understanding how ML platform operators imagine ML platform user groups and the M system development and deployment process. We have not yet identified the specific components of an ML platform to focus our study on. Our intention, however, is to focus on those components that distinguish ML platforms from their underlying cloud-based compute infrastructure, and that most directly impact the ML system development process. We also assume that ML platform operators offer only one perspective, and so additionally intend to consider how ML platform users and ML system development and deployment processes may resist or diverge from the 'norm' of ML platform use imagined by ML platform operators. We provide further detail on this approach below.

ML platform operators provide some explicit information on their ML platform users, in the form of technical documentation, marketing materials, and financial reporting. The services offered by ML platform operators are also suggestive of particular ML platform user groups. For instance, the Microsoft Azure AI platform incorporates a suite of services designed for the financial industry, including ML-based fraud detection tools, which suggest that practitioners working in the financial industry are a potential ML platform user group [Microsoft, a]. Further, where ML platforms incorporate two-sided markets for pre-trained ML models or curated datasets, the data produced by interactions on the two-sided market may suggest additional ML platform user groups. A preliminary review of the datasets available on the HuggingFace ML platform, for example, reveal corpora for numerous languages, including Arabic, Spanish, and Swedish, which also suggest potential ML platform user groups. To collect and analyse these various forms of data, we propose to undertake critical technocultural discourse analysis (CTDA) [Brock, 2018] of an ML platform, most likely one of: Microsoft's Azure AI platform [Microsoft, a], Amazon's AWS AI platform [Amazon, b], or HuggingFace's platform [HuggingFace]. The ML platforms operated by Microsoft and Amazon have been selected as options because both technology firms have produced governing principles for ML system development, and incorporated software packages for measuring fairness into their ML platforms. The ML platform operated by HuggingFace has been selected as an option due to the two-sided market it incorporates, and the platform's incorporation of Datasheets and Model Cards. See Section 2.2 for an overview of key features of

these platforms, and Section 2.7 for descriptions of each of the interventions listed above. We note, however, that final selection of an ML platform to focus on in phase one of our study will be dependent on our outreach efforts to field sites for phase three, as a practical consideration in ML platform selection will be ensuring that the ML platform we study in phase one are used for ML system development by the organisations who we study in phase three. In the following paragraphs we provide a rationale for our proposal to study only one ML platform in this phase of research, and for our selection of CTDA as our method of study.

We propose a deep analysis of an individual ML platform, rather than a comparative analysis between ML platforms for three reasons. First, our preliminary analysis, outlined in Section 2.2 is that there is a large degree of similarity between how ML platforms are structured, the services they offer, and their imagined users. Second, as ML platforms are tools that support the development of ML systems, a comparative analysis between platforms effectively has an exponential impact on the complexity of our research design: we would need to compare both the platforms themselves and the ML systems developed on them, both of which are dynamic and changing rapidly over time. Third, from a practical perspective, we note that CTDA is recognised as challenging to implement [Brock, 2018; Davis, 2020], and so believe that should we determine that a comparative analysis between ML platforms is needed, an alternative approach to data collection and interpretation for this phase of the study will need to be pursued.

CTDA has been developed by André Brock as a technique for studying information technology, particularly Internet-based computing systems [Brock, 2011, 2012, 2018; Davis, 2020]. In CTDA three aspects of computing systems are analysed at once: the materiality of the system – particularly, its interfaces; the production of meaning through use of the system; and, the co-evolution of users of the system, by the system's operators and users themselves [Brock, 2018]. CTDA allows for the iterative application of a conceptual framework. Brock, for example, applies critical race theory in two cycles of analysis [2012]. First, he analyses the "*material, practical and discursive properties of blogs, websites, and video games* – his objects of study [Brock, 2018, p.1013]. Second, he analyses the cultural practices of users of blogs, websites, and video games – the subjects of his study. This iterative approach translates particularly well to the research questions in phase one of our study, where we intend to use the conceptual framework of affordances first to study the material properties of an ML platform, and second to study the process of ML system development and deployment by ML platform users. The CTDA method also enables us to incorporate into our study the potential for ML platform users and ML platform operators to have divergent perspectives on how ML platforms are used, and for different ML platform user groups to have different norms of ML platform use. Divergences between ML platform operators and ML platform users can be revealed by studying the material properties of ML platforms alongside the cultural practices of ML platform users. Cultural practices may include, for example, workarounds or re-purposing that ML platform users adopt to adapt ML platform services to their needs. Differences in use between different ML platform user groups may also be revealed by

studying the cultural practices of ML platform users. Finally, CTDA also offers a path towards studying ML platforms that is not dependent on the support of an ML platform operator. The material aspects of the system are studied through interface analysis, in which the researcher carefully documents how users interact with the system by positioning themselves as a user. As ML platforms are publicly available systems, interface analysis can be undertaken without the need to access any privately held information. The use of the system is studied through analysis of materials produced by users, such as blogs or how-to guides, ML models or datasets shared on marketplaces, and code bases made publicly available. As these are also in the public domain, they do not require the support of an ML platform operator to access.

The specific components of CTDA may vary depending on the computing system to be focused on [Brock, 2018]. For the purposes of answering the three questions focused on in phase one of our study, we propose to conduct the below activities. Although these activities are presented in a list, we intend to undertake them in parallel.

- Responding to question **R1a** and **R1b**:
  - Textual analysis of the technical documentation, marketing materials, and financial reporting of an ML platform operator, and historical contextualisation of their development of an ML platform. Historical contextualisation here refers to an analysis of how a technical system reflects the broad social context in which it was developed. Section 2.3, where we briefly compared ML platforms to two-sided markets, infrastructures, and implementation frameworks, may form the basis of an historical contextualisation of ML platforms.
- Responding to questions **R1b** and **R1c**:
  - Interface analysis of the ML platform's services. Interface analysis treats a computing system as a text, to be read through detailed description of the interface that enables inputs and outputs in a computing system, and probing of the way in which an interface may vary between user groups [Brock, 2018]. We intend to conduct an interface analysis of an ML platform by subscribing to its services and using the platform to develop an ML system. This may include following an ML platform's online training program for using the platform.
- Responding to questions **R1a** and **R1c**:
  - Discourse analysis of how ML platform users describe their use of ML platforms. Discourse analysis treats the artefacts produced by users of a computing system as texts, that reveal the relationship between users and the system [Brock, 2011]. In our context, we intend to extend the application of discourse analysis by applying it to a sample of user-created guides

for using an ML platform (e.g on the Towards Data Science blog [Dille, 2019a,b]), and ML systems created by ML platform users.

From the above activities, we intend to produce a conceptual framework describing the effects of ML platform use on ML system development and deployment, which may expand or re-imagine the four aspects of ML platforms we highlighted in Section 2.2.1. Similarly we also intend to produce a conceptual framework describing ML platform user groups, which may correspond to or invalidate the four categories of ML platform users we described in Section 2.2.2. Our analysis of ML platform user groups will also inform the subsequent phases of our research, as in phase two of our study we propose to organise our research activities around a single category of ML platform users, rather than around a single ML platform.

4.2.1.1 Alternative data collection methods

Should the CTDA data collection and analysis methods we discuss above prove infeasible or inadequate, we may shift our data collection methods for phase one to include a QUAN analysis of ML platform users on an ML platform that operates a publicly accessible two-sided market. Alternatively, we may consider an approach where we select an user group to focus on in phases two and three of our study without first undertaking phase one.

As discussed in Section 2.3, several ML platforms offer web-based marketplaces for ML models or training datasets. The HuggingFace platform, most significantly, enables platform users to upload their own ML models or training datasets to the platform's market, and to download or otherwise incorporate other users' uploaded ML models or training datasets into their own ML systems, with a focus particularly on natural language processing [HuggingFace]. The HuggingFace platform surfaces a variety of data points to platform users regarding ML models and training datasets available on the platform, including number of downloads, and user-generated Model Cards [Mitchell et al., 2019] and Datasheets [Gebru et al., 2018]. The availability of these data points through a web-interface offers the possibility of using web crawling and scraping techniques to collect aggregate data on how ML platform users are interacting with ML platforms. We note, such techniques may require permission of the ML platform operator. Similar techniques have been used in studies of digital two-sided markets, such as AirBnB [Gyódi, 2019] and illicit Darknet markets [Ball et al., 2019]. **R1a** and **R1c** could thus be answered by contrasting the inferences drawn from an analysis of aggregate data on ML platform users interactions with an ML platform with the inferences drawn from the textual analysis we describe above.

We note that an additional option may to be adopt the QUAL walkthrough method, which is a method for generating detailed descriptions of the use of a computing system [Davis, 2020; Light et al., 2018], although it is outside the scope of this proposal to consider that in detail.

4.2.2 Phase two: enactment of ethics considerations

Phase two of our proposed research design focuses on the enactment of ethics considerations by ML platform users. The three research questions we propose to address in phase two are:

- **R2a**: How are ML platform users enacting ethics considerations in the development of ML systems?
- **R2b**: Are there particular fair-ML interventions that ML platform users rely on in their enactment of ethics considerations? If so, why do they rely on them?
- **R2c**: In what ways do the fair-ML interventions that ML platform users rely on, if any, shape the ML system development processes they undertake?

These three questions are focused on exploring the relationship between ML platforms, ML platform users, and the enactment of ethics considerations during the development of ML systems. As discussed in Section 2.6, to the best of our knowledge, the enactment of ethics considerations by ML platform users has not been explored in the literature to date. Given this, and the exploratory nature of these questions, we propose to respond to these questions through QUAL methods [Teddlie and Tashakkori, 2009].

For phase one of our research design we argued that it was conceptually and logistically justified to narrow the scope of our data collection and analysis activities to a single ML platform. For phase two, however, as the primary focus of our research questions are the behaviours of ML platform users, we do not propose to limit our study to one ML platform. Rather we propose to study only one ML platform user group, such as data scientists working in large technology firms, who may use multiple ML platforms. This strategy will enable us to undertake a deeper level of analysis on one group of ML platform users. That said, we note that there is a risk that different ML platforms may introduce into our study nuances not previously captured in phase one. As such, we will need to test this approach with pilot interviews before proceeding. For the subset of ML platform user groups we focus on, we will explore how ethics considerations are enacted through semi-structured interviews of a sample of user group members. To enable inferences drawn from our interviews to have some level of generalisability, we will also seek to substantiate them through a survey of ML platform users. Data collected from these interviews and the survey will address **R2a**, **R2b**, and **R2c**. These research activities will be supported by the hard/soft ethics framework and typology of fair-ML interventions discussed in Section 3.2; the typology of fair-ML interventions will be used to define the space of potential fair-ML interventions to explore during semi-structured interviews and the survey, whilst the hard/soft ethics framework will be used to probe the impact of different organisational contexts (i.e. different regulatory and cultural contexts) on the enactment of ethics considerations.

Combining semi-structured interviews of a small sample of a population of technology users with a survey of a larger sample is an established method within fair-ML discourse for collecting empirical data on technology use. In Section 2.6 we

reviewed four recent studies which explored the engineering practices of ML practitioners through such an approach [Holstein et al., 2019; Kaur et al., 2020; Veale et al., 2018; Orr and Davis, 2020]. Each of these studies interviewed between 11 and 35 ML practitioners, and then used interview results to inform a survey, which was administered to between 197 and 267 respondents. For phase two of our research, we propose to conduct semi-structured interviews with between 10 and 20 ML platform users from one ML platform user group, and to survey between 150 and 300 ML platform users. We note, however, that exact numbers will depend on how convergent interview responses are, and the corresponding number of interviews and survey responses required to reach saturation. As such, between 50 and 60 survey responses may will be sufficient [Dworkin, 2012].

Our participant recruitment strategy will depend on the user group we decide to focus on in this phase of our research, and our efforts to secure a field site for phase three of our research. The studies referenced in the previous paragraph largely recruited participants through purposive sampling [Edmonds and Kennedy, 2020], with the support of large technology firms, or public sector agencies, who were able to circulate invitations to participate to ML practitioners. In some cases, these recruitment methods were augmented by snowball recruitment strategies, under which participants in the research were also asked to nominate further participants [Holstein et al., 2019; Madaio et al., 2020]. At this stage of our research development, it is unclear whether we will be able to secure the partnership of a technology firm to assist with participant recruitment. We have had positive informal conversations with two technology consulting firms, based in Australia, who have indicated they may be supportive of assisting with participant recruitment, and additionally have built relationships with a community of ML practitioners who have indicated their interest in participating in such research – although whether or not they are members of our target ML platform user group remains to be seen. We expect the networking activities we propose in Section 5.3 to enable us to develop a more robust strategy for participant recruitment.

From the data collected in these interviews and surveys we will produce an analysis of how ethics considerations are currently enacted by one category of ML platform users. The significance of this analysis rests on the lack of existing data regarding enactment of ethics considerations during ML system development and deployment by ML platform users, as discussed in Section 2.6.

4.2.2.1 Alternative data collection methods

Should it prove challenging to enroll sufficient participants in semi-structured interviews, we may consider undertaking contextual inquiry with a smaller number of participants. Contextual inquiry protocols have been used in human-computer interaction studies to develop a detailed a picture of ML practitioners' interaction with technology tools [Kaur et al., 2020; Hohman et al., 2019] (a similar protocol is also used in [Richardson et al., 2021], although it is not described as contextual inquiry). The contextual inquiry protocol attempts to enable researchers to collect narrative

data that may usually only be accessible via participant observation studies, which in the context of COVID-19 restrictions on in-person work are particularly challenging to undertake in present times. The contextual inquiry protocol adopted by Kaur et al. featured a combination of semi-structured interviews, short questionnaires, and observation of participants interactions with a quasi-realistic scenario presented in an Jupyter notebook [2020]. In both Kaur et al. and Hohman et al. the contextual inquiry was centered on a technology probe [Hutchinson et al., 2003] created by the researchers for the purpose of enabling observation of ML practitioners undertaking a specific task [Kaur et al., 2020; Hohman et al., 2019].

Were we to apply a similar approach in phase two of this research project, rather than developing a technology probe we would seek to design a contextual inquiry protocol centred on practitioners' use of an ML platform in the process of developing an ML system. Whilst recruiting participants to join a contextual inquiry may be more challenging than recruiting participants for semi-structured interviews, given the additional time commitment required of participants, the depth of narrative data collected during a contextual inquiry would exceed that which we could collect during semi-structured interviews alone. As such, for this approach to be feasible we would likely need to design a contextual inquiry protocol with a particular ML platform user group in mind. Whilst generalising findings to the broader set of ML platform users will be challenging, the findings may provide an opportunity for a deeper level of analysis of how ML platform users enact ethics considerations.

4.2.3 Prototype intervention development

The inferences drawn from phase one and phase two of the research study will be used to inform the development of a prototype intervention for supporting ML platform users to enact ethics considerations in the development and deployment of ML systems. This prototype intervention will be the focus of phase three of our research design. In particular, we expect the inferences drawn from phase one and phase two of the study to help us identify:

- ML practitioner roles within ML platform user groups that are particularly important in the enactment of ethics considerations during the ML system development process.
- Stages in the ML system development process where ethics considerations are most front of mind for ML practitioners, or where existing fair-ML interventions are already widely used by ML practitioners.
- Interaction points in ML practitioners' use of ML platforms where ethics considerations may be salient, but where enactment of such considerations is currently not supported by extant fair-ML interventions.

As such, the development of a prototype tool or process represents the culmination of the inferences we produce in phases one and two of our proposed research design. Using these inferences, we intend to design a focused prototype intervention to

support the enactment of ethics considerations. Whilst we are wary of committing to particular prototypes prior to undertaking phase one and phase two of this study, illustrative examples of existing interventions are well documented within the fair-ML literature. We reviewed two examples, Datasheets and Model Cards, in Section 2.7.3. Further examples include:

- Processes for designing checklists to formalise ML practitioners' discussions of ethics considerations during ML system development [Madaio et al., 2020].
- Frameworks for ensuring accountability in the process of creating, maintaining, and exploiting a dataset for ML system development [Hutchinson et al., 2021].

The prototype intervention we develop will be used, effectively as a design probe [Hohman et al., 2019; Hutchinson et al., 2003], in phase three of our research project.

4.2.3.1 Alternatives to prototype intervention development

We recognise that developing a prototype intervention will be a significant undertaking. An alternative strategy may be to use our phase one and two findings to inform the selection of an existing proposed fair-ML intervention, which we could then focus on evaluating in phase three of our research design. An additional option may be to adapt an existing proposed fair-ML intervention in response to our phase one and two findings.

4.2.4 Phase three: evaluation of interventions

Phase three of our research design is focused on the evaluation of fair-ML interventions for supporting the enactment of ethics considerations. As such, the research question we propose to address in this phase is:

- **R3a**: How can an fair-ML intervention for enacting ethics considerations be evaluated, in the context of ML system development and deployment on ML platforms?

Phase three of our research design is the least developed, and will need to be refined significantly in response to our findings from phase one and two, and the specific prototype we develop. We note that **R3a** may need to be updated in order to ensure that it is tractable within the context of our PhD research. In particular, we expect to refine the research question based on the specific intervention we are focused on evaluating. In other words, we aim to study in phase three how one fair-ML intervention can be evaluated, rather than how fair-ML interventions generally can be evaluated. In future work, we may attempt to generalise our findings from phase three to the evaluation of fair-ML interventions more broadly.

At this stage, we conceive of phase three as being a case study of the evaluation of our prototype intervention, deployed in a small number of field sites. However, the type of prototype developed and the demands the intervention places on ML

platform users and field sites will determine the design of this research phase. For instance, an intervention which requires a significant time commitment from ML platform users may only be able to be evaluated within a single field site, suggesting evaluation can be undertaken through a participatory case study (an example of this is [Wilson et al., 2021], where the researchers evaluated a process for auditing ML systems). Alternatively, an intervention which requires less time from research participants may enable us to evaluate its effectiveness across multiple sites, enabling a comparative study of its performance. An additional consideration in developing phase three of our research design will be our outreach efforts to potential field sites, as any evaluation of an intervention will be dependent on us securing a field site in which the intervention can be deployed.

In this section, Section 4.2, we described our proposed three-phase research design. Phase one and two of the design were described in some detail, with alternative data collection strategies also reviewed. Phase three of the design was described only briefly, as it will need to be refined in response to the prototype intervention we develop, which itself will be informed by the cumulative findings from phase one and two of our research. In the section that follows, Section 4.3, we consider the validity of our proposed methods.

4.3 Validity

In the previous section, Section 4.2, we presented our proposed three-phase research design, and discussed the data collection and analysis activities we plan to undertake under each phase. In this section we address the validity of our propose research design. We note that in addition to demonstrating the validity of our research design, we will need to also demonstrate the reliability of our research methods. We do not, however, address reliability in this research proposal, as we intend to consider reliability in the next phase of our research development, when we begin to refine our research design and make commitments to specific methods for data collection and analysis in each phase of our research design.

In the context of research designs, validity refers to the extent to which the outcomes of research activity "*accurately answers the stated research questions of the study*" [Edmonds and Kennedy, 2020, p.4]. Validity is therefore a complex construction, conceptualised as a continuum, and not a binary feature of a research design [Edmonds and Kennedy, 2020]. As our research design progresses further, and we gain more insight into the feasibility of our planned activities, we will need to consider in detail issues of validity for each phase of our research design. For instance, in phase two, the validity of semi-structured interviews will depend, in part, on our capacity to recruit participants – a capacity we will only be able to evaluate once we have begun more concerted outreach and networking efforts. At this stage of our research development, we consider validity from two perspectives: validity in terms of usefulness of our research to our target academic community, the fair-ML

community; and, validity in terms of the level of generalisability that our proposed research design can aspire to.

In terms of the validity of our research design from the perspective of the fair-ML community, we have already observed, in Section 4.1, that the pragmatist paradigm we adopt is consistent with the nascent methodological commitments of the fair-ML community. We have also highlighted, in Section 4.2.2, that studies with comparable research methods to those we propose for phase two have been considered valid within the fair-ML community, and are often relied upon for the insights they provide with regard to ML practitioners' practices in industry settings. We therefore focus here on considering whether the research method we propose for phase one is likely to be interpreted as valid from the perspective of the fair-ML community. We note that we have not addressed issues of validity for phase three of our research, but contend that our proposed activities for that phase are currently insufficiently formed to merit a discussion of validity. At the same time, we recognise that, as one of the central objectives of our research is to explore how fair-ML interventions can be evaluated, issues of validity for phase three of our study are likely to require significant attention.

The CTDA method we propose for phase one of our research, described in Section 4.2.1, has, to the best of our knowledge, not been applied directly in fair-ML discourse to date. However, each component of the CTDA we propose for phase one – textual analysis of documentation, interface analysis of ML platform services, discourse analysis of ML platform users – can be found in existing fair-ML discourse. Close readings of technical documentation have been employed in critical studies of algorithms and ML systems [Mackenzie, 2015; Caplan and Boyd, 2018; Jo and Gebru, 2020], and form part of emerging practices for the ethnographic study of algorithms and socio-technical infrastructure [Seaver, 2017; Helmond et al., 2019]. Methods comparable to interface analysis have been applied in fair-ML studies of the impact of normative decisions during ML system development on system subjects [Albert and Delano, 2021], and in studies of the development of socio-technical infrastructures [Kelkar, 2018]. Finally, discourse analysis has been applied in studies of the way traditional computer science disciplines conceive of social constructs, such as gender [Keyes, 2018], and to studies of fair-ML literature itself [Barabas et al., 2020]. Given the overlap between CTDA methods and existing work in fair-ML and closely aligned fields, we believe the findings we produce in phase one of our study are likely to be considered valid from the perspective of the fair-ML community.

In terms of the validity of our research design from the perspective of generalisability, we noted in Section 4.2.1 and Section 4.2.2 two key design choices that influence how generalisable our research findings are likely to be: restricting our study in phase one to a single ML platform, and restricting our study in phase two to a single group of ML platform users. Restricting our study in phase one of our research to one ML platform may limit the potential for our findings to inform general claims regarding who are using ML platforms (question **R1a**), how ML platforms generally conceptualise ML system design and development processes (question **R1b**), and the kinds of ML systems supported by ML platforms (question **R1c**). That said, should

the CTDA design proposed for phase one prove to be effective at answering these questions for an individual ML platform, there may be opportunities in future work to extend its application to a comparative study of ML platforms. Additionally, it may be possible in future work to validate the conceptual framework for ML platform use, which we propose to develop during phase one, through QUAN surveys of ML platform users or ML platform operators. Restricting our study in phase two of our research to a single group of ML platform users may have a similar affect in terms of limiting the potential of our findings to inform general claims regarding the enactment of ethics considerations (questions **R2a**, **R2b**, and **R2c**). That said, as one of the assumptions we make regarding enactment, discussed in Section 2.6, is the importance of local context in shaping enactment, it is unclear to what extent any study of enactment in practice can be generalised. As such, the extent to which our findings in phase two can be generalised is likely to be dependent on the level of convergence or divergence we find in how research participants from our target ML platform user group enact ethics considerations.

In this section, Section 4.3, we considered, from a high-level perspective, the validity of our proposed research design. We argued that our research design is likely to be considered valid by our target academic community, but noted that its substantive validity will depend on future work we undertake to refine and further detail our proposed research activities. In the following section, Section 4.4, we provide a similar high-level overview of ethics considerations that have informed our research design.

4.4 Ethics considerations

In the previous section, Section 4.3, we discussed the validity of our proposed research design. In this section we discuss the ethics considerations that have informed our research design. We note that ethics oversight of research at ANU is governed by the *National Statement on Ethical Conduct in Human Research* (2017), which provides detailed guidance regarding ethics considerations for the development of research design, recruitment and consent of participants, collection and use of data, communication of research findings, dissemination of research outputs, and the conclusion of research projects. In future work, as we prepare for formal ethics review, we will need to consider each of these elements in detail. For the purposes of this thesis proposal, we focus primarily on ethics considerations that have informed the development of our research design, and our discussion of participant recruitment. Additionally, we do not discuss in this section ethics considerations regarding the alternative data collection methods we describe for phases one and two of our research design.

We note that ANU has adopted a three-level classification of ethical risk, which is used to determine the protocols required for ethics review. Our assumption, based on our application of the risk table provided by ANU [ANU, 2021], is that our re-

search design will fall into the second level of ethical risks and therefore need to be reviewed by a Delegated Ethical Review Committee following the Expedited Level 2 protocols, which are designed for research involving non-vulnerable participants in activities that have greater than negligible risks. The research participants in our proposed research design are ML platform users, who will be engaged through semi-structured interviews and surveys in phase two of our research. ML platform users will also be engaged in phase three, during our prototype intervention evaluation, although the mode of that engagement is still to be determined, and may include interviews and participant observation. As these research activities do not involve clinical trials, and do not expose participants to unusual research settings, we believe they present a low, but nonetheless greater than negligible level of risk. ML platform users, as discussed in Section 2.2.2, are generally software or data science professionals, working in commercial industry settings. We do not propose to intentionally recruit Aboriginal or Torres Strait Islander peoples. As such, we believe our proposed research activities do not involve participants that ANU's classification system would consider vulnerable.

The ethics concern informing our research design is maintaining, as much as possible, the confidentiality of ML platform users who participate in our research. This is a particular concern for phase two of our research, where we focus on collecting data on ML platform users' enactment of ethics considerations. We recognise that such data may include disclosures that could potentially embarrass research participants, or could jeopardise their relationships with current or future employers. A secondary concern is the potential for research participants to disclose the intellectual property of their employers, such as commercially sensitive aspects of their ML system development process. Incorporated into our proposed research design for phase two are the following measures, designed to protect participant confidentiality and reduce the risk of intellectual property disclosures:

- In all data collection activities, we avoid collecting data that would enable easy identification of participants and describe to participants how data collected will be used. We will only ask for category of workplace rather than specific workplace in surveys or interviews, as in some instances ML platform users may be the only ML practitioner in their workplace. Similarly, we will only ask for country-level location. We will collect gender and race information, as this may be relevant to how ML platform users enact ethics considerations, but will not collect date of birth, name, or any other personally identifying information. We note here that one trade off for us to consider is whether or not to collect participants' email addresses, and store these for a short period after data collection, so as to enable us to follow up with participants should their responses to a survey be unclear, or should we wish to invite them to participate in an interview.

- Additionally, in all data collection activities, we are particularly mindful of avoiding collecting data that would enable third-party identification of the workplaces from which participants are recruited. The categories of work-

places offered to participants in interviews and surveys will be high-level. In phase two, whilst we focus on only one category of ML platform users, we will nonetheless attempt to recruit participants from multiple workplaces, in multiple regions. We will similarly be mindful of collecting commercially sensitive data, and will remind participants to avoid disclosing to us the intellectual property of their employers.

- Finally, consent from participants will be sought for inclusion of any direct quotes or narrative data in thesis or papers prepared for publication. To the extent that it is possible, any identifying information will also be removed from quotes or narrative data.

We also note that the snowballing sampling method we describe for phase two can create conditions in which individuals feel pressured to participate in the research, and also requires the collection of email addresses. As such, we will need to reinforce the principle of voluntary participation in our survey and interview designs and recruitment materials, and consider whether it is appropriate to offer ML platform users incentives to participate in our research.

It may not be possible to mitigate the risk of third-party identification of research participants in phase three, where we propose to evaluate a prototype intervention. Were such an evaluation to occur, for example, through participant observation in a small technology consulting firm, then it may be trivial for interested parties to determine the identities of research participants. As such, we note that our selection of a field site for phase three will have significant consequences for our ethics considerations. In particular, it will be important to ensure that both the field site and individual research participants are informed of the risk of identification through their participation in the research. Further, it will be important to consider whether a potential field site sweeps into our study people who would be considered vulnerable by ANU's risk classification system, or data that is likely to be particularly sensitive. For instance, a small technology consulting firm working with a public sector agency to develop an ML system to be deployed in the criminal justice system is likely to have a different risk profile to the same firm working with a private corporation to develop an ML system to be deployed in their marketing department. As such, we expect to need to revisit our ethics considerations in future work, as we refine our list of potential field sites.

In this section we discussed our preliminary analysis of ethics considerations relevant to our proposed research design. In the following section, Section 4.5, we discuss the research limitations inherent in our proposed research design.

4.5 Research limitations

In Section 4.2 we described the three-phase study we propose to undertake in response to our research questions. In this section we discuss the research limitations

inherent in our three-phase study, building on our discussion of validity in Section 4.3 and ethics consideration in Section 4.4.

An overarching limitation inherent in our proposed research design is the potential for research findings to be generalisable across the full breadth of ML platforms, ML platform users, ML systems, and ML system users and subjects. In Section 4.3 we discussed this limitation in the context of research validity and our decision to study one ML platform and one ML platform user group. While there may be only an handful of ML platforms, each platform has near global scale, and as such ML platform users are a very large and heterogeneous category – they may be located anywhere in the world with high-speed internet access. Further, as ML platforms are designed to offer scalable infrastructure, they are used in firms of all sizes, across small, medium and large businesses, academic institutions, and government agencies. ML platform users are thus situated in a broad range of social, political, regulatory, and cultural contexts, all of which are likely to influence how ML platform users enact ethics considerations.

Our proposed research design is also limited in the breadth of perspectives which it enables to us to directly collect data on. In Section 1.2.2 we described four levels of 'users' who are relevant to our research problem, the enactment of ethics considerations during ML system development on ML platforms:

1 *ML platform operators*, who employ *ML platform developers* to develop ML platforms.

2 *ML system operators*, who employ *ML platform users* to develop ML systems on ML platforms.

3 *ML system users*, who interact with deployed ML systems.

4 *ML system subjects*, who interact with *ML system users*.

Our research design focuses on the role of ML platform users. Data on all other actors will only be indirectly collected during our research. As such, our research design is limited in its capacity to surface the perspectives of these actors. ML platform operators and ML platform developers are indirectly incorporated into phase one of our research, where we focus on exploring and contextualising the affordances of ML platforms through analysis of the documentation produced by ML platform operators and the services offered by ML platforms. ML system operators are also indirectly incorporated into phase two of our research, where we focus on understanding how ML platform users enact ethics considerations through semi-structured interviews and surveys, which will include questions regarding the role of social context (i.e. the role of ML system operators) in the enactment of ethics considerations. ML system users and ML system subjects are similarly indirectly incorporated into phase two of our research, as we expect to also probe ML platform users on whether the context in which an ML system is to be deployed influences how ethics considerations are enacted during ML system development. Limiting research participants to ML platform users risks reinforcing, through our research, a practitioner-centric

perspective on ML system development, which may fail to account for how ML system users and ML system subjects experience the enactment of ethics considerations during ML system development. We believe this concern is most pertinent to phase three of our research design, where we plan to explore how fair-ML interventions in ML system design and development can be evaluated. Clearly, it is necessary to incorporate the perspectives of ML system users and ML system subjects in an evaluation of such interventions. As such, we intend to revisit this concern in future work, where we focus on refining our phase three research activities. We also note that this concern is closely tied to the transformative paradigm, we briefly discussed in Section 4.1, which we also propose to consider further in future work.

Additionally, as we discuss in Section 2.2.1, ML systems themselves are increasingly ubiquitous, incorporated into a vast array of industries and use cases, and very diverse, with a large range of technical features. Our research design assumes, however, that ML system engineering processes are somewhat uniform – that comparable design and development processes are used across different categories of ML system development. If this assumption proves to be false, which may be revealed during our exploration of ML platform users in phase two of our research design, it may limit the scope of our findings regarding the enactment of ethics considerations. In particular, if ML system engineering processes are highly divergent within a single category of ML platform users, then it will be challenging to draw general conclusions regarding how ethics considerations are enacted during ML system development. Whilst such a finding may itself be of some use to fair-ML researchers, as it would suggest that fair-ML interventions in ML system design and development need to be highly contextually dependent, it may necessitate rethinking the scope and design of phase three of our research.

Finally, as we discuss in Section 4.2.4, where our research focuses on the evaluation of proposed fair-ML interventions, we will likely be limited to evaluating only one proposed or prototype intervention in a small number of field sites. As such, whilst our research may be able to define an evaluation process that is contextually appropriate to those field sites, it will be limited in its capacity to reveal how different fair-ML interventions can be evaluated across different contexts. Validating the evaluation process we use in phase three for general use in fair-ML research will need to be undertaken in future research projects.

In this section we described the ways in which our research design is limited. Two themes throughout this discussion were the challenge of incorporating diverse perspectives from people other than ML platform users, and the challenge of producing research findings that may be applicable in contexts other than those which we directly study. We note, however, that these limitations may also be thought of as directions for future work. In Chapter 2 we described a wide gap in the fair-ML discourse: a lack of theoretical or empirical knowledge of how ethics considerations are enacted during ML system development, and of how interventions in this enactment can be evaluated. As the limitations discussed in this section reveal, the research design we have proposed in this chapter can only hope to begin to address this gap.

In the following section, Section 4.6, we conclude this chapter.

4.6 Conclusion

In this chapter we introduced our proposed research design, which responds to the research gaps we identified in our Literature Review in Chapter 2. We discussed, in Section 4.1, the pragmatist paradigm we adopt and describe mixed methods methodology. Following a mixed methods approach, we argued that our research design should follow from our research questions. In Section 4.2 we applied this argument to our research questions. We distinguished between three components of our research problem – ML platforms, the enactment of ethics considerations, and the evaluation of interventions in the enactment of ethics considerations – and developed a three-phase sequential research design. For each phase we described relevant research questions, and then considered what research methods would most effectively address them. Ultimately, we proposed undertaking critical technocultural discourse analysis of an ML platform in phase one, semi-structured interviews and a survey of ML platform users in phase two, and, in phase three, a case study of the evaluation of a prototype intervention in a number of field sites. In the sections that followed the description of our research design we considered the validity (Section 4.3), ethics considerations (Section 4.4), and limitations (Section 4.5) of our research design. Throughout these sections we highlighted areas of ambiguity in our research design, and noted that all phases of our research design require refinement, and that phase three, in particular, is yet to be fully developed.

In the following chapter, Chapter 5, we discuss our thesis plan. Our thesis plan describes how we plan to implement our three-phase research design, including how we plan to further refine and develop each phase. Our thesis plan also considers our research design from the perspective of the objectives of our PhD program, detailing how we will document and aim to publish our research findings for the fair-ML community, our research resourcing requirements, and risks associated with our research design.

Chapter 5

Thesis plan

In this chapter we introduce our plan for undertaking the three phases of the research design we described in the preceding chapter, Chapter 4. In Section 5.1 we present a proposed timeline of activities, aligned to each research phase, in a Gantt chart. In Section 5.2 we discuss our target academic community, the fair-ML community, and note key institutions, conferences, and journals. In Section 5.3 we discuss our plan to build networks within the fair-ML research community, and with ML platform operators, and ML platform users. In Section 5.4 we discuss our plan to publish findings from our research. In Section 5.5 we discuss our skills and professional development needs and training plan. In Section 5.6 we discuss the resourcing requirements of our research design. Finally, in Section 5.7 we present a risk analysis for our thesis plan and mitigation strategies for key risks.

5.1 Projected timetable

Figure 5.1 presents a high-level Gantt chart of the overarching sequence of activities we propose to undertake. Squares filled green represent time periods where a particular research activity is planned to be worked on. Squares filled orange represent time periods that have been set aside should particular research activities require additional time (e.g. if a paper we submit is accepted for publication). Squares filled yellow represent our current focus. A more detailed Gantt chart has been provided to members of our supervisory panel, and can also be found using this hyperlink (access permission required). In general, reflecting the sequential nature of our research design, whereby each phase of research builds on the previous phase, the level of detail in our Gantt chart is highest for phase one of our research design and lowest for phase three and subsequent activities.

Figure 5.2 presents key dependencies on which the proposed sequence of research activities are reliant. In orange are highlighted dependencies on which research activities are reliant. In green are highlighted our proposed research activities, which assume the dependency is satisfied (e.g. that ML platform users can be recruited for a survey). As can be seen, dependencies largely relate to the recruitment of research participants or field sites. This presents the most significant risk to this research project, and is addressed in Section 5.7.

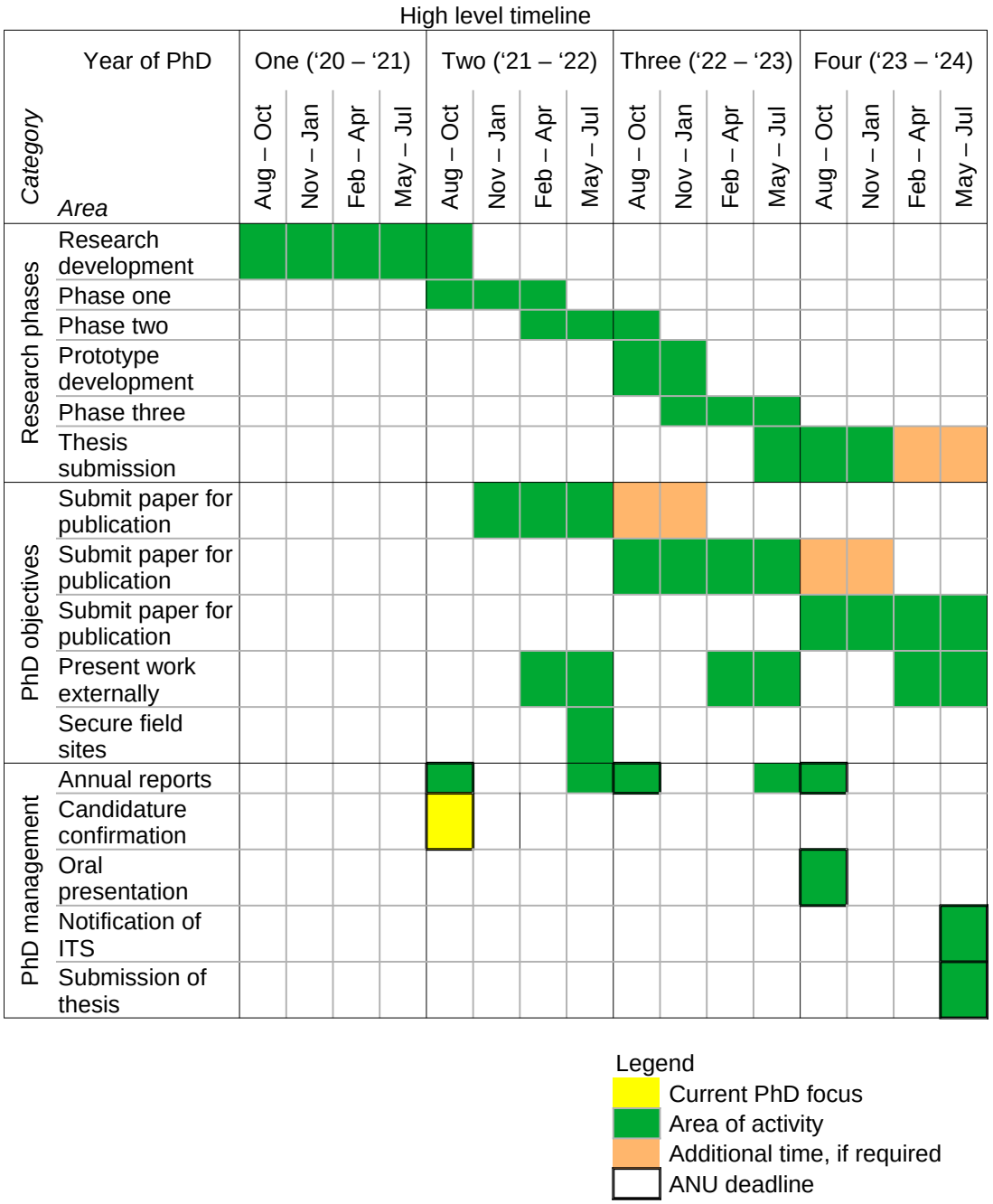

Figure 5.1: High-level Gantt chart of research activities.

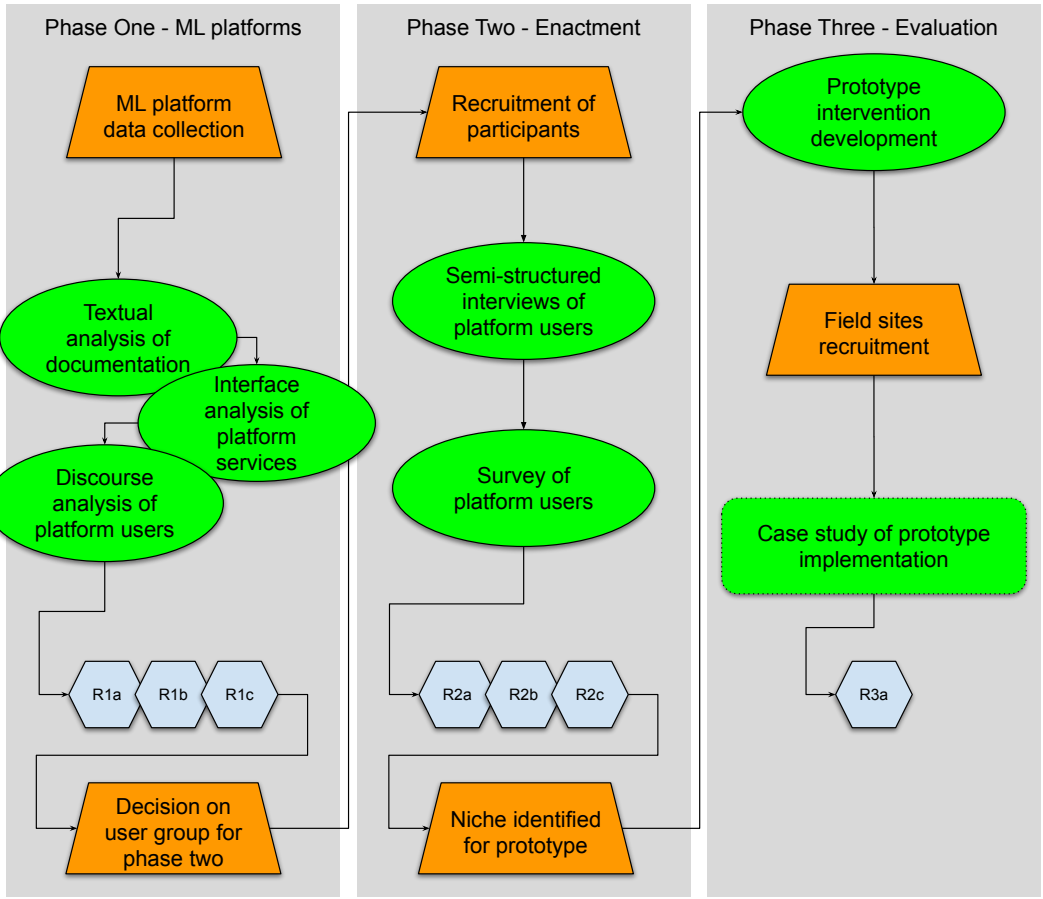

Figure 5.2: Research phases and their dependencies. Shapes coloured green represent research activities. Shapes coloured orange represents decisions or precursor activities that research activities are dependent on.

5.2 Target academic community

The target academic community for this research is the fair-ML community. The fair-ML community is still in its formative stages, and as such the institutions, conferences, and workshops that constitute the community's academic infrastructure are expected to change considerably during this research project's life cycle.

To the best of our knowledge, there are currently two standalone annual international fair-ML conference series: the ACM Conference on Fairness Accountability, and Transparency (ACM FAccT), and the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (ACM AIES). ACM FAccT is of particular relevance as a colloquium for PhD candidates is held alongside the conference. There is also a significant fair-ML track at the annual Web Conference (FATES on the Web). New conferences are regularly being announced – most recently, the ACM Conference on Equity and Access in Algorithms, Mechanism, and Optimization (EAAMO'21) and AI: Law, Ethics, Algorithms, and Politics (AI LEAP), which is hosted by the Humanising Machine Intelligence group at ANU. In addition, there are several annual international conference series focused on the relationship between technology and society, in which fair-ML community members are regular participants. These include: the IEEE International Symposium on Technology and Society (IEEE ISTS), the annual ACM CHI Conference on Human Factors in Computing Systems (CHI), the Society for Social Studies of Science Conference (4S), and NeurIPS.

Table 5.1 presents the schedule of the conferences listed above, although it should be noted that due to the COVID-19 pandemic there is significant uncertainty regarding the scheduling of 2021 and 2022 events. This schedule informs the conference periods and sequence of conference presentation activities detailed in the Gantt chart presented in Section 5.1.

Several academic and non-profit institutions are also closely associated with the fair-ML community. At ANU, two inter-disciplinary organisations have emerged with an interest in the fair-ML community: the 3A Institute, affiliated with the School of Cybernetics, and the Humanising Machine Intelligence (HMI) group. More broadly, within Australia, the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S Centre) launched in 2020 as a national research centre focused on responsible, ethical and inclusive automated decision-making. The ADM+S Centre collaborates with universities and non-profits across Australia, including the HMI group at ANU, the Centre for Artificial Intelligence and Digital Ethics at Melbourne University, the Emerging Technologies Lab at Monash University, and the Gradient Institute. As such, our assessment, informed also by conversations with Australian-based academics, is that the ADM+S Centre is likely to become the hub for the fair-ML community within Australia, and is therefore one of the focal points of our networking plan, discussed further in Section 5.3.

Internationally, there are emerging fair-ML network hubs in North America and Europe. In North America, Data Society is a particularly relevant non-profit research institute, as their AI on the Ground research track studies the materiality of ML system development. AI Now, affiliated with NYU, is a similarly significant

Conference	**Dates**	**Details**
ACM FAccT	Held in March 2021, 2022 dates TBC	Premier standalone fair-ML conference
ACM AIES	Held in May 2021, 2022 dates TBC	Premier standalone fair-ML conference
FATES	Scheduled for April 2022, and May 2023	Fair-ML track focuses on fair-ML issues and web-based technologies
ACM EAAMO	Scheduled for October 2021, 2022 dates TBC	A new conference series
AI LEAP	Scheduled for December 2021, 2022 dates TBC	New fair-ML conference series
IEEE ISTS	Scheduled for October 2021, 2022 dates TBC	Longstanding conference series on technology and society, pre-dates fair-ML
CHI	Scheduled for May 2022, future dates TBC	Premier conference on human-computer interaction
4S	Scheduled for October 2021, July 2022, and November 2023	Long standing Science and Technology Studies conference series
NeurIPS	Held annually, last week November/first week December	Premier conference for ML practitioners

Table 5.1: Table of upcoming fair-ML conferences.

research institute, as their research has previously included surveys of ML practitioners (see [West et al., 2019]) and they are particularly active in supporting research students. Additionally, the Berkman Klein Center for Internet and Society, at Harvard University, is a significant hub for research, particularly with regard to assessing the social impact of ML systems. In the private sector, Google and Microsoft have both invested in the fair-ML community, and offer various internship and research partnership opportunities for graduate students. The surveys of ML practitioners discussed in Section 4.2 ([Holstein et al., 2019; Kaur et al., 2020] are both products of Microsoft research programs, whilst several of the ethics interventions discussed in Section 2.7 ([Mitchell et al., 2019; Gebru et al., 2018; Raji et al., 2020] are products of Google research programs. In Europe, the Oxford Internet Institute, at Oxford University, is a significant hub for research on governance of the ML systems and digital infrastructure. The reviews of ethics interventions and ethical AI principles cited in Section 2.7 ([Morley et al., 2019a,b]) are products of Oxford Internet Institute research tracks. Additionally, the European Association for the Study of Science and Technology (EASST) is a longstanding association for Science and Technology Studies scholars and is an hub for sharing workshop, fellowship and post-graduate opportunities in Europe. Spanning Europe and North America, the Society for the

Social Studies of Science (4S) brings together Science and Technology Studies scholars globally, and has a dedicated graduate student program and student presentation track at their annual 4S conference. This review of international fair-ML community hubs, however, is largely the product of desk research. As such, in the networking plan discussed in the following section, Section 5.3, an important initial step is to validate the analysis presented in this section. We also recognise that this review has focused narrowly on North America and Europe, and note that fair-ML researchers in other regions are also developing new research hubs, which we hope to identify and connect with in future work.

In this section we described our target academic community, the fair-ML community. In the section that follows, Section 5.3, we describe how we plan to build relationships and networks in the fair-ML community, and with industry partners.

5.3 Networking plan

The networking plan we propose reflects our analysis of the fair-ML community, discussed in the previous section, Section 5.2, and the needs of our research design, discussed in Section 4.2. In the immediate future, our networking plan is focused on:

- Developing relationships with the ADM+S Centre and affiliated academics and industry partners, with the aim of securing ADM+S Centre graduate student membership; and
- Developing relationships with ML platform operators and ML platform users, with the aim of securing support for participant recruitment for phase two of our research design, and field sites for phase three.

Over the longer term, our networking plan is focused on:

- Building relationships internationally, with the aim of securing a graduate student fellowship at one of the premier international institutions discussed in Section 5.2; and
- Deepening relationships with graduate students and post-doc academics with similar research interests to our own, with the aim of identifying research and publishing collaboration opportunities.

We note that our networking activities will need to be responsive to opportunities as they emerge, and that each of the networking objectives listed above will need to be pursued concurrently. As such, we think of the networking activities described below more as a collection of possible directions to pursue than a rigid plan of actions. We also recognise that this networking plan is dependent on our publishing plan, described in Section 5.4, because writing high-quality work, and finding ways to make this work visible to the networks we describe in this section, will provide us with opportunities to develop and solidify new relationships.

5.3.1 ADM+S Centre graduate student membership

ADM+S Centre graduate student membership is a significant objective for us for two reasons: first, we anticipate that active participation with the ADM+S Centre's activities will enrich our research and serve as a useful bridge to relevant academics and other graduate students in Australia; second, we anticipate that ADM+S Centre graduate student membership will be an helpful credentialing exercise, in that academics, ML platform operators, and ML platform users in Australia and internationally are all likely to expect that a research project like ours would be affiliated in some way with the ADM+S Centre. However, as ANU is not an institutional partner of the ADM+S Centre, graduate student membership may not be possible. If that transpires to be the case, we will aim instead to build informal relationships with academics at the ADM+S Centre and seek a visiting researcher opportunity.

To obtain ADM+S Centre graduate student membership we require the formal support of a Chief Investigator at ADM+S Centre, which is most likely obtained through the support of an Associate Investigator. As such, the first networking activity we plan to undertake is outreach to Associate Investigators. To date, we have met informally with Jake Goldenfein, an Associate Investigator who has a strong interest in the fair-ML community and has previously published advice for public sector organisations purchasing ML systems [Goldenfein, 2019]. In addition, whilst ANU is not an institutional partner of ADM+S Centre, the HMI group at ANU are affiliated with it. Accordingly, we also plan to undertake outreach to HMI academics at ANU, particularly Jenny Davis, whose work on affordances [2020] has been a substantial influence on our research (see Section 3.3 and Section 4.2.1).

5.3.2 Industry support for phase two and three of our research design

Phase two and three of our research design cannot proceed without the support of industry; phase two is dependent on industry support for participant recruitment, and phase three is dependent on industry support for securing field sites to evaluate our prototype fair-ML intervention. As such, securing industry partners for these phases is a critical objective of our networking plan. Fortunately there are a large number of potential partners, including ML platform operators, as well as consulting firms who are formal technology partners with ML platform operators and develop ML systems, government agencies and large corporations who are ML system operators, and technology focused startups (see Section 2.2.2 for a more detailed description of these partners). Government agencies and corporations may commission consulting firms to develop ML systems for them, or may be developing ML systems with in-house ML platform users. We believe consulting firms and ML platform operators are the most high-potential partners, as these organisations have a more ongoing financial interest in best practice ML system development compared to government agencies and corporations, for whom ML system development may be a means to an end, rather than a core organisational focus. Additionally, while technology focused startups may have a financial interest in best practice ML system development, they are likely to be more resource constrained, particularly in terms of staff time, than

consulting firms and ML platform operators. That said, some technology focused startups are specifically focused on developing and marketing ML systems that are fairness-aware, and similarly some government agencies are particularly interested in ensuring that ML systems they operate meet societal expectations for fairness. As such, these startups and agencies may be interested in participating in our research (see [Sánchez-Monedero et al., 2020] for an example of a technology focused startup supporting fair-ML research and [Abuhamad et al., 2021] for an example of a government agency supporting fair-ML research).

Securing the support of an ML platform operator, however, may be challenging. ML platform operators, particularly Microsoft and Google, have formal processes for supporting student-led research, usually through participation in internship programs. Microsoft Research, for instance, runs annual PhD internship programs in the United States, China, and Europe. Students, however, are required to be enrolled in universities in one of those regions. Confidentiality requirements may also be an obstacle to publishing research undertaken with an ML platform operator. That said, Microsoft and Google both employ several high-profile fair-ML researchers, and our understanding is that these researchers have a certain degree of flexibility in terms of the PhD students they choose to support. We therefore believe the best approach to exploring the potential of an ML platform operator supporting our research is to focus on building relationships with fair-ML researchers already working at ML platform operators. These researchers (particularly Hanna Wallach, Madeleine Clare Elish, Jenn Wortman Vaughan, and Ben Hutchinson) are all active in the ACM FAccT conference series, and as such we plan to apply to participate in the Doctoral Colloquium at the 2022 FAccT conference, assuming that it goes ahead. We also plan to reach out directly to these researchers, although believe it is prudent to wait until after our Thesis Proposal Review to do so.

There are a large number of consulting firms who are technology partners with ML platforms. Whilst we are not constrained geographically, we assume that Australian-based consulting firms are more likely to be interested in supporting our research than firms based elsewhere. To date, we have tested interest in our research via informal conversations with senior team members at Eliiza, a Melbourne-based consulting firm who are technology partners for the Google and AWS ML platforms. The feedback we have received is that our research is of interest and relevance, and that firms would prefer to deepen their relationship with us through a short internship before exploring the opportunity to become a field site in phase three of our research. As such, we plan to draft a short internship proposal (3 week internship), which will be focused on deploying an existing ethics intervention in the consulting firm, and to circulate this to relevant Australian-based consulting firms. The objective of this internship will be to build relationships, rather than to collect data or contribute directly to our research design.

In addition to direct outreach to ML platform operators and consulting firms, we are currently focused on building relationships with ML practitioners who have an interest in the fair-ML community. We believe these ML practitioners, who are generally software engineers or data scientists working in industry, are likely both

to be ML platform users themselves and to be connected to organisations who may be suitable research partners. To date, we have presented at two Melbourne-based ML practitioner meetups, and are an active participant of the Melbourne fair-ML reading group. We plan to circulate the internship proposal discussed above in these networks.

5.3.3 Research and publishing collaboration opportunities

In the long term, the success of our research will depend on our ability to build relationships with peers – other graduate students and early career academics – who will become our collaborators. As such, our networking plan is also focused on developing peer relationships through participation in relevant graduate student networks.

Within Australia, the Science and Society Network at Deakin University runs the AusSTS Graduate Network, which we are an active member of. Most recently, we presented at the annual AusSTS workshop, which proved to be a fruitful networking opportunity. We plan to continue to engage with the AusSTS Graduate Network. Several of the academic institutions listed in Section 5.2 also run PhD programs whose participants have research interests that overlap with ours. We are already informally connected to the 3A Institute's PhD program, and meet regularly with participating students. Through the AusSTS workshop we also met students who are members of the ADM+S Centre PhD program and the PhD program at the Emerging Technologies Lab at Monash University. Our plan is to continue participating in student workshops, and continue to be proactive in reaching out to peers who we meet through these events. Internationally, as mentioned already, ACM FAccT brings PhD students together though their Doctoral Colloquium. In addition, the 4S conference has an active student network and student presentation track. We plan to apply to participate in both of these in 2022.

Finally, to increase our visibility to potential collaborators, as well as to support our industry partner outreach, we plan to develop a personal website and engage more with members of the fair-ML community who are active on Twitter.

In this section we described our networking plan, focusing in particular on our outreach to the ADM+S Centre and to developing industry partnerships. In the following section, Section 5.4, we detail our publishing plan.

5.4 Publishing plan

Our publishing plan has been designed to align with the three phases of our research design, and is summarised in Table 5.2. As shown in the high-level timeline presented in Figure 5.1, we aim to prepare and submit a paper for publication at the conclusion of each phase of research. Reflecting the emergent nature of our research design, the aspects of our publishing plan that correspond to phase one of our proposed research design are currently more detailed than those that correspond to phase two and phase three.

Research phase	**Conference objective**	**Journal objective**
Phase one	At beginning of phase one: present at FAccT doctoral colloquium At conclusion of phase one: present at FAccT, AIES, or CHI workshop	Publish results of ML platform user analysis in digital humanities journal
Phase two	Present at FAccT, AIES, or CHI workshop	Publish results of ML platform user interviews and survey in an applied ethics journal
Phase three	Present at FAccT or AIES main conference	Publish results of prototype evaluation in applied ethics or digital humanities journal

Table 5.2: Table of conference presentation and publishing objectives for each phase of research.

As yet, to the best of our knowledge, there are no dedicated fair-ML journals. The *AI and Ethics* journal recently launched, and has published two issues so far, demonstrating a focus on the development of ML systems (and AI technologies more generally) that accord with societal expectations [MacIntyre et al., 2021]. However, it is unclear whether the journal has the support of the fair-ML community, as very few of the established fair-ML researchers are represented on the journal's editorial board. Meanwhile, many of the most highly cited and influential fair-ML papers have been published in the ACM FAccT, AIES, and CHI conference proceedings (examples, cited in this thesis proposal include [Buolamwini and Gebru, 2018; Green and Viljoen, 2020; Holstein et al., 2019; Kaur et al., 2020; Keyes, 2018]). As such, a large part of our publishing plan mirrors our networking plan; we focus, in 2021, on submitting proposals to the student or workshop tracks of these conferences, and then working towards submitting a conference paper in late 2022.

Fair-ML researchers with social science backgrounds have, however, published in a number of digital humanities and media studies journals. Relevant digital humanities and media studies journals, in descending order of journal impact factor, include: *New Media and Society*, *Big Data and Society*, and *Information Communication and Society*. These journals have published influential papers on the critical study of algorithms [Kitchin, 2017; Boyd and Crawford, 2012; Brock, 2018], the role of algorithms and computing systems in contemporary life [Bucher, 2017; Sadowski, 2019], the social and technical production of ML systems [Felzmann et al., 2019; Bechmann and Bowker, 2019; Orr and Davis, 2020], and the social dynamics and impacts of digitally mediated two-sided markets [Nieborg and Poell, 2018; Cotter, 2019]. We do not expect digital humanities journals to be relevant to all aspects of our research. However, we believe that our research design and efforts to explore who ML platform users are, how ML platforms are being used to develop ML systems, and how

ML platform users are enacting ethical considerations, may be of relevance to these journals. Additionally, *Information Communication and Society* has published a special issue on algorithms [Beer, 2017], and we believe it is likely some of these journals may announce special issues of relevance to our research during the course of our PhD studies.

There are also a number of applied ethics journals that have a particular focus on technology. In these journals fair-ML researchers have published analyses of ethical issues associated with ML systems, and evaluations of ethics interventions in ML system development. Relevant journals, in descending order of journal impact factor, are: *Ethics and Information Technology*, *Science and Engineering Ethics*, *AI and Society*, and *Philosophy and Technology*. *Philosophy and Technology* is of particular interest to us, as its Editor-in-Chief is currently Luciano Floridi, whose framework for conceptualising ethics interventions is central to our research design [2018], and discussed in Section 3.2. Additionally, *Philosophy and Technology* has recently published several studies of software development processes and ethics interventions from relatively junior academics outside of North America [Ahuja and Kumar, 2021; Loi and Christen, 2021; Gogoll et al., 2021]. We believe phase two and phase three of our proposed research design may be relevance to applied ethics journals, as they have published similar studies of the enactment of ethics interventions [Ryan et al., 2021], the development and evaluation of new ethics interventions [Mökander et al., 2021; Tomalin et al., 2021], and analyses of intervention opportunities within the software development process [Rochel and Evéquoz, 2020].

In this section we discussed our publishing plan, which closely aligns with our proposed research design. In the following section, Section 5.5, we discuss our training and professional development plans, which also are informed by our research design.

5.5 Training and professional development

The research design we described in Section 4.2 incorporates a broad range of research methods: CTDA in phase one, semi-structured interviews and surveys in phase two, and potentially participant observation in phase three. Our training and professional development plan is focused on improving our research skills in these methods. A summary of our training and professional development plan can be found below in Table 5.3. Our training and professional development plan currently extends only to the end of 2022. We expect to update and expand our plan, through conversations with our supervisors and in future annual reports.

At ANU, there are a number of courses offered by the School of Regulation and Governance, the ANU Centre for Social Research Methods, the School of Humanities, and the Research School of Humanities and the Arts which focus on methods for studying socio-technical systems. These include: REGN9076, Methods in Regulation and Governance; HUMN6003, Digital Humanities Methods and Practices; and

Year	**Training activities**	**Details**
2021	Research ethics and management	Undertake ANU training in privacy awareness, ethics approval, and research management
	Use of statistics	Meet with ANU Statistical Consulting Unit, undertake their online courses
	Emerging research methods	Apply to participate in CAIDE's 2021 summer school program
	CTDA research methods	Review Centre for Critical Race and Digital Studies syllabus
2022	Research methods for socio-technical systems	Participate in, or audit, relevant courses (e.g. HUMN6003, Digital Humanities Methods and Practices)
	Research management	Participate in ANU Researcher Development Team trainings, as relevant
	Emerging research methods	If possible, apply to participate in European or United States-based summer schools

Table 5.3: Table of planned training activities for 2021 and 2022.

SOCY2166, online research methods. We plan to review the syllabus for these courses and select appropriate courses to undertake as coursework in 2022. The ANU Statistical Consulting Unit also provides training services to higher degree candidates, including face-to-face consultation and four online courses. We plan to consult with the Statistical Consulting Unit in October 2021, and undertake the relevant online courses in the last quarter of 2021. Outside of ANU, the Centre for Artificial Intelligence and Digital Ethics (CAIDE) at the University of Melbourne offers one of the first graduate-level courses in fair-ML research, which it may be possible for us to audit in 2022. Finally, the Center for Critical Race and Digital Studies has developed a syllabus for researchers interested in the study of new media, which includes CTDA [Lopez and Land, 2019], which we intend to review in in the last quarter of 2021.

Outside of training on research methods, we also need to develop our research management and ethics skills. The ANU Research Services Division offers training in privacy awareness and the ethics approval process, which we intend to undertake in September 2021, to inform the preparation of our ethics approval documentation. The ANU Researcher Development Team also offers training in research management and various aspects of completing a PhD. In early 2021 we participated in the Research Development Team's Thesis Proposal bootcamp, and we intend to participate in future trainings, where they are relevant.

Finally, we note that summer schools are traditionally valuable training and network development opportunities for PhD students. Due to the ongoing COVID-19 crisis, we are currently unclear which summer school programs will be going ahead in the near future. That said, the CAIDE Summer Research Academy, which runs

from November to December, is proceeding for 2021 and we intend to apply to participate.

In this section we discussed our training and professional development plan, focusing particularly on activities we intend to undertake in 2021 and 2022. In the following section, Section 5.6, we consider the resourcing requirements for participation in these activities, and for our research project more broadly.

5.6 Resourcing requirements

The research activities we describe in our research design, in Section 4.2, are largely low-cost. They do not require special equipment, or significant compute time, and, due to COVID-19 restrictions, are not reliant on travel. That said, there are costs associated with the networking, publishing, and training activities we have outlined in the previous sections of this chapter. We detail these costs, and our revenue sources, in Table 5.4, below. We note that our revenue and expenditure values are only estimates at this stage, and that we will need to develop a more detailed budget in future work.

Category	**Value**	**Details**
Revenue	\$6000	Budget provided by ANU College of Engineering and Computer Science
	\$2000	Budget awarded through small grants, e.g. for travel to participate in conferences
	\$1500	Budget awarded through small scholarships
	\$9500	**Total projected revenue**
Expenditure	\$3000	Travel to attend conferences (in late 2022 onwards) or participate in fellowships
	\$1000	Fees for conferences, summer schools, etc.
	\$1500	Fees for publication submissions and preparation
	\$1500	Budget for research administration (e.g. interview transcription)
	\$1750	Recruitment costs
	\$750	Contingency
	\$9500	**Total projected expenditure**

Table 5.4: Table of projected costs and revenues for our research project.

5.7 Risks and mitigation strategies

In Table 5.5 we summarise the risks we have identified with this research project. These include risks that may prevent us from fulfilling our planned research, net-

working, or publishing activities, and risks arising from these activities. For each risk listed, we have assessed the likelihood and impact of it being realised, and begun to develop a mitigation strategy. Risks that have both a high likelihood of being realised, and would have a high impact should they do so, are those that we will need to be particularly proactive in managing. In future work, for each phase of our research, as we further refine our research design and prepare for ethics approval, we intend to produce an updated risk analysis, that focuses in particular on the risks arising from that phase of research activities.

Risk	**Likelihood**	**Impact**	**Mitigation strategy**
Unable to secure field sites for phase three implementation of prototype intervention	High	High	Begin outreach and networking well before phase three of research is planned to commence. Design prototype intervention in way that aligns with platform users we are most likely to secure access to.
Unable to recruit ML platform users to participate in phase two semi-structured interviews	Medium	High	Begin outreach and networking well before phase two of research is planned to commence. Design research question so that relatively small sample of users are necessary for interviews. Develop alternative research activities, should recruitment prove impossible.
Field site, or research participants, withdraw consent to publish findings from phase three prototype implementation and evaluation	Medium	High	Identify back up field sites for phase three prototype implementation. Select prototype use case cautiously, seeking to avoid likely contentious uses.
Research participants withdraw consent to publish findings from phase two survey	Low	High	Proactively seek consent for publication, and ensure that personally identifiable information is not collected during survey. Account for some attrition when setting recruitment targets.

Risk	Likelihood	Impact	Mitigation strategy
Research participants withdraw consent to publish findings from phase two semi-structured interviews	Low	High	Proactively seek consent for publication and explain procedures for ensuring as much anonymity as possible. Design interview questions with the disclosure constraints of ML platform users in mind. Account for some attrition when setting recruitment targets, and aim to conduct more semi-structured interviews than may be strictly necessary.
Confidentiality of a research participant is breached	Low	High	Maintain high standards of data security. Avoid collection of personally identifiable information unless absolutely necessary.
A research participant is reprimanded by an employer for their participation	Low	High	Engage employers proactively in the research, and seek their consent for participant engagement in phase two and phase three. In survey only collect data on type of employer, rather than place of employment.
Data collected during any phase of research is leaked or inappropriately accessed	Low	High	Maintain high standards of data security and management. Ensure data management plan includes deletion or archiving of data, as appropriate, at conclusion of project.
COVID-19 outbreaks or restrictions prevent in person gatherings at conferences or workshops	High	Medium	Assume participation in conferences and workshops will largely be virtual. Seek to establish personal relationships that are not dependent on in-person meetings. Focus in 2022 on building relationships locally.
COVID-19 outbreaks or restrictions prevent in person work during phase three prototype implementation and evaluation	Medium	Medium	Choose prototype intervention and field sites for phase three such that remote work will not be a barrier to proceeding. Design data collection methods that can be adapted to remote work.

Risk	Likelihood	Impact	Mitigation strategy
Unable to recruit participants for phase two survey	Medium	Low	Identify multiple paths to engaging ML platform users. Can pivot to alternative method, not reliant on user participation. Can constrain engagement to smaller category of ML platform users.
ML platform structures, services, or use change substantially during course of research	Medium	Low	Phased structure of research design and pragmatic approach enables us to pivot research activities should this become necessary.
Assumptions regarding significance of ML platform use by ML practitioners prove to be unfounded	Low	Low	Test assumption early through ML platform documentation analysis. If necessary, can pivot focus to ML system development more generally.

Table 5.5: Table of risks and mitigation strategies.

5.8 Conclusion

Thus concludes our Thesis Plan. We believe this plan, and our Thesis Proposal more broadly, expresses a valid set of research questions and a plausible plan for addressing them.

In Chapter 1 we introduced our research proposal. We situated ourselves within the fair-ML research community, provided an overview of Machine Learning terminology, and introduced our research questions.

In Chapter 2 we reviewed fair-ML literature, and sought to demonstrate the significance and novelty of our research questions. We described the key features of ML platforms, and contextualised these with comparisons to two-sided markets, infrastructure, and implementation frameworks. We highlighted three ethics issues associated with ML systems, and reviewed fair-ML proposals for addressing these during ML system development. We argued that the relationship between the enactment of ethics considerations during ML system development and the tools used to support ML system development has been under-studied by fair-ML researchers to date.

In Chapter 3 we discussed the conceptual frameworks that underpin our research design and will shape our research activities. We highlighted how we interpret the enactment of ethics considerations within the frame of discourse ethics, how we

plan to draw on existing typologies of proposed fair-ML interventions to support our research, and how we plan to apply affordance theory to interpret the relations between ML platforms, ML platform users, and the ML systems they develop.

In Chapter 4 we introduced the mixed methods methodology we adopted to develop our three-phase research design, and described the research activities we propose to undertake in each phase. Finally, in the preceding sections of Chapter 5 we detailed how we plan to undertake our proposed research activities and how we intend to share the findings from our research activities with the fair-ML community.

We look forward to feedback from our supervisory panel, and guests at our Thesis Proposal presentation. On this note, whilst this document is itself a proposal for future work, it also represents the culmination of work we have undertaken over the last year. This work would not have been possible without the generous support and patient feedback of our supervisory panel members. It therefore seems most appropriate to conclude this document by acknowledging and thanking them.

Professor Jochen Trumpf, Dr Elizabeth Williams, and Dr Niels Wouters: thank you.

Bibliography

2021a. ACM FAccT Conference. https://facctconference.org/index.html. (cited on page 30)

2021b. OpenCV: Object Detection. https://docs.opencv.org/4.5.3/d5/d54/group__objdetect.html. (cited on page 4)

Abebe, R.; Levy, K.; Barocas, S.; Raghavan, M.; Kleinberg, J.; and Robinson, D. G., 2020. Roles for Computing in Social Change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 252–260. doi:10.1145/3351095.3372871. (cited on pages 46 and 48)

ABS, 2016. Estimates of Aboriginal and Torres Strait Islander Australians. Technical report, Australian Bureau of Statistics. https://www.abs.gov.au/statistics/people/aboriginal-and-torres-strait-islander-peoples/estimates-aboriginal-and-torres-strait-islander-australians/jun-2016#key-statistics. (cited on page 33)

Abuhamad, G. E. A.; Brunet, M.-E. E. A.; Institute), V.; McCalman, L. G. I.; and Steinberg, D. G. I., 2021. Implications Tutorial: Using Harms and Benefits to Ground Practical AI Fairness Assessments in Finance. In *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://facctconference.org/2021/acceptedtuts.html#harms. (cited on page 94)

Ahuja, S. and Kumar, J., 2021. The Influence of Business Incentives and Attitudes on Ethics Discourse in the Information Technology Industry. *Philosophy & Technology*, (2021), 1–26. (cited on page 97)

Albert, K. and Delano, M., 2021. This Whole Thing Smacks of Gender. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 342–352. ACM, New York, NY, USA. doi:10.1145/3442188.3445898. https://dl.acm.org/doi/10.1145/3442188.3445898. (cited on pages 2, 34, 35, and 79)

Amazon, a. Amazon SageMaker Clarify – Bias Detection and Explainability. https://aws.amazon.com/sagemaker/clarify/. (cited on page 25)

Amazon, b. Artificial Intelligence Services. https://aws.amazon.com/machine-learning/ai-services/. (cited on pages 15 and 70)

Amazon, c. Learn Machine Learning on AWS – Amazon Web Services. https://aws.amazon.com/machine-learning/learn/. (cited on page 18)

Amazon, d. Machine Learning Competency Partners. https://aws.amazon.com/machine-learning/partner-solutions/?partner-solutions-cards.sort-by=item.additionalFields.partnerNameLower&partner-solutions-cards.sort-order=asc&awsf.partner-solutions-filter-partner-type-finserv=*all&awsf.partner-solutions-filter-. (cited on page 17)

Amazon, e. Machine Learning customers. https://aws.amazon.com/machine-learning/customers/. (cited on page 18)

Anable, A., 2018. Platform Studies. *Feminist Media Histories*, 4, 2 (2018), 135–140. doi:10.1525/fmh.2018.4.2.135. (cited on page 23)

Andreessen, M., 2007. Analyzing the Facebook Platform, three weeks in. https://web.archive.org/web/20071002070223/http://blog.pmarca.com/2007/06/analyzing_the_f.html#comment-73194948. (cited on page 28)

Andrejevic, M. and Selwyn, N., 2020. Facial recognition technology in schools: critical questions and concerns. doi:10.1080/17439884.2020.1686014. (cited on page 34)

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L., 2016. Machine Bias — ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. (cited on pages 2, 29, and 46)

ANU, 2021. Risk Level Classification Table. https://services.anu.edu.au/files/guidance/RiskLevelTableJan2021_2.pdf. (cited on page 80)

Ball, M.; Broadhurst, R.; Niven, A.; and Trivedi, H., 2019. Data Capture and Analysis of Darknet Markets. *SSRN Electronic Journal*, , March (2019), 1–14. doi:10.2139/ssrn.3344936. (cited on page 73)

Bandy, J., 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. (2021), 1–34. http://arxiv.org/abs/2102.04256. (cited on page 46)

Barabas, C.; Doyle, C.; Rubinovitz, J. B.; and Dinakar, K., 2020. Studying up: Reorienting the study of algorithmic fairness around issues of power. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, (2020), 167–176. doi:10.1145/3351095.3372859. (cited on page 79)

Barocas, S.; Guo, A.; Kamar, E.; Krones, J.; Morris, M. R.; Vaughan, J. W.; Wadsworth, D.; and Wallach, H., 2021a. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. http://arxiv.org/abs/2103.06076. (cited on page 45)

Barocas, S.; Hardt, M.; and Narayanan, A., 2021b. *Fairness and Machine Learning*. https://fairmlbook.org. (cited on pages 1, 2, 5, 29, and 32)

Bauer, J. M. and Herder, P. M., 2009. Designing socio-technical systems. In *Philosophy of Technology and Engineering Sciences* (Eds. D. M. Gabbay; P. Thagard; J. Woods; and A. W. M. Meijers). Elsevier Science & Technology. ISBN 9780080930749. (cited on pages 3 and 27)

Baxter, K.; Schlesinger, Y.; Aerni, S.; Baker, L.; Dawson, J.; Kenthapadi, K.; Kloumann, I.; and Wallach, H., 2020. Bridging the Gap from AI Ethics Research to Practice. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, (2020), 3375680. (cited on page 44)

Baylor, D.; Breck, E.; Cheng, H. T.; Fiedel, N.; Foo, C. Y.; Haque, Z.; Haykal, S.; Ispir, M.; Jain, V.; Koc, L.; Koo, C. Y.; Lew, L.; Mewald, C.; Modi, A. N.; Polyzotis, N.; Ramesh, S.; Roy, S.; Whang, S. E.; Wicke, M.; Wilkiewicz, J.; Zhang, X.; and Zinkevich, M., 2017. TFX: A TensorFlow-based production-scale machine learning platform. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. Part F1296, 1387–1395. doi:10.1145/3097983.3098021. (cited on page 28)

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 and Society*, 6, 1 (2019), 1–11. doi:10.1177/2053951718819569. (cited on page 96)

Beer, D., 2017. The social power of algorithms. doi:10.1080/1369118X.2016.1216147. (cited on page 97)

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Mitchell, M., 2020. *On the dangers of stochastic parrots: can language models be too big?*, vol. 1. Association for Computing Machinery. ISBN 9781450375856. doi:10.1145/3442188.3445922. (cited on page 30)

Benjamin, R., 2019. *Race after technology: Abolitionist tools for the new jim code*. Wiley. ISBN 978-1-509-52643-7. (cited on page 30)

Bernardi, L.; Mavridis, T.; and Estevez, P., 2019. 150 successful machine learning models: 6 lessons learned at Booking.com. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2019), 1743–1751. doi:10.1145/3292500.3330744. (cited on page 2)

Bietti, E., 2020. From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, (2020), 210–219. doi:10.1145/3351095.3372860. (cited on page 41)

Binns, R., 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149–159. (cited on page 2)

Bird, S.; Dudik, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; and Walker, K., 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. (2020), 1–6. (cited on pages 25, 26, 35, and 49)

Bogost, I. and Montfort, N., 2007. New media as material constraint: An introduction to platform studies. In *Electronic Techtonics: Thinking at the Interface. Proceedings of the First International HASTAC Conference*, 176–193. ISBN 9781435713628. (cited on page 28)

Boxall, K.; Nyanjom, J.; and Slaven, J., 2018. Disability, hospitality and the new sharing economy. *International Journal of Contemporary Hospitality Management*, 30, 1 (2018), 539–556. doi:10.1108/IJCHM-09-2016-0491. (cited on page 22)

Boyd, D. and Crawford, K., 2012. Critical questions for big data - Provocations for a cultural, technological, and scholarly phenomenon. *Informacios Tarsadalom*, , 2 (2012), 7–23. (cited on page 96)

Brock, A., 2011. "When keeping it real goes wrong": Resident evil 5, racial representation, and gamers. *Games and Culture*, 6, 5 (2011), 429–452. doi:10.1177/1555412011402676. (cited on pages 71 and 72)

Brock, A., 2012. From the Blackhand Side: Twitter as a Cultural Conversation. *Journal of Broadcasting and Electronic Media*, 56, 4 (2012), 529–549. doi:10.1080/08838151.2012.732147. (cited on page 71)

Brock, A., 2018. Critical technocultural discourse analysis. *New Media and Society*, 20, 3 (2018), 1012–1030. doi:10.1177/1461444816677532. (cited on pages 70, 71, 72, and 96)

Brown, S.; Davidovic, J.; and Hasan, A., 2021. The algorithm audit: Scoring the algorithms that score us. doi:10.1177/2053951720983865. (cited on pages 45, 46, and 47)

Bucher, T., 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information Communication and Society*, 20, 1 (2017), 30–44. doi:10.1080/1369118X.2016.1154086. (cited on page 96)

Buolamwini, J. and Gebru, T., 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR. (cited on pages 34 and 96)

Burrell, J., 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society*, 3, 1 (2016), 1–12. doi:10.1177/2053951715622512. (cited on pages 25, 30, and 31)

Bynum, T. W., 2001. Computer ethics: Its birth and its future. doi:10.1023/A:1011893925319. (cited on page 36)

Caplan, R. and Boyd, D., 2018. Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data & Society*, 5, 1 (2018), 205395171875725. doi:10.1177/2053951718757253. http://journals.sagepub.com/doi/10.1177/2053951718757253. (cited on pages 16 and 79)

Chouldechova, A. and Roth, A., 2018. The Frontiers of Fairness in Machine Learning. http://arxiv.org/abs/1810.08810. (cited on page 2)

Christin, A., 2020. The ethnographer and the algorithm: beyond the black box. *Theory and Society*, 49, 5-6 (10 2020), 897–918. doi:10.1007/s11186-020-09411-3. (cited on page 30)

Cobbe, J.; Lee, M. S. A.; and Singh, J., 2021. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. (2021). http://arxiv.org/abs/2102.04201. (cited on page 45)

Cofone, I., 2019. Algorithmic Discrimination Is an Information Problem. *Hastings Law Journal*, 70, 6 (2019), 1389. (cited on page 34)

Corbett-Davies, S. and Goel, S., 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. (7 2018), 1–5. http://arxiv.org/abs/1808.00023. (cited on page 43)

Cotter, K., 2019. Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media & Society*, 21, 4 (4 2019), 895–913. doi:10.1177/1461444818815684. http://journals.sagepub.com/doi/10.1177/1461444818815684. (cited on page 96)

Creswell, J.; Shope, R.; Clark, V. L. P.; and Green, D. O., 2006. How interpretive qualitative research extends mixed methods research. *Research in the Schools*, 13, 1 (2006), 1–11. (cited on page 65)

Creswell, J. W., 1999. Mixed-Method Research. In *Handbook of Educational Policy*, 455–472. Elsevier. doi:10.1016/B978-012174698-8/50045-X. https://linkinghub.elsevier.com/retrieve/pii/B978012174698850045X. (cited on page 62)

Davis, J. L., 2020. *How Artifacts Afford: The Power and Politics of Everyday Things*. Design thinking, design theory. MIT Press. ISBN 9780262044110. (cited on pages 1, 58, 59, 60, 70, 71, 73, and 93)

Davis, J. L. and Chouinard, J. B., 2016. Theorizing Affordances: From Request to Refuse. *Bulletin of Science, Technology & Society*, 36, 4 (2016), 241–248. doi:10.1177/0270467617714944. (cited on pages 58 and 59)

DeCuir-Gunby, J. T. and Schutz, P. A., 2017. Developing a Mixed Methods Proposal: A Practical Guide for Beginning Researchers. doi:10.4135/9781483399980. https://methods.sagepub.com/book/developing-a-mixed-methods-proposal. (cited on pages 63 and 64)

Demange, G. and Gale, D., 1985. The Strategy Structure of Two-Sided Matching Markets. *Econometrica*, 53, 4 (7 1985), 873. doi:10.2307/1912658. https://www.jstor.org/stable/1912658?origin=crossref. (cited on page 20)

DeVries, T.; Misra, I.; Wang, C.; and van der Maaten, L., 2019. Does Object Recognition Work for Everyone? (2019). http://arxiv.org/abs/1906.02659. (cited on pages 32 and 33)

Dieterich, W.; Mendoza, C.; and Brennan, T., 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. *Performance of the COMPAS Risk Scales in Broward County*, (2016), 1–37. (cited on page 29)

Dille, S., 2019a. How Microsoft Azure Machine Learning Studio Clarifies Data Science | by Steve Dille | Towards Data Science. https://towardsdatascience.com/how-microsoft-azure-machine-learning-studio-clarifies-data-science-8e8d3e6ed64e. (cited on page 73)

Dille, S., 2019b. How to Decide Between Amazon SageMaker and Microsoft Azure Machine Learning Studio | by Steve Dille | Towards Data Science. https://tinyurl.com/a2r7ybz2. (cited on page 73)

Dourish, P., 2016. Algorithms and their others: Algorithmic culture in context. *Big Data and Society*, 3, 2 (2016), 1–11. doi:10.1177/2053951716665128. (cited on page 31)

Duda, R. and Shortliffe, E., 1983. Expert Systems Research. *Science*, 220, 4594 (4 1983), 261–268. doi:10.1126/science.6340198. https://www.sciencemag.org/lookup/doi/10.1126/science.6340198. (cited on page 37)

Dworkin, S. L., 2012. Sample size policy for qualitative studies using in-depth interviews. *Archives of Sexual Behavior*, 41, 6 (2012), 1319–1320. doi:10.1007/s10508-012-0016-6. (cited on page 75)

Edmonds, W. A. and Kennedy, T. D., 2020. *An Applied Guide to Research Designs: Quantitative, Qualitative, and Mixed Methods*. ISBN 9781483317274. doi:10.4135/9781071802779. (cited on pages 62, 65, 75, and 78)

Eubanks, V., 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press. ISBN 1-4668-8596-3. (cited on pages 17 and 30)

Evans, D. S.; Schmalensee, R.; Noel, M. D.; Chang, H. H.; and Garcia-Swartz, D. D., 2011. Platform economics: Essays on multi-sided businesses. *Competition Policy International*, (2011), 459. (cited on page 20)

Felzmann, H.; Villaronga, E. F.; Lutz, C.; and Tamò-Larrieux, A., 2019. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data and Society*, 6, 1 (2019), 1–14. doi:10.1177/2053951719860542. (cited on pages 44 and 96)

Floridi, L., 2018. Soft Ethics and the Governance of the Digital. doi:10.1007/s13347-018-0303-9. (cited on pages 44, 57, and 97)

Flyvbjerg, B., 2001. *Making social science matter: Why social inquiry fails and how it can succeed again*. Cambridge university press. ISBN 1139429922. (cited on page 37)

Forsythe, D. E., 1993. Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence. *Social studies of science*, 23, 3 (1993), 445–477. (cited on page 37)

Fultner, B., 2014. Jurgen Habermas: Key concepts. *Jurgen Habermas: Key Concepts*, (2014), 1–264. doi:10.4324/9781315711461. (cited on pages 55 and 56)

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; Crawford, K.; Jamie Morgenstern, G.; Vecchione, B.; and Wortman Vaughan, J., 2018. Datasheets for datasets. *arXiv*, (2018). (cited on pages 47, 48, 51, 73, and 91)

Gogoll, J.; Zuber, N.; Kacianka, S.; Greger, T.; Pretschner, A.; and Nida-Rümelin, J., 2021. Ethics in the Software Development Process: from Codes of Conduct to Ethical Deliberation. *Philosophy & Technology*, (2021), 1–24. (cited on page 97)

Goldenfein, J., 2019. Algorithmic Transparency and Decision-Making Accountability: Thoughts for Buying Machine Learning Algorithms. *Closer to the Machine: Technical, Social, and Legal aspects of AI*, (2019), 41–145. https://ssrn.com/abstract=3445873. (cited on pages 17 and 93)

Google, a. AI Platform | Google Cloud Platform. https://cloud.google.com/ai-platform. (cited on pages 15 and 17)

Google, b. Google Cloud Partner Directory | Google Cloud. https://cloud.withgoogle.com/partners/. (cited on page 17)

Gorwa, R., 2019. What is platform governance? *Information Communication and Society*, 22, 6 (2019), 854–871. doi:10.1080/1369118X.2019.1573914. (cited on page 22)

Green, B., 2021. The Contestation of Tech Ethics: A Sociotechnical Approach to Ethics and Technology in Action. (6 2021). http://arxiv.org/abs/2106.01784. (cited on page 45)

Green, B. and Hu, L., 2018. The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. *Presented at the Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning*, (2018). (cited on page 34)

Green, B. and Viljoen, S., 2020. Algorithmic realism: Expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 19–31. doi:10.1145/3351095.3372840. (cited on pages 2, 38, 63, and 96)

Greene, D.; Hoffmann, A. L.; and Stark, L., 2019. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, (2019), 2122–2131. doi:10.24251/hicss.2019.258. (cited on page 44)

Gyódi, K., 2019. Airbnb in European cities: Business as usual or true sharing economy? *Journal of Cleaner Production*, 221 (2019), 536–551. doi:10.1016/j.jclepro.2019.02.221. (cited on page 73)

Hanna, A.; Denton, E.; Smart, A.; and Smith-Loud, J., 2020. Towards a critical race methodology in algorithmic fairness. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, (2020), 501–512. doi:10.1145/3351095.3372826. (cited on page 2)

Harrison, G.; Hanson, J.; Jacinto, C.; Ramirez, J.; and Ur, B., 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. doi:10.1145/3351095.3372831. (cited on page 63)

Hazelwood, K.; Bird, S.; Brooks, D.; Chintala, S.; Diril, U.; Dzhulgakov, D.; Fawzy, M.; Jia, B.; Jia, Y.; Kalro, A.; Law, J.; Lee, K.; Lu, J.; Noordhuis, P.; Smelyanskiy, M.; Xiong, L.; and Wang, X., 2018. Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective. *Proceedings - International Symposium on High-Performance Computer Architecture*, 2018-Febru (2018), 620–629. doi:10.1109/HPCA.2018.00059. (cited on pages 17 and 22)

Helmond, A., 2015. The Platformization of the Web: Making Web Data Platform Ready. *Social Media and Society*, 1, 2 (2015). doi:10.1177/2056305115603080. (cited on page 27)

Helmond, A.; Nieborg, D. B.; and van der Vlist, F. N., 2019. Facebook's evolution: development of a platform-as-infrastructure. doi:10.1080/24701475.2019.1593667. (cited on pages 27 and 79)

Henderson, P.; Hu, J.; Romoff, J.; Brunskill, E.; Jurafsky, D.; and Pineau, J., 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21 (2020), 1–43. (cited on page 30)

Hesse-Biber, S. and Johnson, R. B. (Eds.), 2015. *The Oxford Handbook of Multimethod and Mixed Methods Research Inquiry*. ISBN 9780199933624. (cited on page 67)

High-Level Independent Group on Artificial Intelligence (AI HLEG), 2019. A Definition of AI: Main Capabilities and Disciplines. *European Commission*, (2019), 7. https://ec.europa.eu/digital-single-. (cited on pages 1 and 44)

Hinze, J. and Parker, H. W., 1978. Safety: Productivity and Job Pressures. *ASCE J Constr Div*, 104, 1 (1978), 27–34. doi:10.1061/jcceaz.0000754. (cited on page 48)

Hoad, T. F., 2003. enact. doi:10.1093/acref/9780192830982.013.4998. https://www.oxfordreference.com/view/10.1093/acref/9780192830982.001.0001/acref-9780192830982-e-4998. (cited on page 40)

Hohman, F.; Head, A.; Caruana, R.; DeLine, R.; and Drucker, S. M., 2019. Gamut: A design probe to understand how data scientists understand machine learning models. *Conference on Human Factors in Computing Systems - Proceedings*, (2019), 1–13. doi:10.1145/3290605.3300809. (cited on pages 75, 76, and 77)

Holstein, K.; Vaughan, J. W.; Daumé, H.; Dudík, M.; and Wallach, H., 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16. doi:10.1145/3290605.3300830. (cited on pages 41, 42, 49, 75, 91, and 96)

HuggingFace. Hugging Face - The AI community building the future. https://huggingface.co/. (cited on pages 20, 22, 26, 47, 51, 70, and 73)

Hughes, T. P., 1987. The evolution of large technological systems. *The social construction of technological systems: New directions in the sociology and history of technology*, 82 (1987). (cited on page 27)

Hutchinson, B. and Mitchell, M., 2018. 50 years of test (Un)fairness: lessons for machine learning. *arXiv*, (2018), 49–58. (cited on page 2)

Hutchinson, B.; Smart, A.; Hanna, A.; Denton, E.; Greer, C.; Kjartansson, O.; Barnes, P.; and Mitchell, M., 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (2021), 560–575. doi:10.1145/3442188.3445918. (cited on page 77)

Hutchinson, H.; Mackay, W.; Westerlund, B.; Bederson, B. B.; Druin, A.; Plaisant, C.; Beaudouin-Lafon, M.; Conversy, S.; Evans, H.; Hansen, H.; Roussel, N.; Eiderbäck, B.; Lindquist, S.; and Sundblad, Y., 2003. Technology probes: Inspiring design for and with families. *Conference on Human Factors in Computing Systems - Proceedings*, , 5 (2003), 17–24. (cited on pages 76 and 77)

IBM. SPSS Software - Australia | IBM. https://www.ibm.com/au-en/analytics/spss-statistics-software. (cited on page 15)

IEEE, 2019. Ethically Aligned Design: First Edition Overview. Technical report, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e-overview.pdf. (cited on pages 44 and 47)

Ingram, D., 2010. *Habermas: introduction and analysis*. ISBN 9780801448799. (cited on page 56)

Isdahl, R. and Gundersen, O. E., 2019. Out-of-the-box reproducibility: A survey of machine learning platforms. In *Proceedings - IEEE 15th International Conference on eScience, eScience 2019*, i, 86–95. IEEE. doi:10.1109/eScience.2019.00017. (cited on page 50)

Islind, A. S.; Norström, L.; Vallo Hult, H.; and Olsson, S. R., 2021. Socio-Technical Interplay in a Two-Sided Market: The Case of Learning Platforms BT - Digital Transformation and Human Behavior. 33–53. Springer International Publishing, Cham. (cited on page 20)

Jackson, K. M.; Pukys, S.; Castro, A.; Hermosura, L.; Mendez, J.; Vohra-Gupta, S.; Padilla, Y.; and Morales, G., 2018. Using the transformative paradigm to conduct a mixed methods needs assessment of a marginalized community: Methodological lessons and implications. *Evaluation and Program Planning*, 66, October 2017 (2018), 111–119. doi:10.1016/j.evalprogplan.2017.09.010. https://doi.org/10.1016/j.evalprogplan.2017.09.010. (cited on page 62)

Jacobs, A. Z. and Wallach, H., 2021. *Measurement and fairness*, vol. 1. Association for Computing Machinery. ISBN 9781450383097. doi:10.1145/3442188.3445901. (cited on pages 34, 35, 43, and 63)

Jakobson, R., 1960. Closing statement: linguistics and poetics. In *The Lyric Theory Reader: A Critical Anthology (2014)*, 234. (cited on page 25)

Jasanoff, S., 2004. *States of knowledge: The co-production of science and the social order*. ISBN 0203413849. doi:10.4324/9780203413845. (cited on page 40)

Jo, E. S. and Gebru, T., 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, (2020), 306–316. doi:10.1145/3351095.3372829. (cited on page 79)

Jobin, A.; Ienca, M.; and Vayena, E., 2019. Artificial Intellgience: the global landscape of ethics guidelines. *Nature Machine Intelligence*, 1, 9 (2019), 389–399. https://arxiv.org/pdf/1906.11668. (cited on page 44)

Johnson, R. B. and Onwuegbuzie, A. J., 2004. Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher*, 33, 7 (2004), 14–26. doi:10.3102/0013189X033007014. (cited on page 62)

Joler, V. and Crawford, K., 2018. Anatomy of an AI system: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources. Technical report. https://anatomyof.ai. (cited on pages 18 and 30)

Kasy, M. and Abebe, R., 2020. Fairness, equality, and power in algorithmic decision-making. (2020), 1–14. http://www.cs.cornell.edu/~red/fairness_equality_power.pdf. (cited on pages 2, 43, and 44)

Katell, M.; Young, M.; Dailey, D.; Herman, B.; Guetler, V.; Tam, A.; Binz, C.; Raz, D.; and Krafft, P. M., 2020. Toward situated interventions for algorithmic equity: Lessons from the field. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, (2020), 45–55. doi:10.1145/3351095.3372874. (cited on page 2)

Kaur, H.; Nori, H.; Jenkins, S.; Caruana, R.; Wallach, H.; and Wortman Vaughan, J., 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. *Conference on Human Factors in Computing Systems - Proceedings*, (2020), 1–14. doi:10.1145/3313831.3376219. (cited on pages 32, 35, 41, 49, 75, 76, 91, and 96)

Kelkar, S., 2018. Engineering a platform: The construction of interfaces, users, organizational roles, and the division of labor. doi:10.1177/1461444817728682. (cited on page 79)

Kenney, M.; Bearson, D.; and Zysman, J., 2019. The Platform Economy Matures: Pervasive Power, Private Regulation, and Dependent Entrepreneurs. *SSRN Electronic Journal*, (2019). doi:10.2139/ssrn.3497974. (cited on page 22)

Keyes, O., 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2, CSCW (2018), 1–22. doi:10.1145/3274357. (cited on pages 35, 79, and 96)

Kitchin, R., 2017. Thinking critically about and researching algorithms. *Information Communication and Society*, 20, 1 (2017), 14–29. doi:10.1080/1369118X.2016.1154087. https://doi.org/10.1080/1369118X.2016.1154087files/2753/Kitchin-2017-Thinkingcriticallyaboutandresearchingalgorith.pdffiles/2752/1369118X.2016.html. (cited on page 96)

Krafft, P. M.; Young, M.; Katell, M.; Lee, J. E.; Narayan, S.; Epstein, M.; Dailey, D.; Herman, B.; Tam, A.; Guetler, V.; Bintz, C.; Raz, D.; Ousman Jobe, P.; Putz, F.; Robick Bissan Barghouti, B.; Robick, B.; and Barghouti, B., 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. (2021). https://doi.org/10.1145/3442188.3445938. (cited on pages 45 and 46)

Kruchten, P.; Fraser, S.; and Coallier, F. (Eds.), 2019. *Agile Processes in Software Engineering and Extreme Programming*, vol. 355 of *Lecture Notes in Business Information Processing*. Springer International Publishing, Cham. ISBN 978-3-030-19033-0. doi:10.1007/978-3-030-19034-7. http://link.springer.com/10.1007/978-3-030-19034-7. (cited on page 62)

Kshetri, N., 2021. Data Labeling for the Artificial Intelligence Industry: Economic Impacts in Developing Countries. *IT Professional*, 23, 2 (2021), 96–99. doi:10.1109/MITP.2020.2967905. (cited on page 30)

Kyriakou, K.; Barlas, P.; Kleanthous, S.; and Otterbacher, J., 2019. Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*, (2019), 313–322. (cited on page 50)

Larkin, B., 2013. The politics and poetics of infrastructure. *Annual Review of Anthropology*, 42 (2013), 327–343. doi:10.1146/annurev-anthro-092412-155522. (cited on pages 23, 24, 25, 26, and 27)

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J., 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. (cited on page 29)

Li, E. L.; Chen, E.; Hermann, J.; Zhang, P.; and Wang, L., 2017. Scaling Machine Learning as a Service. In *International Conference on Predictive Applications and APIs*, 14–29. (cited on pages 16, 17, 18, 22, and 28)

Light, B.; Burgess, J.; and Duguay, S., 2018. The walkthrough method: An approach to the study of apps. *New Media and Society*, 20, 3 (2018), 881–900. doi:10.1177/1461444816675438. (cited on page 73)

Lima, L.; Furtado, V.; Furtado, E.; and Almeida, V., 2019. Empirical Analysis of Bias in Voice-Based Personal Assistants. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, 533–538. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3308560.3317597. https://doi.org/10.1145/3308560.3317597. (cited on page 34)

Loi, M. and Christen, M., 2021. Choosing how to discriminate: navigating ethical trade-offs in fair algorithmic design for the insurance sector. *Philosophy & Technology*, (2021), 1–26. (cited on page 97)

Lopez, L. K. and Land, J. (Eds.), 2019. *Critical Race and Digital Studies Syllabus, Volume 1*. Center for Critical Race and Digital Studies. criticalracedigitalstudies.com. (cited on page 98)

Lwakatare, L. E.; Raj, A.; Bosch, J.; Olsson, H. H.; and Crnkovic, I., 2019. A taxonomy of software engineering challenges for machine learning systems: An empirical investigation. In *International Conference on Agile Software Development*, 227–243. Springer, Cham. (cited on pages 6 and 16)

Lwakatare, L. E.; Raj, A.; Crnkovic, I.; Bosch, J.; and Olsson, H. H., 2020. Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions. *Information and Software Technology*, 127 (2020). doi:10.1016/j.infsof.2020.106368. (cited on page 16)

MacIntyre, J.; Medsker, L.; and Moriarty, R., 2021. Past the tipping point? *AI and Ethics*, 1, 1 (2 2021), 1–3. doi:10.1007/s43681-020-00016-1. https://doi.org/10.1007/s43681-020-00016-1http://link.springer.com/10.1007/s43681-020-00016-1. (cited on page 96)

Mackenzie, A., 2015. The production of prediction: What does machine learning want? *European Journal of Cultural Studies*, 18, 4-5 (2015), 429–445. doi:10.1177/1367549415577384. (cited on page 79)

Mackenzie, A., 2019. From API to AI: platforms and their opacities. *Information Communication and Society*, 22, 13 (2019), 1989–2006. doi:10.1080/1369118X.2018.1476569. https://doi.org/10.1080/1369118X.2018.1476569. (cited on page 28)

Madaio, M. A.; Stark, L.; Wortman Vaughan, J.; and Wallach, H., 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Conference on Human Factors in Computing Systems - Proceedings*, (2020). doi:10.1145/3313831.3376445. (cited on pages 75 and 77)

Marda, V. and Narayan, S., 2020. Data in New Delhi's predictive policing system. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, vol. 4, 317–324. doi:10.1145/3351095.3372865. (cited on page 2)

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A., 2019. A survey on bias and fairness in machine learning. (cited on pages 29 and 33)

Mertens, D. M., 2003. Mixed methods and the politics of human research: The transformative-emancipatory perspective. *Handbook of mixed methods in social and behavioral research*, (2003), 135–164. (cited on page 61)

Mertens, D. M., 2007. Transformative Paradigm. *Journal of Mixed Methods Research*, 1, 3 (7 2007), 212–225. doi:10.1177/1558689807302811. http://journals.sagepub.com/doi/10.1177/1558689807302811. (cited on page 62)

Metselaar, S. and Widdershoven, G., 2016. Discourse Ethics. In *Encyclopedia of Global Bioethics*, 895–902. Springer International Publishing, Cham. doi:10.1007/978-3-319-09483-0{_}145. http://link.springer.com/10.1007/978-3-319-09483-0_145. (cited on page 56)

Microsoft, a. AI Platform | Microsoft Azure. https://azure.microsoft.com/en-au/overview/ai-platform/. (cited on pages 15, 17, and 70)

Microsoft, b. Application for Gated Services. https://tinyurl.com/44sv5whu. (cited on pages 18, 22, and 49)

Microsoft, c. Azure on Microsoft Learn | Microsoft Docs. https://docs.microsoft.com/en-us/learn/azure/. (cited on pages 18 and 51)

Midgley, G.; Munlo, I.; and Brown, M., 1998. The theory and practice of boundary critique: developing housing services for older people. *Journal of the Operational Research Society*, 49, 5 (1998), 467–478. doi:10.1057/palgrave.jors.2600531. https://www.tandfonline.com/doi/full/10.1057/palgrave.jors.2600531. (cited on page 38)

Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. doi:10.1016/j.artint.2018.07.007. (cited on page 31)

Mingers, J., 2011. Ethics and OR: Operationalising discourse ethics. *European Journal of Operational Research*, 210, 1 (2011), 114–124. doi:10.1016/j.ejor.2010.11.003. http://dx.doi.org/10.1016/j.ejor.2010.11.003. (cited on page 56)

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T., 2019. Model cards for model reporting. In *Proceedings of the 2019 Conference on Fairness, Accountability and Transparency*, Figure 2, 220–229. doi:10.1145/3287560.3287596. (cited on pages 47, 51, 73, and 91)

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; and Lum, K., 2021. Algorithmic fairness: Choices, assumptions, and definitions. doi:10.1146/annurev-statistics-042720-125902. (cited on pages 32, 33, and 43)

Mitchell, T. M., 1997. *Machine Learning*. McGraw-Hill. ISBN 0070428077. (cited on page 1)

Mökander, J.; Morley, J.; Taddeo, M.; and Floridi, L., 2021. Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics*, 27, 4 (2021), 1–30. (cited on page 97)

Montfort, N. and Bogost, I., 2009. *Racing the beam: The Atari video computer system*. Mit Press. ISBN 0262261529. (cited on page 28)

Moor, J. H., 1979. Are there decisions computers should never make? *Nature and System*, 1, 4 (1979), 217–229. doi:10.4324/9781315259697-39. (cited on page 48)

Morley, J.; Elhalal, A.; Garcia, F.; Kinsey, L.; Mokander, J.; and Floridi, L., 2021. Ethics as a service: a pragmatic operationalisation of AI Ethics. *Ethics*, 24, 3 (2 2021), 265. doi:10.1086/206828. http://arxiv.org/abs/2102.09364. (cited on pages 44, 45, 46, and 47)

Morley, J.; Floridi, L.; Kinsey, L.; and Elhala, A., 2019a. From What to How: An Overview of AI Ethics Tools, Methods and Researchto Translate Principles into Practices. (cited on pages 43 and 91)

Morley, J.; Floridi, L.; Kinsey, L.; and Elhalal, A., 2019b. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *arXiv*, (5 2019). http://arxiv.org/abs/1905.06876. (cited on pages 47, 57, and 91)

Nieborg, D. B. and Poell, T., 2018. The platformization of cultural production: Theorizing the contingent cultural commodity. *New Media & Society*, 20, 11 (11 2018), 4275–4292. doi:10.1177/1461444818769694. http://journals.sagepub.com/doi/10.1177/1461444818769694. (cited on page 96)

Noble, S. U., 2018. *Algorithms of oppression: How search engines reinforce racism*. nyu Press. ISBN 1479837245. (cited on page 30)

Obermeyer, Z. Z.; Powers, B.; Vogeli, C.; and Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366, 6464 (2019), 447–463. doi:10.1530/ey.17.12.7. http://science.sciencemag.org/. (cited on pages 2, 29, 34, and 46)

O'Neil, C., 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA. ISBN 0553418815. (cited on page 30)

Orr, W. and Davis, J. L., 2020. Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information Communication and Society*, 23, 5 (2020), 719–735. doi:10.1080/1369118X.2020.1713842. (cited on pages 41, 42, 75, and 96)

Oswald, M.; Grace, J.; Urwin, S.; and Barnes, G. C., 2018. Algorithmic risk assessment policing models: Lessons from the Durham HART model and 'experimental' proportionality. *Information and Communications Technology Law*, 27, 2 (2018), 223–250. doi:10.1080/13600834.2018.1458455. 10.1080/13600834.2018.1458455. (cited on pages 2, 31, and 32)

Parekh, A. B. and Natarajan, S., 2021. Just and Equitable Data Labeling: Toward a Responsible AI Supply Chain. *Interactions*, 28, 4 (6 2021), 80. doi:10.1145/3470544. https://doi.org/10.1145/3470544. (cited on page 30)

Pasquale, F., 2015. *The black box society*. Harvard University Press. ISBN 0674368274. (cited on page 31)

Patro, G. K.; Chakraborty, A.; Ganguly, N.; and Gummadi, K. P., 2020. Incremental fairness in two-sided market platforms: On smoothly updating recommendations. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, , iii (2020), 181–188. doi:10.1609/aaai.v34i01.5349. (cited on page 20)

Perkowitz, S., 2021. The Bias in the Machine: Facial Recognition Technology and Racial Disparities. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, (2021), 1–16. doi:10.21428/2c646de5.62272586. (cited on page 34)

Perrow, C., 2011. *Normal Accidents: Living with High Risk Technologies-Updated edition*. Princeton university press. ISBN 140082849X. (cited on page 46)

Pichai, S., 2018. AI at Google: our principles. *The Keyword*, 7 (2018). (cited on page 44)

Plantin, J. C. and de Seta, G., 2019. WeChat as infrastructure: the techno-nationalist shaping of Chinese digital platforms. *Chinese Journal of Communication*, 12, 3 (2019), 257–273. doi:10.1080/17544750.2019.1572633. https://doi.org/10.1080/17544750.2019.1572633. (cited on page 27)

Plantin, J. C.; Lagoze, C.; Edwards, P. N.; and Sandvig, C., 2018. Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media and Society*, 20, 1 (2018), 293–310. doi:10.1177/1461444816661553. (cited on pages 19, 23, and 27)

Plantin, J. C. and Punathambekar, A., 2019. Digital media infrastructures: pipes, platforms, and politics. *Media, Culture and Society*, 41, 2 (2019), 163–174. doi:10.1177/0163443718818376. (cited on pages 9 and 27)

Polyzotis, N.; Roy, S.; Whang, S. E.; and Zinkevich, M., 2017. Data management challenges in production machine learning. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Part F1277 (2017), 1723–1726. doi:10.1145/3035918.3054782. (cited on page 16)

Rahman, K. S., 2018. Regulating Informational Infrastructure: Internet Platforms as the New Public Utilities. *GEO. L. TECH. REV*, 2.2 (2018), 234–251. (cited on page 27)

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P., 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. doi:10.1145/3351095.3372873. (cited on pages 45, 46, and 91)

Rességuier, A. and Rodrigues, R., 2020. AI ethics should not remain toothless! A call to bring back the teeth of ethics. doi:10.1177/2053951720942541. (cited on page 45)

Richardson, B.; Garcia-Gathright Spotify, J.; Way, S. F.; Jennifer Thom, S.; Henriette Cramer, S.; Garcia-Gathright, J.; Thom, J.; and Cramer, H., 2021. Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits; Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. https://doi.org/10.1145/3411764. (cited on pages 47, 63, and 75)

Robert, L. P.; Pierce, C.; Marquis, L.; Kim, S.; and Alahmad, R., 2020. Designing fair AI for managing employees in organizations: a review, critique, and design agenda. *Human-Computer Interaction*, 35, 5-6 (2020), 545–575. doi:10.1080/07370024.2020.1735391. https://doi.org/10.1080/07370024.2020.1735391. (cited on page 58)

Rochel, J. and Evéquoz, F., 2020. Getting into the engine room: a blueprint to investigate the shadowy steps of AI ethics. *AI & SOCIETY*, (2020), 1–14. (cited on page 97)

Rochet, J.-c. and Tirole, J., 2003. Platform Competition in Two-Sided Markets. *Journal of the European Economic Association*, 1, 4 (6 2003), 990–1029. doi:10.1162/154247603322493212. https://academic.oup.com/jeea/article/2280902/Platform. (cited on page 20)

Rochet, J.-c. and Tirole, J., 2004. Two-Sided Markets: An Overview. (cited on pages 20 and 22)

Roson, R., 2009. Two-Sided Markets: A Tentative Survey. *Review of Network Economics*, 4, 2 (2009), 142–160. doi:10.2202/1446-9022.1070. (cited on pages 20 and 21)

Roy, A.; Qureshi, S.; Pande, K.; Nair, D.; Gairola, K.; Jain, P.; Singh, S.; Sharma, K.; Jagadale, A.; Lin, Y. Y.; Sharma, S.; Gotety, R.; Zhang, Y.; Tang, J.; Mehta, T.; Sindhanuru, H.; Okafor, N.; Das, S.; Gopal, C. N.; Rudraraju, S. B.; and Kakarlapudi, A. V., 2019. Performance comparison of machine learning platforms. *INFORMS Journal on Computing*, 31, 2 (2019), 207–225. doi:10.1287/ijoc.2018.0825. (cited on pages 2 and 50)

Russell, S. J. and Norvig, P., 2021. *Artificial Intelligence: A Modern Approach*. Pearson, 4th editio edn. ISBN 978-0-13-461099-3. (cited on page 2)

Ryan, M.; Antoniou, J.; Brooks, L.; Jiya, T.; Macnish, K.; and Stahl, B., 2021. Research and Practice of AI Ethics: A case study approach juxtaposing academic discourse with organisational reality. *Science and Engineering Ethics*, 27, 2 (2021), 1–29. (cited on page 97)

Sadowski, J., 2019. When data is capital: Datafication, accumulation, and extraction. *Big Data and Society*, 6, 1 (2019), 1–12. doi:10.1177/2053951718820549. (cited on page 96)

Sambasivan, N.; Arnesen, E.; Hutchinson, B.; Doshi, T.; Prabhakaran, V. V.-K.; and Prabhakaran, V. V.-K., 2021. Re-imagining algorithmic fairness in India and beyond. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (2021), 315–328. doi:10.1145/3442188.3445896. https://doi.org/10.1145/3442188.3445896http://arxiv.org/abs/2101.09995. (cited on pages 31, 34, and 45)

Sanchez-Cartas, J. M. and León, G., 2021. Multisided Platforms and Markets: a Survey of the Theoretical Literature. *Journal of Economic Surveys*, 35, 2 (2021), 452–487. doi:10.1111/joes.12409. (cited on page 20)

Sánchez-Monedero, J.; Dencik, L.; and Edwards, L., 2020. What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 458–468. New York, NY. doi:10.1145/3351095.3372849. (cited on pages 2 and 94)

Schiff, D.; Biddle, J.; Borenstein, J.; and Laas, K., 2020. What's Next for AI Ethics, Policy, and Governance? A Global Overview. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 153–158. ACM, New York, NY, USA. doi:10.1145/3375627.3375804. https://dl.acm.org/doi/10.1145/3375627.3375804. (cited on pages 43 and 44)

Schoonenboom, J. and Johnson, R. B., 2017. How to Construct a Mixed Methods Research Design. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 69, S2 (10 2017), 107–131. doi:10.1007/s11577-017-0454-1. http://link.springer.com/10.1007/s11577-017-0454-1. (cited on pages 62, 64, and 67)

Seaver, N., 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data and Society*, 4, 2 (2017), 12. doi:10.1177/2053951717738104. https://doi.org/10.1177/2053951717738104files/2750/Seaver-2017-AlgorithmsascultureSometacticsfortheethnog.pdf. (cited on pages 31 and 79)

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J., 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 59–68. doi:10.1145/3287560.3287598. (cited on pages 2, 34, and 38)

Shove, E., 2003. *Comfort, Cleanliness and Convenience: The Social Organization of Normality*. Berg. ISBN 9516981208. doi:10.5040/9781474214605. (cited on page 26)

Silipo, R., 2020. Machine Learning Algorithms and The Art of Hyperparameter Selection | by Rosaria Silipo | Towards Data Science. https://towardsdatascience.com/machine-learning-algorithms-and-the-art-of-hyperparameter-selection-279d3b04c281. (cited on page 5)

Silva, S. and Kenney, M., 2018. Algorithms, Platforms, and Ethnic Bias: An Integrative Essay. *Phylon (1960-)*, 55, 1 & 2 (2018), 9–37. https://www.jstor.org/stable/26545017. (cited on page 33)

Simon, J., 2020. New – Amazon SageMaker Clarify Detects Bias and Increases the Transparency of Machine Learning Models | AWS News Blog. https://tinyurl.com/3nf27bh7. (cited on page 49)

Smith, M., 2016. In Wisconsin, a Backlash Against Using Data to Foretell Defendants' Futures. https://www.nytimes.com/2016/06/23/us/backlash-in-wisconsin-against-using-data-to-foretell-defendants-futures.html. (cited on page 29)

Srnicek, N., 2017. *Platform capitalism*. John Wiley & Sons. ISBN 1509504885. (cited on pages 9, 21, 22, 23, and 30)

Star, S. L., 1999. The Ethnography of Infrastructure. *American Behavioral Scientist*, 43, 3 (1999), 337–791. (cited on pages 23, 24, 25, 26, and 27)

Stevenson, A., 2010a. enact. doi:10.1093/acref/9780199571123.013.m{_}en{_}gb0264020. https://www.oxfordreference.com/view/10.1093/acref/9780199571123.001.0001/m_en_gb0264020. (cited on page 41)

Stevenson, A., 2010b. enactment. doi:10.1093/acref/9780199571123.013.m{_}en{_}gb0264030. https://www.oxfordreference.com/view/10.1093/acref/9780199571123.001.0001/m_en_gb0264030. (cited on page 41)

Suresh, H. and Guttag, J. V., 2019. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. http://arxiv.org/abs/1901.10002. (cited on page 33)

Tashakkori, A. and Creswell, J. W., 2007. Editorial: The New Era of Mixed Methods. *Journal of Mixed Methods Research*, 1, 1 (2007), 3–7. doi:10.1177/2345678906293042. (cited on page 61)

Teddlie, C. and Tashakkori, A., 2006. A general typology of research designs featuring mixed methods. *Research in the Schools*, 13, 1 (2006), 12–28. (cited on page 65)

Teddlie, C. and Tashakkori, A., 2009. *Foundations of Mixed Methods Research*. SAGE. ISBN 9780761930112. http://marefateadyan.nashriyat.ir/node/150. (cited on pages 61, 62, 63, 64, 65, 66, 67, 70, and 74)

Tomalin, M.; Byrne, B.; Concannon, S.; Saunders, D.; and Ullmann, S., 2021. The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. *Ethics and Information Technology*, (2021), 1–15. (cited on page 97)

Turilli, M., 2008. Ethics and the practice of software design. *Frontiers in Artificial Intelligence and Applications*, 175, 1 (2008), 171–183. (cited on pages 36, 41, and 48)

Uber, 2021. Make driving more rewarding. https://developer.uber.com/products/drivers. (cited on page 22)

Vaughan, D., 1989. Regulating Risk: Implications of the Challenger Accident. doi:10.1111/j.1467-9930.1989.tb00032.x. (cited on page 48)

Veale, M.; Van Kleek, M.; and Binns, R., 2018. 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*, 1–14. ACM, New York, NY, USA. doi:10.1145/3173574.3174014. https://dl.acm.org/doi/10.1145/3173574.3174014. (cited on pages 41, 42, and 75)

West, S. M.; Whittaker, M.; and Crawford, K., 2019. Discriminating Systems: Gender, Race and Power in AI. Technical report, New York, NY. https://ainowinstitute.org/discriminatingsystems.html. (cited on pages 35 and 91)

Whittaker, M.; Alper, M.; Bennett, C. L.; Hendren, S.; Kaziunas, L.; Mills, M.; Morris, M. R.; Rankin, J.; Rogers, E.; Salas, M.; and Myers, 2019. Disability, Bias, and AI. Technical report, AI Now Institute At NYU. (cited on page 34)

Wiener, N., 1947. A scientist rebels. *Atlantic Monthly*, 179, 1 (1947), 46. (cited on page 37)

Wiener, N., 1961. *Cybernetics or control and communication in the animal and the machine*. MIT Press, Cambridge, Mass, 2. ed., 10 edn. ISBN 978-0-262-23007-0 978-0-262-73009-9. http://www.gbv.de/dms/ilmenau/toc/331213087.PDF. (cited on page 36)

Wilson, C.; Ghosh, A.; Jiang, S.; Mislove, A.; Baker, L.; Szary, J.; Trindel, K.; and Polli, F., 2021. Building and auditing fair algorithms: A case study in candidate screening. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (2021), 666–677. doi:10.1145/3442188.3445928. (cited on pages 2, 45, 46, and 78)

Wright, C., 1986. Routine deaths: fatal accidents in the oil industry. *The Sociological Review*, 34, 2 (1986), 265–289. doi:10.1111/j.1467-954X.1986.tb02702.x. (cited on page 48)

Yao, Y.; Viswanath, B.; Xiao, Z.; Zheng, H.; Wang, B.; and Zhao, B. Y., 2017. Complexity vs. Performance: Empirical analysis of machine learning as a service. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, Part F1319, 119 (2017), 384–397. doi:10.1145/3131365.3131372. (cited on page 2)

Zuboff, S., 2020. The age of surveillance capitalism: The fight for a human future at the new frontier of power. *Yale Law Journal*, 129, 5 (2020), 1460–1515. doi:10.1093/sf/soz037. (cited on pages 21 and 30)