From things to relations: co-construction of machine learning work with labels, ethics, and biomedical data

Oksana Dorofeeva

From things to relations: co-construction of machine learning work with labels, ethics, and biomedical data

PhD Dissertation

Politica

© Forlaget Politica and the author 2026

ISBN: 978-87-7335-363-9

Cover: Svend Siune
Print: Fællestrykkeriet, Aarhus University
Layout: Annette Bruun Andersen

Submitted 31 August 2026
The public defence takes place 27 November 2026
Published November 2026

Forlaget Politica
c/o Department of Political Science
Aarhus BSS, Aarhus University
Bartholins Allé 7
DK-8000 Aarhus C
Denmark

Table of Contents

Acknowledgements 7

Chapter 1 Introduction 11
1.1. Machine learning in a societal context 11
1.2. ML work in relation to (particular) regimes of normativity 12
1.3. Casing the study and arriving at the research questions 15
1.4. Outline of the dissertation 16

Chapter 2 Machine learning and machine learning work: a short introduction 19
2.1 Artificial intelligence and machine learning: a short history 19
2.2 Machine learning: a brief introduction 22
2.3 Infrastructural elements and supply chains underscoring machine learning 25
2.4 Machine learning work 27
2.5 A note on terminology 32

Chapter 3 Theory 35
3.1 The landscape of social research about science and technology 35
3.2 Meanings and representations of ML work 36
3.3 Labels as representations 39
3.3.1 Stuart Hall's codes and encoding-decoding 40
3.4 Representations of ethics in ML work 42
3.4.1 From ethics to 'ethics' 42
3.4.2 Developing the constellations metaphor: spatial metaphors and narrative infrastructures 44
3.4.3 Constellations 45
3.5 Friction in relations of ML work and (biomedical) data 46
3.5.1 From friction to friction in working relationships 47
3.6 Summary and the role of theory in this dissertation 49

Chapter 4 Methodology 53
4.1 Overall research design 53
4.2 Ontological-epistemological grounding 53
4.3 Expert interview pre-study 55
4.4 Interview main study 56
4.4.1 Sampling approach and recruitment 56
4.4.2 Sample 59
4.4.3 Interview guide 62
4.4.4 Ethical considerations 64
4.4.5 Transcription 65
4.4.6 Coding 65
4.5 Secondary data: conference ethical procedures 66
4.6 Analytical strategy 67

Chapter 5 Paper 1 – The Wrong Shortcut: 'AI' and representations of machine learning work 69

Chapter 6 Paper 2 – Where is ethics in machine learning? Constellations of ethics in representations of machine learning work 89

Chapter 7 Paper 3 – Where machine learning work meets biomedical data: friction in data settings 117

Chapter 8 Discussion and conclusion 137
8.1 Analysing vernacular concepts in computational cultures 138
8.2 ML work as both situated and embedded in broader narratives and structures 141
8.3 ML work and/in the societal context 144
8.4 Limitations of the study 145
8.5 Directions for future research 147
8.6 Conclusion 148

Summary 151

Dansk resumé 153

Bibliography 155

Appendix 175
Appendix I: Interview guide 175
Appendix II: Ethical approval for the interview study 179
Appendix III: Consent form for the interview study 181
Appendix IV: Secondary data sources list 183
Appendix V: Coding list for Paper 1 184
Appendix VI: Coding list for Paper 2 191
Appendix VII: Coding list for Paper 3 204
Appendix VIII: Generative AI use declaration 208

Acknowledgements

One of the tropes of a PhD dissertation's acknowledgements is to draw a parallel between the topic or findings of one's research and one's experience producing it. While you see a PDF or a physical booklet with one name on the cover, both the contents of this dissertation and the person named its author have been significantly shaped by various relations. It is a great joy to verbalise them – to the best of my ability – here.

First and foremost, I want to express my deepest gratitude to the people without whose time and stories this research would not exist – to researchers, data scientists, and engineers who participated in interviews and opened a window into their world for me. I would also like to thank everyone who helped with recruitment. While protection of the participants' anonymity makes unsung heroes out of those who connect the researcher and participants, please know that I remember and that this help has meant a great deal to me. Moreover, I want to thank various people I have bothered with questions about technical matters or other aspects of the world of machine learning (you know who you are – and thank you for your patience). I am also very grateful to experts who agreed to be interviewed at the very beginning of my project to help me find my footing and formulate the right guiding questions.

Several people have dedicated many working hours to this dissertation. Mads P. Sørensen and Tine Ravn, thank you both for your supervision in this quest. Thank you for making time for meetings and for allowing them to run long because I had more things to fret about, for getting through very (!) rough drafts with very little complaining, and most of all, for supporting me in pursuing this PhD in my own way. I am also very grateful to the student assistants who helped me with transcriptions: Simon, Signe, and Rasmus. I might always feel slightly awkward around you, knowing you had to listen to my voice for hours, but I am very glad I had your help. My sincere thanks also go to Annette Bruun Andersen for language editing. I would also like to thank my committee – Louise Amoore, Nanna Bonde Thylstrup, and Alan Irwin – for the time and effort dedicated to reading and evaluating my thesis, and for the great care with which you have done so.

To the place where it all mostly happened: CFA, or whatever name you will choose (you will always be CFA to me). To centre leader Carter Bloch: I thank you for hiring me and then sticking with me through the various bureaucratic adventures that neither of us had anticipated. To all of my CFA colleagues, former and current: you have all been a key part of this experience, and the small size of our unit has meant that with all of you we have shared our

moments, from academic discussions (thank you for all the feedback I got over the years!), to lunch conversations (often deliciously niche!), to chats in the hallway or near a coffee machine. It has been sad not to see my face on the new wall with pictures, and this is how I know I still feel a part of your collective. But I have extra-special thanks to give to the three people without whom I would genuinely not have survived this PhD journey: Emer, Simon, and Lucilla. Emer, I am so glad that we got paired in the office at a time when both of us were going through challenges. I am so happy sharing the office has only made us keener to spend time together outside of it. Thank you for being so patient with me, especially in the more manic times. Simon and Lucilla, my PhD elders (yes, I know you hate this word – this is the last time, I promise): thank you for welcoming me into CFA and for being the people to whom I still come with research dilemmas, social anxieties, or pop culture items I need to get out of my brain by sharing. Insert an appropriate Crazy Ex-Girlfriend song here.

The broader environment of the Political Science department, with its amazing human (and coffee and vibes) resources, has also added a lot to my PhD experience. I am very grateful for the opportunity to have taught a course here; to Martin Bækgaard for coming up with the summer course idea that has allowed me to teach, and to Florian Keppeler for showing me the ropes of course coordination here. I am also incredibly thankful for the students who participated in the course – for stepping out of their comfort zones and being such good sports about it (your zine is still one of my most prized possessions, and I brag about it every chance I get). I would also like to thank the energetic secretariat team at the Department (and at CFA) for all the support I have gotten in teaching and other endeavours, including accounting for my complicated train rides. And despite all its coffee, the lounge would not be the warm place it is without Njall and the service and kitchen staff: you're this place's critical infrastructure. Moreover, I am grateful for the many social activities organised by department members. Even someone as socially challenged as me has enjoyed many great hours discussing books and perfumes, playing badminton, doing crafts, and just having (more or less organised) drinks – thank you to everyone who organised and/or participated in these moments. A special shout-out has to go to MJ: it is a little funny that it has taken moving to Aarhus for us to meet and become friends, but I am very glad this happened. Keep them on their toes.

One of the great joys of a PhD is that you get to have your academic village, but you also get to meet so many wonderful people beyond it – and I think I have been exceptionally lucky in this regard. Some of the notable figures who shaped me as a researcher, I have met before the PhD: here, I send a ping of gratitude to Polina Kolozaridi, for my introduction to internet (and other

technology) studies, for treating me as an academic grownup from the start, and for a point of feedback I come back to at my most socially anxious. I also want to thank multiple wonderful teachers and colleagues I have had the honour to learn from at the University of Amsterdam. From my Aarhus years, I want to thank Marie Østergaard Møller: you have been so generous with your time and warmth, and I have so enjoyed talking to you about methods and other things, in the Netherlands and in Denmark. In the first year of my PhD, I was lucky to attend the WTMC summer school on exactly a topic I needed, taught by Tarleton Gillespie and other brilliant scholars. Insights (I have kept my paper notes!), discussions, and connections from that week have been disproportionately meaningful. Among them, a special place goes to Dwayne – I still don't know if PhD is truly a journey, but if it is, thank you for having been my fellow traveller. Vienna waits for you. The critical data studies course taught by Nanna Bonde Thylstrup, Klaus Hoeyer, and Helene Ratner (and co-organised by Louis Ravn) was also an invigorating moment for my research when I was lost in the woods. The same year, I managed to observe EASST 2024 and met two brilliant and kind people to whom I am indebted beyond what I can express. Vassilis Galanos, thank you for (co-)organising the vibrant hype community around you, for being such a champion for young scholars like me and believing in me when I did not believe in myself. Benedetta Catanzariti, there are no words for what you have done for me – your brilliance is only matched by your kindness, and your presence gives me hope for academia. Thank you for always being so welcoming, for introducing me to the work of Diana Forsythe (things just clicked!), for paving the way for my visit to STIS, all the guidance and opportunity-making during it, and for continuing to respond to my emails and pleas for help after it. I only hope I did not disappoint you too much as my informal, chosen (in a way a stray cat chooses a person) supervisor. Speaking of STIS – I have been lucky to visit the Department of Science, Technology and Innovation Studies at the University of Edinburgh in a beautiful spring. I want to thank Morgan Currie for patronage and say thank you to the many people in (and around!) the department who have been so generous with their time, insights, and coffee. I so hope we cross paths again soon. Special thanks go to the PhD dwellers of the 'library' room: Jana, Max, and Beren, you have all been wonderful pals, and I wish you the best of luck with finishing your respective PhDs (let me know if I can be of any help) and look forward to hanging out again.

I owe everything – including this dissertation and the title it should grant me – to the support of my friends and family. To Maddie, for being a caring friend in Denmark, a Teams work session buddy, and for helping me capture the constellations idea. To Cyrilein: I am so lucky to have you, and rarely sure I deserve you. Thank you for your long train rides, for noting I was talking in

Stuart Hall-isms before I had the bravery to commit to it, for being my wife when I needed one, and for accepting me with all my quirks; I hope in the future I can show up more for you and not just continually be saved by you. To Masha: you know there would not be this PhD and probably academic research in my life (for worse and for better), without you. Thank you for all the conceptual work we do together: seriously, but also coming up with academic 'love kernels', 'theory soup', and the upcoming serpentarium (tbd). I hope we get to properly work and write together soon and get to see each other and have non-academic fun. Моей Анюте: самой долгой дружбе в моей жизни, такой долгой, что мы уже больше семья. Ты мой лучик и глоток свежего неакадемического воздуха.

И самое главное: лучи любви и благодарности моей семье. Спасибо , что вы такой надёжный причал и тыл, что в мою молодость я могла и могу заниматься непонятной карьерой с неясными перспективами, просто потому что интересно. Всем-всем-всем за поддержку, заботу, и веру в мое будущее. Маме и бабушкам, за пример сильных женщин которые горят своим делом. Дашке, за то что ты есть.

Chapter 1
Introduction

1.1. Machine learning in a societal context

> The power of learning machines in society seems to be steadily on the rise. However, contrary to the feeling of wonder and magic that often surrounds machine learning, the politics and practices of learning machines are mundane. There is a huge amount of work and tinkering by the people that make machine learners. (Lee, 2025, p. 6144)

In the years I worked on my PhD research, machine learning, or rather its 'public-facing term' (Joyce et al., 2021), artificial intelligence, was all around me, and not just because this research focused on people who do this machine learning work. Machine learning (ML) and artificial intelligence (AI) concepts were prominently featured in people's lives and discourse: in books and movies, in traditional and social media publications, in emails and casual conversations.

A lot of this scientific and public discourse was exciting and hopeful. For instance, machine learning made the long-existing dream of self-driving vehicles feel close to realisation (Norton, 2021). In healthcare, intensive data sourcing and sophisticated data-driven techniques gave hope for personalised medicine and drug discovery, as well as answers to the pressures faced by healthcare systems (Al Kuwaiti et al., 2023). These hopes are also reflected in the valorisation of expertise in building and using these technologies, with data science named the 'sexiest job of the 21st century' (Davenport & Patil, 2012, October).

On the other hand, datafication and algorithmic systems have been accumulating concerns since entering the public spotlight. Concerns about state surveillance after Snowden revelations and the surveillance economy of user analytics and advertising, potential manipulation of public opinion related to the claims of Cambridge Analytica, and fake news have all been important elements of the cultural fabric of the 2010s (Elish & boyd, 2018). In 2016, tensions around tech companies manifested in the outburst of rage and public critique now termed 'techlash'(Su et al., 2021). As shown by Su et al., these critiques and emotions seeped into the lives of tech workers in Silicon Valley, changing their relationships with their social milieu and how they performed their professional identities, e.g., hiding their occupations on dating apps to avoiding displays of corporate memorabilia in public (Su et al., 2021). More

recently, the generative AI tools renewed controversies around data extraction and actualised new concerns: from the science fiction-like possibilities of existential threat from sentient artificial intelligence, to discussions about the planetary and infrastructural costs of computation at scale (Valdivia, 2025) and renewed 'automation anxiety' (Bassett & Roberts, 2023). Some of these concerns have also fuelled new movements of AI resistance (*The AI Resist List*).

Both discursive streams portray the significance ascribed to computational technologies as powerful and controlling our lives and as ineffable, opaque, and esoteric (Ziewitz, 2016). Trying to avoid further mystification of data and AI, scholars have argued for studying the mundane practices of making machine learning models (Lee, 2025) and treating AI as a 'normal technology' (Narayan & Kapoor, 2025). At the same time, it is important to acknowledge that ML/AI, as it is now, is not 'normal' because it is not treated as such. Politicians and policy-makers look at AI as the solution to economic growth and a tool of a new arms race (Bareis & Katzenbach, 2022). Computation based on data reuse speaks to dreams of various stakeholders (Stevens et al., 2022). Users anthropomorphise and ascribe intelligence to various technologies based on their combination of affordances and interface designs (Natale, 2021) and develop complicated affective relations with and folk theories about algorithms around them (Ruckenstein, 2023). As such, contemporary computational technologies are developed, deployed, and maintained under the contradictory valence of, in the words of Galanos and Stewart (2024), 'hypes, horrors, and hopes'. Considering technology as *socio*technical, co-constructed with society (Bijker & Law, 1992), means acknowledging this valence as part of the current ML/AI's sociotechnical system.

In this dissertation, I study this sociotechnical system by focusing on the ascribed meanings and practices of machine learning work. In doing so, I also take account of how meanings and practices of ML work relate to the broader social dynamics surrounding these technologies.

1.2. ML work in relation to (particular) regimes of normativity

In the last ten years, efforts to address societal concerns about data and algorithms have often coalesced around promoting *ethics*, resulting in AI ethics exploding both as an academic field and a form of governance (Stark, 2023). In addition to research papers, ML/AI ethics has been the topic of various guidelines (Franzke, 2022; Jobin et al., 2019) and AI strategies and policies (Kerr et al., 2020). These numerous projects show why Andrejevic (2020) described AI as undergoing an 'ethical turn'. Some of the studies focused on the

novel ethical issues presented by the expansion of computational prediction to various spheres of life and human activity data; they also noted that these developments bring data science and machine learning closer to the areas of concern of research ethics and other modes of governance that computer science has historically been distant from, raising questions about whether these oversight systems can handle the emerging technologies (Ienca et al., 2018; Metcalf & Crawford, 2016). Another stream of research critically questioned what ML/AI ethics is, who or what it works for, and whether it can lead to meaningful change and truly beneficial sociotechnical practices (Greene et al., 2019; Munn, 2023) and engaged in what Bietti (2020) termed 'ethics-bashing'. Some problematised ethics frameworks' abstraction and distance from practice (Munn, 2023; van Maanen, 2022); others noted that ethics is apolitical, and that framing ML/AI harms as ethical issues obscures the power dynamics, which is why inquiry and discussions need to shift from ethics to politics (Andrejevic, 2020; D'ignazio & Klein, 2020). Moreover, a stream of critical scholarship raised concerns about corporate capture and 'ethics-washing' and argued that ML/AI ethics is shaped by and promotes the interests of tech companies (Phan et al., 2022; Steinhoff, 2024; Wagner, 2018).

While these debates are important, for me they also raised questions about whether these different discourses and practices reached researchers and developers 'on the ground' and contributed to how practitioners understood and went about their work. Previous findings indicate that computational cultures abstract the social and political dimensions of technology (Malazita & and Resetar, 2019) and portray ethics as absent from their work (Barocas & boyd, 2017; Slota et al., 2023). I was curious whether the 'ethical turn' changed that.

Social and humanistic studies of computational technologies have produced a rich body of research on how practitioners construct ethics (e.g., Avnoon et al., 2024; Bennett, 2025; Orr & Davis, 2020; Widder & Nafus, 2023; Widder et al., 2023). However, less is known about what shapes these emic notions and how they interact with institutionalised ethics approaches. Notable examples of studies eliciting such connections are the analyses of how Silicon Valley culture shapes ethics approaches developed there (Metcalf et al., 2019), and how views of ethical issues vary across the AI supply chain (Widder & Nafus, 2023). Moreover, Samuel et al. (2021) found that health AI researchers constructed two distinct notions of ethics, one of which is synonymous with data governance, while the other is about broader concerns about the uptake of AI in society and its implications.

As noted by Helm and Gerlek (2026), STS sensibilities would suggest that as a knowledge and/or governance endeavour, AI ethics co-constitutes its object. Ziewitz's ethnography of search engine optimisation (2019) highlights that when an obligation to 'be ethical' arises as a form of regulation, it

introduces situated negotiations of what that means in practice. Hence, at the start of my doctoral studies, my hope was to better understand the emic notions of ethics in machine learning by considering their relationships to this broader set of discourses and practices *around* ML/AI, and especially the proliferating ethics-focused endeavours.

Another point of discussion in debates around ML/AI ethics was that it promises self-regulation to avoid 'real' regulation (Wagner, 2018). This is why addressing the issue of ML/AI ethics involves considering its relationship with regulation (Delacroix & Wagner, 2021). Here, Delacroix and Wagner (2021) bring up biomedicine as an exemplary sphere in which ethics and regulation have a defined 'interface'. Again, a thorough discussion of how this interface should work is well beyond the focus of this dissertation. However, these debates highlight that ethics is a particular 'regime of normativity' (Pickersgill, 2012) among others. These other regimes of normativity could be based on legal forms of governance, such as, in the European context, the General Data Protection Regulation (EU), 2016/679 (GDPR) and (at that point upcoming) Artificial Intelligence Act (EU), 2024/1689. Hence, I was interested in how these different 'regimes of normativity' would interact in practice and affect ML development.

Importantly, these debates emphasised that machine learning models are not built and deployed in a vacuum. Thinking about ethics and regulations helped me consider ML work as embedded in various institutional and cultural contexts and has directed my analytical attention to these connections. For example, academic ML/AI research is situated within the broader dynamics of the scientific system, such as the uneasy marriage between publication-centric regimes of valuation and the promotion of ethical, legal, and societal aspects of technology (ELSA) and responsible research and innovation (RRI) programs (Felt, 2017; Sigl et al., 2020). On the other hand, medical ML/AI work would interact with medical regimes of normativity, from different approaches to responsibility (Stevens & Beaulieu, 2024) to the more explicit connection to legal regulation mentioned by Delacroix and Wagner (2021). Such contextual factors would contribute to the diversity within the distributed knowledge infrastructures and communities of practice working on ML/AI.

This interest in machine learning's embeddedness in these broader contexts has inspired and guided this dissertation project. In the next section, I will show how this focus transformed throughout my doctoral studies, resulting in this dissertation. By explaining the journey, I will also introduce the resulting research questions and provide the roadmap to the dissertation.

1.3. Casing the study and arriving at the research questions

As described by sociologist Annette Lareau (2021), qualitative research is an exercise in thinking as you go. In a similar vein, political scientist Joe Soss (2021) argues that interpretive research is often more about *casing a study* through analysis than about *studying* a predetermined *case*. This also applies to my doctoral research as it gradually transformed through my engagement with the literature and, most importantly, with ML practitioners' narratives. It thus also grew into new research questions from those I formulated at the beginning of the PhD.

Initially, I set out to explore how big data researchers understand and navigate ethics and governance. Already then, the conceptualisation and naming of the community I was interested in gave me 'terminological anxiety' (Seaver, 2017). I had a feeling that even though 'big data' captured the datafied knowledge production at scale (Bonde Thylstrup et al., 2019) and continued to appear in the literature, as a more experience-close term it was dying or already dead (Elish & boyd, 2018). Within the first few months, an inspirational critical data studies scholar giving feedback on my project proposal said: you probably mean data scientists? – and urged me to consider what different terms would do for my study. Based on my reflection on the fleeting nature of these technology labels (Pollock & Williams, 2011) and not wanting to rely on buzzwords as organising concepts for my study, I moved away from 'big data'. Being interested in work done across contexts, I also did not want to limit the inquiry to a specific occupational title like data scientist (Avnoon, 2021). At the end of this deliberative process, machine learning (ML) became my way of focusing on a community of practice connected by shared knowledge objects (ML models) and practices across cultures and institutions (Clarke & Star, 2008). So, I designed the interview study to examine the relationship between machine learning work and ethics and governance.

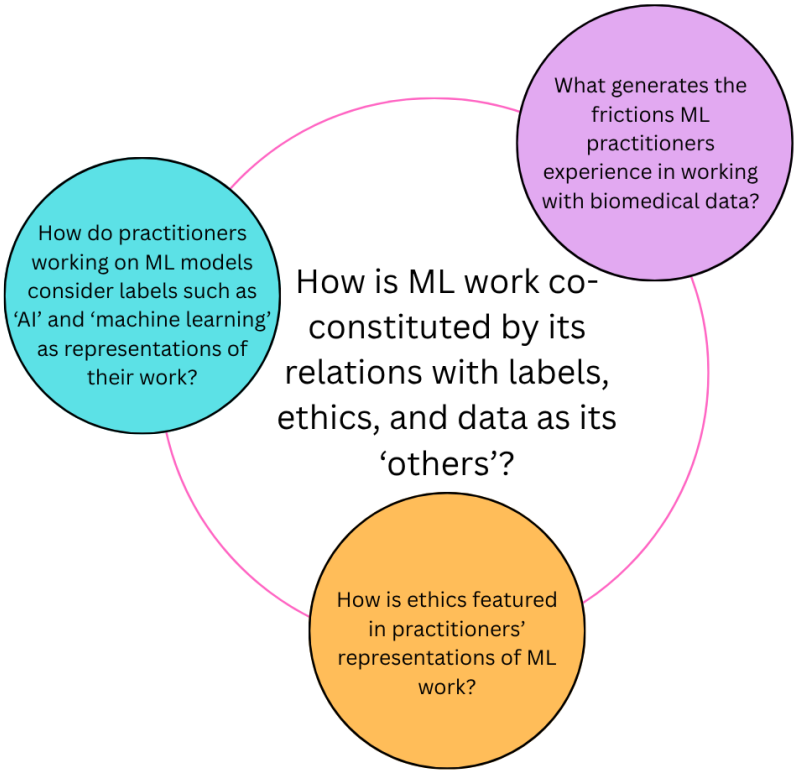
Throughout the data collection and dialectic conversation between data and theory in analysis, the focal points and emphases of the research project have shifted. The resulting integrative question of the dissertation is the following:

> How is machine learning work co-constituted by its relations with labels, ethics, and (biomedical) data as its 'others'?

This question articulates my approach to ML work as embedded in sociotechnical systems that include 'others', related but (analytically) distinct elements (Dourish, 2016). Drawing on the shared interview data corpus, three empirical studies within the project investigate ML work in its relation to one such

'other': labels ('machine learning' and 'AI'), ethics, and biomedical data. Figure 1.1 illustrates the individual research questions for each analysis together with the integrative one.

Figure 1.1. Overview of the three individual and one integrative research questions



In framing the study this way, I approach machine learning work not as a thing in itself but as made of relations among various meanings, materialities, and practices. The exploration of each 'other' then foregrounds one of such relations as an analytical angle from which I look at the work of machine learning. By analysing these relationships based on the narratives of machine learning practitioners, I demonstrate how machine learning work is both situated and embedded in broader contexts.

The remainder of this chapter provides a roadmap for reading the rest of the dissertation.

1.4. Outline of the dissertation

Having introduced the dissertation's subject and research questions in this first chapter, I contextualise the inquiry in Chapter 2 by providing more background information on machine learning and machine learning work. First, I outline a brief history of the interrelated development of machine learning and artificial intelligence in the 20th and 21st centuries (which will turn up again in

the discussion of these categories in Chapter 5). Second, I provide a concise introduction to machine learning algorithms and models and consider the broader supply chains and knowledge infrastructures needed to make these models practically doable. Next, I locate ML practitioners in these supply chains and discuss the stages and elements of ML work and their interactions with other categories of actors. I end the chapter on a terminological note.

In Chapter 3, I build the theoretical toolkit of the dissertation, introducing the general constructivist sensibilities I take from social studies of science and technology and developing the foundation of conceptualising meanings and representations of ML work, in relation to which I explore all the topics of the individual articles. Then, I delve into the conceptual frameworks I developed to analyse the 'others' of the dissertation. First, I introduce my conceptualisation of practitioners' relationship with labels in the context of the historical dynamics around the concept of 'artificial intelligence' and the technologists' active role in reconstructing the meaning of categories in which they define their work. To do so, I draw on Stuart Hall's theory of representation and the encoding-decoding model of communication. Within this framework, I incorporate the concept of boundaries. Second, I describe my conceptualisation of the practitioners' representations of the relationship between 'ethics' as a situated and socially constructed concept and their work. For that, I draw on the use of geographical metaphors in STS to develop the conceptual metaphor of constellations. Third, I outline how my attention to notions and meanings of ML work has formed my approach to friction through the notion of relationality, for which I draw on infrastructural thinking. I finish the chapter by discussing the integrative approach these theoretical tools form and discuss the role of the theory in my knowledge production process.

In Chapter 4, I discuss the data and methods of the empirical analyses I executed in the project. First, I introduce the ontological-epistemological basis of the project, drawing from interpretivist social sciences and grounded in the notion of situated knowledge. Based on this, I describe the expert interview pre-study that informed the design of the main empirical basis of the dissertation: an interview study. After discussing interviewing as a method, I get into the specific choices and contingencies of the study: the sampling and recruitment approaches, the resulting sample of the interviewees and its limitations, the interview guide and topics. I also discuss the limitations and the ethical considerations of the study. I describe data handling and coding and introduce the secondary data source I added during analysis: ethical requirements of computational conferences. Finally, I outline my abductive analytical strategy.

Chapters 5 to 7 comprise the analysis part of the dissertation in the form of academic articles. Table 1.1 provides an overview of the three articles and the research questions they answer. All three articles draw on the main

interview study but explore different facets of the data. This single vantage point unites the thematically diverging chapters. Paper 2 also draws on the secondary data (ethical requirements of computational conferences).

Table 1.1. Overview of empirical analyses included in this dissertation, all single-authored

Paper/ Chapter	Title	Research Question	Publication information
Paper 1/ Chapter 5	The Wrong Shortcut: 'AI' and representations of machine learning work	How do practitioners working on ML models consider labels such as 'AI' and 'machine learning' as representations of their work?	Submitted
Paper 2/ Chapter 6	Where is ethics in machine learning? Constellations of ethics in representations of machine learning work	How does ethics – the way ML practitioners themselves understand it – fit into representations of their work?	Under review
Paper 3/ Chapter 7	Where machine learning work meets biomedical data: friction in data settings	What generates the frictions ML practitioners experience in working with biomedical data?	Not submitted

Finally, Chapter 8 summarises the results of the empirical analyses and discusses the overall findings, contributions, and limitations of the study. I finish by charting some possible future research directions and offering some concluding reflections.

Chapter 2
Machine learning and machine learning work: a short introduction

The first chapter has just cracked a window into the topics of the dissertation. In this chapter, I provide a contextual backdrop against which my inquiry into ML work and its others can unfold. The chapter also introduces the reader to vocabulary that will be featured in the following chapters. I start by outlining a brief history of machine learning and its relations to artificial intelligence and statistics in the 20th century and in contemporary times. Thereafter, I give an introduction to machine learning models and the infrastructures and supply chains that afford machine learning practice. Finally, I locate machine learning work in these supply chains and outline its stages and dynamics, including interactions with other actors.

2.1 Artificial intelligence and machine learning: a short history

Any history is a construction of the history's subject and the definition of its significant (and thus worth tracing) events or features. In the context of proclamations of the revolutions imposed by big data, machine learning, or (generative) AI, historical approaches are also critical tools to specify and demystify shiny new technological projects (Suchman, 2023). However, the aim of this section is not to provide a thorough genealogy of machine learning but to set the scene for the contemporary dynamics discussed in the empirical chapters. To do so, it provides a brief history of ML and AI as research subfields and technologies (Suchman, 2023) in the 20th and into the current century. This is not to deny the 'long now' of these sociotechnical systems, drawing on the centuries-long interweaving of technology and power (Crawford & Joler, 2023), long-standing dreams of automation (Jones-Imhotep, 2020), histories of quantification in science and governance (Hacking, 1990; Porter, 1995), and more.

The origins of contemporary machine learning are often traced to the statistical classification methods developed in the 1930s and 1950s (Dobson, 2023) and the early cybernetics movement of the 1940s-1950s, which studied information processing and transfer in both human brains and machines (Olazaran, 1996). In the 1950s, two main computational approaches to information processing emerged (Cardon et al., 2018; Olazaran, 1996). The 'connectionist' approaches modelled information processing by 'learning' from

statistically representative input-output pairs and iteratively adjusting the weighted connections between them. This approach was embodied in neural networks as the model of 'learning' inspired by nervous systems, where information is processed through patterns of activation in an interconnected structure of artificial neurons, and the 'learning' takes place through adjusting the weights connecting them (Dobson, 2023). Another paradigm, reinforcement learning, drew on an explicitly behaviourist approach, focused on trial-and-error interaction to produce associations between stimuli and responses; notably, its elements were also used in (some) neural networks (Dobson, 2023). By contrast, the 'symbolic' approach emphasized explicit knowledge representations and rule-based reasoning (e.g., true-false, if-then) and manipulation of arbitrary symbols. The proponents of the latter, John McCarthy and colleagues, coined the 'artificial intelligence' term in a 1956 proposal for a two-month Dartmouth Summer Research Project (McCarthy et al., 2006) and led this new field until the early 1990s with symbolism as a dominant paradigm (Cardon et al., 2018).

Over the next decades, both symbolists and connectionists went through 'the tides of investment, interest, and perceived progress' that are often described as AI summers and winters (Galanos, 2023, section 2, para. 1). Both approaches saw significant developments in the 1950s-60s. However, as their ambitious promises outpaced results, funding dropped, and scepticism grew from the early 1970s onward, a period referred to as the first 'AI winter' (Cardon et al., 2018; Olazaran, 1996). The connectionist paradigm was further marginalised by the publicised critique from the leaders of the symbolic camp (for a detailed analysis of the controversy, see Olazaran, 1996). From then on, the connectionist stream continued, but often in affiliation with other disciplines and fields (Galanos, 2023; Olazaran, 1996). Many of the techniques associated with today's deep learning, such as backpropagation algorithms and convolutional neural networks, were developed or revised during the 1980s-1990s (Cardon et al., 2018). Notably, neural networks of the second half of the 20th century were often developed for visual pattern recognition (such as convolutional neural networks), which embodies the entanglement of computer vision and machine learning research (Dobson, 2023). Meanwhile, the symbolic perspective regained resources in the 1980s with the development of 'expert systems' that encoded professional knowledge to perform real-life tasks. These systems were actively studied and critiqued by STS and other social-science approaches (Collins, 1990; Forsythe, 1993; Star, 1989; Suchman & Trigg, 1993). Expert systems' limited performance and unfulfilled promise of commercial applications resulted in a second AI winter in the 1990s (Cardon et al., 2018).

In parallel, developments in computational methods and statistical theory have furthered the development of approaches often referred to as 'statistical machine learning' (e.g., naive Bayes classifiers, decision trees, kernel methods, support vector machines) (Cardon et al., 2018). The distinction between statistical modelling and statistical machine learning is contested but often attributed to the modelling assumptions and priorities (Breiman, 2001). Statistical modelling would begin with a specific stochastic model of the relationship between variables and seek to estimate its fit and parameters. Statistical ML approaches would more often treat this relationship as unknown, placing greater emphasis on prediction, empirical performance, and generalisation than on explanation and parameter inference (ibid.). Throughout the 1990s and 2000s, statistical machine-learning approaches such as kernel methods often outperformed neural networks and were regarded as more computationally efficient while offering stronger theoretical guarantees regarding generalisation (Breiman, 2001; Cardon et al., 2018).

The connectionist paradigm came back into the spotlight in 2012, after a 'deep' – as opposed to 'shallow' architectures such as support vector machines, a term coined by Hinton in the noughties – convolutional neural network approach called AlexNet won the ImageNet Large Scale Visual Recognition Challenge, outperforming competitive approaches by a large margin on an image classification task (Cardon et al., 2018). On unstructured data (such as images comprising ImageNet) and by using larger volumes of training data and computational resources, deep learning approaches proved they can outperform statistical machine learning approaches. This breakthrough led to further acceleration of research and development in deep learning, both in prediction and generative tasks: production of new data units (for example, images) based on an original data set and prescribed parameters. These developments are underscored by broader infrastructures around them: the produced large training datasets (often produced through extensive and precarious human annotation labour), advances in computing hardware, and the growing resources of technology companies and research labs. Following the successes in deep learning, many fields, such as computer vision, natural language processing, emotion recognition, and more, have increasingly used deep learning methods, establishing connections and blurring the boundaries between fields based on the use of shared methods (Berman et al., 2025). Based on these developments, several studies have noted that computational methods are being positioned as universal solutions for any and all 'domains' of knowledge and activity (Ribes et al., 2019; Rotem & Shani, 2025; Slota, Hoffman, et al., 2020). The current state-of-the-art for many generative tasks, especially with language and multimodal models, draws on the transformer architecture introduced in 2017 for sequence modelling in natural language processing

(Vaswani et al., 2017). In that period, 'AI' becomes almost synonymous with neural networks, while the symbolic origins of the term fade into the background (Galanos, 2023) or, rather, are repositioned as a previous stage in a linear development trajectory, as 'good old-fashioned AI' (GOFAI) (Smith, 2019). However, despite this distancing from the 'AI' notion, the symbolic and especially the statistical machine learning methods continue to be used both in research and in practice.

During the 2010s, ML/AI models were, in the words of Sheikh et al. (2023), actively 'leaving the lab and entering society'. By then, related developments in internet technologies (on the mutual production between AI and internet, see Galanos, 2023) and growing computer and internet use have digitalised many world activities, from science to entertainment, resulting in significant volumes of data. The buzzword 'big data' captures the 2012-2015 discourse surrounding data sourcing and analysis, presenting them as revolutionary sources of insight and revenue (Elish & boyd, 2018; Pentzold & Knorr, 2024). This 'data imperative' helped consolidate data science as a distinct occupation that integrates data gathering, analysis, and machine learning (Avnoon, 2024). Data scientists and machine learning engineers and other practitioners do predictive analytics and build algorithmic systems actively making decisions and shaping the online and offline worlds around them, from content moderation (Gorwa et al., 2020) and recommendation systems for platforms (Seaver, 2022), financial trading methods (Hansen, 2020; MacKenzie, 2021) to border control (Amoore, 2020, 2024). In doing so, they engineer ML systems as the 'nuts and bolts' that build certain orders of our world (Winner, 1980). These practices have continued, even though the 'big data' buzz faded, at least partially due to surveillance connotations and concerns (Elish & boyd, 2018). So, 'by late 2015, technology companies that were once seen as being at the forefront of Big Data began rebranding their efforts as "AI"'(Elish & boyd, 2018, p. 61). In November 2022, the company OpenAI released a new type of AI product in ChatGPT, a large language model based on the transformer architecture, made accessible to users without coding skills via a chatbot interface. After its rapid adoption, the market for generative AI expanded with models covering more and more use cases. Today, we are witnessing the large-scale 'industrialisation of AI' (van der Vlist et al., 2024), accompanied by intensive discourse about (generative) AI in academic as well as media outlets (Lammar et al., 2025; Magalhães & Smit, 2026; Markelius et al., 2024; Munk et al., 2024).

2.2 Machine learning: a brief introduction

In this section, I sketch the components and types of machine learning. I borrow extensively from the accessible overview created by Pasquinelli and Joler

(2021), as well as other work. Pasquinelli and Joler define the composition of machine learning as a knowledge instrument as comprising three main elements: 'object to be observed (*training dataset*), an instrument of observation (*learning algorithm*), and a final representation (*statistical model*)' (2021, p. 1265, original emphasis). These elements are intertwined, and the distinction is more analytical and/or discursive. For instance, Pasquinelli and Joler note that it is difficult to distinguish between algorithm and model; in the developer community, these terms are often conflated (p. 1268). Moreover, scholars have argued that data and algorithms entangle in iterative, non-linear ways during training, which is why they cannot be isolated from each other (Amoore, 2020; Thylstrup et al., 2022).

The dataset would include input data as well as potentially target output data (Jaton, 2025). Modelling aims to represent the patterns based on this training dataset in ways that would generalise beyond it. Hence, scholars have argued that datasets play a formative role in machine learning, as algorithms are derivatives of these referential datasets (Jaton, 2017, 2021). As such, patterns and categories included in the dataset constitute some (not all) of its key sociopolitical features and sources of potential harms of algorithmic systems developed on this dataset (Crawford, 2021). In the following sections, I will elaborate on the production of these datasets.

ML approaches are also differentiated based on the relationship between data and algorithm in 'learning' and evaluation. The supervision criterion classifies methods into unsupervised, semi-supervised, supervised, and reinforcement learning; 'classical' or 'statistical' ML as well as deep learning both include supervised and unsupervised methods. Examples of unsupervised approaches that aim to derive patterns directly from the data are exploratory approaches familiar to social scientists, such as the k-means technique (Jaton, 2025) and unsupervised or self-supervised neural-network algorithms such as autoencoders (Bengesi et al., 2024). In supervised approaches, the referential dataset would include both input data (for instance, images or texts of emails) and one or multiple labels (indicating the content of the images and whether a given email is spam), which can be treated as 'ground truth'. Based on a subset of these data (training dataset), ML algorithms approximate functions that map inputs to outputs; their performance is then evaluated on another subset of the referential dataset. As these referential datasets shape both 'learning' and evaluation (which informs the choice of resulting model), we, as evocatively characterised by Jaton, 'get algorithms of our ground truths' (Jaton, 2017). While the influence of ground truth data is most obvious in supervised learning, it is also highly important for self-supervised learning, such as transformer models (including large language models), although in a different way (Jaton, 2025). As Jaton argues, instead of separate labels, supervision signals

in these approaches reside within the input data itself. For example, in training data for large language models, the ground truth references are the relation between consecutive word/tokens, based on which models learn to predict the next word.

The learning algorithm can take various architectures, some of which were mentioned above and in the history section: 'statistical' ML algorithms (kernel methods, support vector machines, decision trees) and neural networks. Both 'learn' patterns from training data, but 'statistical' ML algorithms are more constrained in their ways of representing the data. For instance, decision-tree algorithms organise the data through a series of branching decisions that partition observations in the data into progressively more homogeneous groups; support vector machine (SVM) algorithms create decision boundaries that separate categories in a feature space (Breiman, 2001). SVMs and neural networks both rely on vector representations. However, the latter do not rely on a specific representational strategy such as recursive partitioning or construction of decision boundaries in a predetermined feature space but iteratively adjust parameters, creating multiple intermediate representations of data. Through gradual adjustment of parameters, this approach is said to be capable of approximating a wide range of input-output functions with sufficient training and architectural complexity (Pasquinelli & Joler, 2021, p. 1271). Neural-network algorithms also have different architectures and can vary in complexity; contemporary neural-network architectures often involve millions of parameters controlled by *hyper*parameters that determine key architectural and training choices, e.g., number of neurons and 'layers' in the network (Pasquinelli & Joler, 2021, p. 1269, my emphasis).

The goal of this training is to produce a model that will 'elegantly *fit* only the important patterns of the training data and apply those patterns also to new data "in the wild"' (Pasquinelli & Joler, 2021, p. 1271, original emphasis). Hence, this resulting model can be used for further tasks, such as classification and prediction/generation of new patterns. It can also be further developed by training on additional data and/or tinkering with the algorithm architecture.

Machine learning and especially deep learning have been discussed and critiqued for their opacity and susceptibility to bias. These characteristics are generally attributed to algorithmic modelling that treats the input-output relationship agnostically (Breiman, 2001) and the transformation of input data into multidimensional vectors for modelling (Pasquinelli & Joler, 2021). In addition, the complex architectures of neural networks and their entanglement with datasets through iterative processing (Thylstrup et al., 2022) make it difficult to know the models' chains of information processing and pattern

recognition, including bias and mistakes. This opacity is often discussed as machine learning's 'black box' problem.

Having outlined the broad contours of machine learning, I now turn to the infrastructures and work needed to make machine learning 'doable' (Fujimura, 1987) in practice. In doing so, I also address the materiality and the social and environmental costs of machine learning.

2.3 Infrastructural elements and supply chains underscoring machine learning

Early machine learning models often operated on highly specialised computation devices. For instance, Perceptron, an influential early neural network, was created as both an algorithm and a physical device that would implement it (Dobson, 2023). Digital and distributed, today's computation seems more ethereal; however, ML/AI models are underscored by complex global supply chains and infrastructures, with unequally distributed environmental and social costs (Crawford, 2021; Valdivia, 2025). This section gives a broad-strokes description of the three key resources involved in the production of machine learning (hardware, data, and software) and their supply chains. Access to data and to hardware and software resources, often gathered under the umbrella term 'compute', significantly shape development and contribute to inequalities *among* ML projects (Vipra & West, 2023).

First, machine learning draws on the electronics supply chain for all the hardware involved in the distributed ML systems, from processing chips to cameras and sensors, cables, and more. All these incorporate minerals produced by mining, for instance, lithium, a key component of rechargeable batteries, and various rare minerals used in the production of electronic devices (Crawford, 2021). Training of large deep learning models is powered by Graphic Processing Units (GPUs), microprocessors made of silicon and other minerals that perform parallel (rather than sequential) computing, allowing them to handle computationally intensive tasks (such as training a deep learning model) much faster (Valdivia, 2025). With the spread of deep learning, the demand for GPUs has been accelerating; they are an expensive (one of the highest costs in development) and key resource for training deep learning models, and access to them shapes the tech industry in significant ways (Vipra & West, 2023). Compared to them, 'classic'/'statistical' machine learning methods such as decision trees and SVMs require significantly less computational power and have lower energy costs (Lottick et al., 2019). The mining and electronics manufacturing industries also require shipping, consume energy, produce carbon emissions, and release electronic and toxic waste, which is often regarded as the final step of the supply chains (Valdivia, 2025).

The previous section posited that machine learning relies on data to draw patterns from. While computational methods were often positioned as a solution to the existing 'data deluge' overwhelming existing knowledge and informational infrastructures (Kitchin, 2014), data often needs to be produced, travel, and/or transform to be (re)used in ML/AI. For example, the ImageNet dataset, used in the training of the AlexNet, the neural network approach that was seen as a revolution in deep learning (see section 2.1, this chapter), was a massive project led by Princeton and Stanford computer vision researchers (Deng et al., 2009). The dataset combined large amounts of images from the internet with the categorical system of a database of English words (WordNet) by relying on the newly available crowd work platform Amazon Mechanical Turk to annotate the images according to the concepts they featured (for a critical history of ImageNet, see Denton et al., 2021). Depending on the context, the data production and data 'journey' to where they meet machine learning algorithms can look very different. For example, a computer vision researcher can record a video as 'toy data' for training. Data can come from repositories of existing open-source datasets, created specifically for training and evaluating models (often called 'benchmark' datasets), such as ImageNet; they can also be extracted from the Internet. Another source of data is datafied organisational processes (such as electronic hospital records), which might be tailored to produce large volumes of data for (re)use (Hoeyer, 2023). Data can also come from data vendors and brokers (Kitchin, 2022) and other actors. Recently, generating synthetic data has gained attention as a solution to data limitations (Offenhuber, 2024). Data often need to be labelled for machine learning. These tasks are sometimes done by ML practitioners (Muller et al., 2019) but are often outsourced to collaborators (see the next section), specialised annotation companies, and crowd workers on platforms like Amazon Mechanical Turk and Prolific. These practices can contain extractivist and exploitative dynamics due to power imbalances between annotators and platforms and clients (Miceli & Posada, 2022). Access to (re)usable data or resources to produce new datasets also differentiates companies and labs in the AI landscape (Orr & Crawford, 2024).

Hardware (GPUs) and data are commonly stored in data centres, digital warehouses that provide the infrastructure for data storage and processing at a distance, or what is colloquially referred to as 'the cloud'. Large data centres host supercomputers, clusters of GPUs, which provide intensive computation used in training large-scale deep learning models (Vipra & West, 2023). Cloud computing is often provided as a service by Big Tech companies, and other developers have to rely on these infrastructures (van der Vlist et al., 2024; Vipra & West, 2023). Through their infrastructural involvement in this and

other elements of the supply chains, big companies shape technology development (van der Vlist et al., 2024).

The AI supply chain also includes software (Valdivia, 2025), some of which is proprietary and is sold as products and services. van der Vlist et al. (2024) give an overview of some of the pre-trained models, APIs, and other developer tools provided by Big Tech that can be used for training and deploying machine learning models and building systems. However, ML/AI development also draws on an array of open-source tools and models, which can be free to use with some licensing restrictions (Widder et al., 2024). They can be developed by various actors, including the Big Tech companies. For instance, Thomas (2019) mentions that computer vision developers use software libraries like Google's TensorFlow for deep learning.

These supply chains illustrate that ML engineers do not build the technology from scratch but rather draw together distributed resources, artefacts, and expertise. Thomas (2019) describes that even though computer vision development is done in small, co-located, organisationally distinct teams, it is by no means isolated, as 'these professionals turn to external smart camera manufacturers, competitive multinational and regional cloud service providers, and a rapidly maturing open source software and artificial intelligence (AI) algorithm community to acquire the necessary computer vision software know-how and resources' (p. 12). As ML practitioners draw on knowledge and tools developed by other ML researchers and practitioners, Widder and Nafus (2023) characterise AI development as a supply chain of its own from basic knowledge to end-use software. Upstream in that chain, researchers develop general ideas, new algorithms, and architectures, 'foundational' models that can be further developed for more particular uses, and proofs-of-concept (proving that something can be done). Practitioners downstream from them draw on these ideas, technologies, and models to create their own ML models and systems for particular use contexts.

With the context of this broader infrastructure supporting ML work, I will now delve into what that work involves.

2.4 Machine learning work

Based on the stages of ML and data science projects noted by previous research (Avnoon, 2024; Bechmann & Bowker, 2019; Wirth et al., 2025; Zhang et al., 2020), the life cycle of machine learning development can be portrayed as consisting of the following elements: 1) problem formulation and goal-setting, 2) data selection and acquisition, 3) data preparation, 4) model-building, 5) model evaluation, 6) creating the final 'product', which could range from communication of results to producing the model to deploy, and 7) deployment and maintenance: monitoring of performance and potentially re-

training. While these stages can help navigate the world of machine learning work, researchers have argued that development is more iterative than such a linear 'pipeline' (Bennett et al., 2025; Wirth et al., 2025). As I will show below, these elements also differ depending on the type of project.

Project formulation and goal-setting also include the operationalisation of the project into machine learning terms, e.g., choosing the output variables. Based on the project, practitioners decide what data they need to create, procure, or make use of (if they are already available to them). Empirical studies have demonstrated the intricacies and contingencies of creating new 'ground truth' datasets (Henriksen & Bechmann, 2020; Jaton, 2017), as well as how various contextual factors can affect the generated data's quality (Zając et al., 2023). These data also need to be prepared for modelling. Muller et al. (2019) demonstrate that even though the idea is that the sophisticated data analysis methods would derive patterns from any data, only some data (such as specialised 'benchmark' datasets) are algorithm-ready. Different sets of data might need to be combined and aligned (for example, data from different sensors); data might need to be cleaned, anomalies detected and transformed. For training and evaluation, data often needs to be annotated, e.g., images paired with the category label(s). This labour-intensive task is often done by other actors (see below), with ML practitioners potentially creating the annotation interfaces and/or schemas.

Scholars have argued that data cleaning, filtering, and transformation create particular representations of the world. For training, practitioners often select, transform, and generate what are called features of the data (certain variables) to be used in the modelling (Lee, 2025). While the training dataset could potentially contain all possible solutions, each additional feature also increases the dimensionality of the data representation and computation costs. Hence, reducing the dimensionality of data would make computation more feasible as well as avoid potential spurious correlations. As Lee describes, by including certain features, technologists create representations of the phenomenon of interest and what is important in it (the chosen features). Depending on the data and algorithm used, a practitioner would do more or less manual feature engineering (the creation of informative variables from unstructured data). Practitioners often select a subset of existing data for ML development, thus creating 'slices' as particular representations of the overall dataset and the datafied phenomena (Schjøtt, 2026). As described by Lee (2025), the transformation of features into mathematical objects (such as vectors and numbers) includes decisions about what and how to quantify; these decisions form specific ways of seeing and understanding the world; for instance, vectorisation promotes spatial and geometrical conceptualisation of the phenomena of interest (p. 6139-6140). These insights highlight the

amount of work, choice-making, and tinkering in ML work. Through these practices, to borrow an expression from Bechmann and Bowker (2019), even 'a seemingly unsupervised model becomes extremely supervised' (p.7).

Model development starts with model (type) selection, for instance, starting with a new (not yet connected to any data) or a pre-trained or foundational model, which would then be further developed (Bechmann & Bowker, 2019). Engdahl (2024) ethnographically describes how adapting open-source 'foundational' neural networks for more specific purposes requires the contextual reconstruction of the model. This work can involve generating supplementary datasets for re-training, adding algorithms, finding new ways to use metadata, tinkering with hardware (e.g., cameras), or navigating local regulatory and infrastructural restrictions, eventually resulting in recalibrating the weights of the pre-trained neural networks. This study showcases that while ML research, development, and deployment involve reuse of data and models, these elements are not, in the words of Thylstrup et al. (2022, p. 1), 'traceable, stable, and contained' but part of reconfigured assemblages.

From the choice of the initial model or architecture, model training is an iterative process, which involves continuous tinkering with the architecture, choosing techniques (e.g., for clustering), and validating and interpreting results (Bechmann & Bowker, 2019; Lee, 2025). Dataset and model development can also go hand in hand, e.g., more data can be generated and/or labelled to capture some features or cases the current model is struggling with (Wirth et al., 2025).

Model development can be finished when a 'working' or 'good enough' result is achieved. In a general sense, 'a "good enough" solution is one that achieves some level of optimisation in the relationship between a given target and the actual output of a model' (Amoore, 2020, p. 67). And in practice, determining whether a machine learning solution is good enough and choosing the metrics and thresholds of success involves situated and pragmatic criteria and ongoing negotiations (Schjøtt & Blanke, 2026; Wirth et al., 2025). This varies by types of projects (Wirth et al., 2025); for instance, for academic researchers in ML as well as many of the connected fields, the dominant focus of the work and contribution is often developing models that would outperform state-of-the-art (SOTA) techniques on the standardised 'benchmark' datasets used for comparing performance between models. For a lot of industrial development, however, novelty in the algorithm architecture would not matter, but creating a reliable solution will (Thomas, 2019). In certain contexts, there might be regulatory criteria and requirements posed by, for example, medical device certification (can be relevant for medical software that involves machine learning models) (Niemiec, 2022).

The ‘good enough’ model would be written up and shared as a research output, or used (deployed), potentially integrated into larger software systems (Wirth et al., 2025). The resulting products, however, are rarely completely ‘finished’. Even academic researchers’ work might require maintenance, for example, updating servers and code for the published solution to work. Even more, maintenance is characteristic of applied machine learning. This includes dealing with the changing nature of the world the model should represent (Lee, 2025; Sambasivan et al., 2021). Over time, concepts and phenomena, as well as physical environments where data-producing equipment like sensors and cameras are situated, as well as featured concepts and phenomena, can drift; accounting for drifts would require further tinkering. This maintenance would also include performance monitoring and dealing with failure cases (Wirth et al., 2025).

Notably, these stages and practices vary depending on the context of development: types of projects, models, and data, involved actors, and the position in the supply chain (Widder & Nafus, 2023). For example, Thomas (2019) noted two types of computer vision work: algorithm-invention in prestigious labs (academic and industry) and more engineering work that uses the algorithms produced by the former to build working solutions, though the distinction has blurred as both have started to draw on (and/or contribute to) the pool of open-source tools. However, these supply chains, and the use of common workflows and tools, connect technologists across the distributed community of practice.

Another common feature of ML work is collaboration. In addition to ML practitioners working in teams, they often interact with a variety of other actors: managers and product owners, data stewards, compliance experts, data annotators, as well as researchers and/or practitioners specialised in a certain topic of interest, often referred to as ‘domain experts’ or ‘domain specialists’ (Wirth et al., 2025; Zhang et al., 2020). Depending on the organisational setup and working relationship, these actors can interact with ML practitioners at different stages of the latter’s work mentioned above (Diaz & Smith, 2024; Zhang et al., 2020). Social studies have explored how these dynamics shape the outcomes of ML systems, including risks and biases (Miceli et al., 2022; Sambasivan et al., 2021; Zając et al., 2023). As managers and product owners (unless that role is played by domain specialists) will appear to a very limited extent in the empirical studies of the dissertation, I will only point to some of the studies that provide insight on these actors and their interactions with technologists – Zhang et al. (2020) and Wirth et al. (2025) – but provide more information on the other three types of actors that will make appearances in the following chapters.

Organisations where ML practitioners work, especially in industry, need to ensure their compliance with binding legislative acts, international standard organisations, or industry standards (Zając et al., 2023, p. 355). Hence, ML practitioners interact with various compliance and quality assurance specialists within organisations, as well as data protection officers and data stewards, and privacy and security specialists. Data stewards take care of the data, ensuring processing compliance and managing access to certain datasets. They can also prepare and aggregate the data for further (re)use (Thakkar et al., 2022).

As mentioned above, data annotation can be done by a variety of actors, from ML practitioners to domain specialists, crowd workers, or employees of specialised annotation teams, internal to the organisation or subcontracted. In addition to labelling training data, annotators can also support model evaluation and even maintenance (Wirth et al., 2025). Previous studies have problematised simplified views of annotation work often held by others, noting that its intricacy is obscured or realised in the process (Bennett et al., 2025). The degree to which annotators can develop skills and how much that knowledge can be incorporated into the developed dataset depend on the organisational relationship and power dynamics; for instance, whether there is a communication line from annotators to developers, and/or the degree of mutual understanding (Miceli et al., 2022; Wirth et al., 2025; Zając et al., 2023). Among actors involved in producing knowledge objects in machine learning, annotators are often in the most precarious position, as this work is often outsourced from the Global North to the Global South, poorly compensated, and performed under stark inequality between annotators and their clients (Gray & Suri, 2019; Shestakofsky, 2024).

As mentioned, specialists in a certain knowledge area for which a machine learning solution is developed are colloquially referred to as ‘domain experts’. This notion refers to the separation between general methods and particular domains of knowledge and activity; this separation positions computational methods as potential solutions to questions and problems in different ‘domains’ of knowledge and activity, with the idea to infuse abstract methods with the domain context (Ribes et al., 2019; Slota, Hoffman, et al., 2020). Domain experts can contribute to problem formulation, be involved in data generation and labelling (e.g., creating label schemas, doing primary data collection, and labelling data), inform feature selection, and participate in evaluation and testing (Diaz & Smith, 2024). Reports on how involved ‘domain experts’ are throughout the different stages of ML projects have varied (Diaz & Smith, 2024; Zhang et al., 2020), probably reflecting the variation among projects and working relationships. However, while this workflow might be seen as a boundary object enabling collaboration between two different epistemic

communities (Star & Griesemer, 1989), critical scholars have problematised the epistemic hierarchy arising in these computational collaborations, as 'domain knowledge' is either deprioritised or abstracted from the experts (Diaz & Smith, 2024; Sambasivan & Veeraraghavan, 2022). Moreover, they note underuse or misuse of domain knowledge and potential gaps in mutual understanding between ML practitioners and domain specialists as posing risks for future systems' performance (Sambasivan et al., 2021; Zając et al., 2023).

This and previous sections have given a glimpse into the machine learning and artificial intelligence projects across time, and the knowledge practices, infrastructures, and relations that comprise machine learning as we know it today. This section has also pointed to the diversity of ML projects and practice across contexts. Before finishing this contextual chapter, I address how I deal with this diversity conceptually and terminologically.

2.5 A note on terminology

So far, I have relatively freely mixed the terms (and the studies about) machine learning and artificial intelligence when describing ML work. This mix is indicative of terminological turns in the literature I am drawing from, which rarely featured distinctions between them or their rationales for selecting the terms they use. While I initially found this frustrating, I realised that this flux reflects the fuzzy and changeable meanings of these terms in practice, as indicated by previous research as well as some of my interviewees (see Chapter 5).

For example, reviewing academic literature on data science as well as discussions within the professional online communities in places like Kaggle, Muller et al. (2019) noted a diversity of roles and disagreement over their contents. Finding that 'many people are doing the work of data science, across multiple job categories and titles, and that this diversity is likely to increase over time', the researchers change the wording for their studied group from 'data scientists' to 'data science workers' (p.2). Avnoon (2024) characterises this heterogeneity as a feature of 'alternative professionalisation' of data science that is not based on delineating a turf of tasks and gatekeeping access to it through credentialism but is seemingly open to everyone while inheriting the normativities of computer science and engineering. Avnoon notes that this openness is far from full and still reflects the privileging of 'technical' backgrounds like computer science and engineering. The interviewees from the industry told me that what the 'data science' role includes varies too much by company to give a full indication of the tasks and required companies. Other studies noted that the same specialists might choose to describe themselves as working on machine learning but use 'artificial intelligence' when talking to stakeholders or media (Crawford, 2021; Joyce et al., 2021).

In this context, I settled on machine learning and machine learning practitioners. I borrow this descriptor from Bennett et al. (2025), although they apply it to engineers and data annotators alike, to acknowledge the diverse knowledge and work that goes into AI developments and problematise the existing hierarchies of recognition in tech. The 'data scientists' vs. 'ML/AI researchers' reflected a divide across positions in the supply chain (though again, not fully – for example, one of the interviewees had a data science title while doing a research and development job in industry). The distinction between research and engineering also felt too blurry, given that AI research has historically been torn between science and engineering (Forsythe, 2001) and the bridging focus on building technologies rather than only providing disembodied (written) algorithms (Thomas, 2019). Moreover, focusing on ML/AI researchers did not reflect the links between fields like computer vision, natural language processing, etc., facilitated by the spread of deep learning methods (Berman et al., 2025). Thus, I focus on ML practitioners as the concept that helps me see connections across organisations (which differently define roles), fields, and sectors. In doing so, I acknowledge that while developing and/or applying ML models is what determines ML work as understood in the community and in this dissertation, many others' work is necessary to make machine learning 'doable' both as a research and technological object.

With this background information on machine learning and machine learning work, I now move to the theoretical framework of the dissertation.

Chapter 3
Theory

This chapter outlines the theoretical toolkit of the dissertation. First, I introduce the general landscape of social studies of science and technology across multiple fields as well as the dissertation's general focus on representation and understandings of work. With this foundation, I describe the themes I explore in the empirical research and theoretical frameworks I employ to do that. First, I zoom in on the theories I employed to study representations related to labels of ML work and representations of ethics: Stuart Hall's theories of representation and the encoding-decoding model, and the conceptual metaphor of constellations based on the use of spatial metaphors of representation in STS. Next, I delve into the lens of friction and how I complement it with relational thinking from infrastructure studies. The chapter ends with a reflection on the general approach this eclectic theoretical framework engenders and the role of theory in this dissertation.

3.1 The landscape of social research about science and technology

This dissertation inherits its analytical sensibilities from work in science and technology studies (STS) and connected social and humanistic approaches. Establishing that science and technology are neither neutral nor independent of the cultures and societies in which they are produced has been foundational to social studies of science and technology (Bijker & Law, 1992). In doing so, STS inquiries have moved from the idea of disembodied scientific 'knowledge' to researching the cultures, practices, work, and infrastructures of knowledge production (Edwards et al., 2013; Pickering, 1992). Furthermore, they have recognised knowledge work and expertise existing 'outside the mainstream scientific community' (Collins & Evans, 2007, p. 609) and in globally distributed locations (Knorr Cetina, 2007). So they have extended their analytical focus from local epistemic cultures of isolated labs to distributed knowledge infrastructures: 'networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds' (Edwards, 2010, p. 17).

These developments are also reflected in the studies of computational practice. A wave of laboratory ethnographies has examined the culture and practices of artificial intelligence labs producing 'expert systems' (Collins, 1990; Forsythe, 2001; Suchman & Trigg, 1993). Researchers in fields from STS

to human-computer interaction (HCI) and Computer Supported Cooperative Work (CSCW) have explored computers and various information and communication technologies as elements of knowledge production (e.g., Star & Ruhleder, 1996), and the larger knowledge and information infrastructures have grown with the use of the internet around the world (Bowker et al., 2010). In the 2010s, the descendants of the laboratory studies tradition focused on machine learning development (Jaton, 2017, 2021) and extended this approach beyond scientific research, following the spread of machine learning algorithms and work (Schjøtt & Blanke, 2026). Scholars across disciplines focused their analytic attention on the work algorithms (and those producing them) performed in various spheres of life, from finance (Hansen & Borch, 2021) to music recommendation (Seaver, 2022). The exploration of various subjects related to digitalisation and computation has formed several fields named after them: critical data (Iliadis & Russo, 2016) and dataset studies (Thylstrup, 2022), critical algorithm studies (Seaver, 2017), and, most recently, critical AI studies (Raley & Rhee, 2023). Here, I will not make a strong distinction between STS and these fields, or connected work in HCI and CSCW. Instead, I will treat them as a polylogue of approaches to what Lee and Ribes (2025) call 'computational relationalities'. Moreover, while I draw from these topically relevant literatures, I do not limit myself theoretically to frameworks and concepts developed within this context.

3.2 Meanings and representations of ML work

My use of the concept of representation indicates a constructivist approach to meaning as actively produced through interpretive practices, rather than phenomena and their meanings existing independently of each other, or the latter objectively reflecting the former one-to-one (Hall, 1997). In science and technology studies, an influential conceptualisation of representation was developed by Thomas F. Gieryn (1983, 1999). Gieryn denaturalised the notion of science as necessarily possessing certain traits or including certain approaches, claims, and concepts while excluding others. He argued that what we know as science is (re)constructed through credibility debates among scientists and between scientists and other actors. What is invoked in these contestations is not 'real science' (in its multiplicity) but various representations of science, which are drawn, like maps, by what he termed boundary-work: 'the discursive attribution of selected qualities to scientists, scientific methods, and scientific claims for the purpose of drawing a rhetorical boundary between science and some less authoritative residual non-science' (Gieryn, 1999, pp. 4-5). While Gieryn focused his analysis on the credibility-contest discourses that employ and produce representations of science through boundary-drawing, he also argued that the resulting representations become part of how other

people understand science and where they put their epistemic trust when consulting 'scientific' advice and authority. As such, they have real consequences not only for the participants of the credibility contests but also for those drawing on these representations.

Social studies of science and technology have also applied a constructivist approach to representation to the question of how actors in technoscience see their and others' work. In that line of thinking, Lucy Suchman proposes to approach representations of work 'not as proxies for some independently existent organizational processes but as part of the fabric of meanings within and out of which all working practices – our own and others' – are made' (1995, p. 58). The inclusion of 'our own' emphasises that while an insider perspective is unique, it is also partial (Haraway, 1988; Suchman, 2002) and socially constructed. For instance, in her ethnographic study of AI labs, Forsythe (2001) noticed that when she asked engineers to describe their work, their portrayals centred on problem-solving done by building systems and writing code. Observing that these descriptions do not cover many of the engineers' activities, e.g., conducting interviews with experts to elicit their decision-making schemes, Forsythe argued that these understandings are shaped by the narrow definition of the 'work' category that excludes less technical tasks. In addition to such cultural notions, particular positions in networked relations and interests contribute to specific and partial representations of work (Suchman, 1995). These representations inform practice and have real consequences for the developed technology. Forsythe (2001), for instance, connected AI engineers' narrow understanding of 'work' and 'knowledge' to their approach to knowledge elicitation interviews and modelling experts' decision-making, and the resulting brittleness of 'expert systems'.

In analysing how ML practitioners understand their work, I also draw on the narrative approach developed and used by Ulrike Felt and her collaborators (Felt, 2017; Felt & Fochler, 2012; Sigl et al., 2020). Felt and colleagues foreground narratives with which scientists think and talk about their work as 'windows to the social, epistemic, political, symbolic and many other facets of life in research' (Felt, 2017, p. 55). Felt terms these narratives 'epistemic living spaces' and posits that they inform practice by giving definition to the situations around scientists and shaping their sense of agency as well as the options for action that seem available to them (Felt, 2017). The researchers then use this framework to examine different aspects of life in research through narratives. For instance, Felt and Fochler (2012) consider how, through the proliferation of communication activities, promotional representations of science produced in media communication – 'short and exciting tales about one's work and its outcomes' (p. 142) – become part of how researchers think and talk about their work. Repeated in communications such as grant applications

and recruitment of new generations of scientists, these narratives shape the (re)production of science.

Moreover, this framework argues that personal and situated narratives are not individualistic but reflect both broadly circulating narratives and structural conditions. Felt and colleagues make this connection by introducing the concept of the narrative infrastructure (Felt, 2017; Sigl et al., 2020). The narrative infrastructure consists of a set of stable and coherent narrative points that support local meaning-making. For instance, the idea that scientific institutions reward publications and not outreach or engaging with society forms part of the narrative infrastructure that deprioritises societal responsibilities as a part of doing science (Sigl et al., 2020). In that, narrative infrastructures are similar to the constructed boundaries of science: they are shared cultural notions that are drawn on in local meaning-making and practice (Gieryn, 1999).

These approaches, focused on situated and intertwined with practice meanings, provide a foundation for the dissertation's theoretical framework. I approach ML practitioners' understandings of their work as shaped by both practice and the relations it enacts and circulating narratives or emic categories like 'science' (Gieryn, 1999) or 'work' (Forsythe, 2001). These understandings, in turn, contribute to their working practices and sociotechnical systems these practices co-produce.

For each of the three empirical analyses (Chapters 5-7), I use an eclectic set of hybrid theoretical approaches I develop based on the theoretical foundation described above as well as the concepts and theories I discuss in the rest of the chapter. The approaches and their building blocks are illustrated in Table 3.1.

Table 3.1. The theoretical frameworks in empirical analyses

Paper/Chapter	Theoretical framework
1/5	Boundaries (Gieryn, 1983, 1999) + encoding-decoding (Hall, 1973)
2/6	The map metaphor (Gieryn, 1999) + narrative infrastructure (Felt, 2017) → constellations metaphor
3/7	Friction (Edwards, 2010; Tsing, 2005) + relational and infrastructural thinking (Star & Ruhleder, 1996)

In Chapter 5, I employ Stuart Hall's concept of codes (Hall, 1997) and the encoding-decoding model of communication (Hall, 1973) to theorise communication and practitioners' relationship with labels such as 'ML' and 'AI' as signifiers of their work. Into that theory, I integrate the concept of boundaries,

drawing on Gieryn's work (1983, 1999) while also acknowledging the sociological genealogy of the concept (Lamont & Molnár, 2002).

In analysing representations of ethics in Chapter 6, I draw on the lesser-known aspects of Gieryn's (1999) conceptual framework: the geographical metaphor of representation. I also borrow the concept of narrative infrastructures to consider the sources of these representations.

In Chapter 7, my analysis of friction theorises it as relative to ML work (as understood by ML practitioners) and considers the notions and expectations of ML work as crucial elements of working relations in which friction can generate.

Above, I outlined the theoretical foundation of the project, from the broad landscape of social and humanistic approaches to studying science and technology to the theories that formed my focus on meanings of ML work. Now, I will discuss the theoretical frameworks I used in the empirical analyses.

3.3 Labels as representations

From this foundation and through dialectical conversation with data, I developed the theoretical framework to analyse ML practitioners' meaning-making around the labels 'machine learning' and 'artificial intelligence' as descriptors of their work.

Current STS research has problematised the naturalisation of computational categories, including *within* STS research (Lee & Ribes, 2025). Lucy Suchman has specifically called out the 'uncontroversial thingness of AI', a combination of 'interpretive flexibility' (Pinch & Bijker, 1984) and naturalisation that presupposes a concrete and stable, singular referent for this symbol. And studies analysing the history of this concept have noted how its meaning and relation to 'machine learning' changed over time (Cardon et al., 2018; Dobson, 2023; Galanos, 2023). A change in the meaning of technology is not unique to AI or ML: for instance, the concept of a bicycle, fairly agreed upon now, earlier referred to different and competing technologies (Pinch & Bijker, 1984). However, as noted by Galanos (2023), artificial intelligence is a project that so far has not narrowed its interpretive flexibility down to something stable and limited, as posited by the social construction of technology (SCOT) model (Pinch & Bijker, 1984). Moreover, it has been noted that AI changes its meaning whenever it achieves a new goal (McCorduck, 1979, in Galanos & Stewart, 2024, p. 28).

Researchers were active participants in the reconstruction of the machine learning and AI categories (Dobson, 2023; Galanos, 2023). One dynamic of this participation is the competition between the connectivist and symbolist approaches to artificial intelligence, expulsion of the connectivists from the 'AI' category, and their 'revenge' with the success of deep learning in the 2010s

(Cardon et al., 2018; Galanos, 2023). This competition could be put in Gieryn's terms (1999) as a re-drawing of the boundaries of AI. Another dynamic is related to the changing expectations of 'AI', the waves of excitement and disappointment commonly understood as AI summers and winters (Galanos, 2023). When applying for grants or publishing research results, researchers took note of the connotations of the approaches and technologies they worked on, avoiding terminologies associated with technologies and approaches that have disappointed expectations (Dobson, 2023; Galanos, 2023). For example, Dobson characterises computer vision as one of the projects in which researchers redefined their work 'when it had become no longer fashionable or politically viable to be associated with artificial intelligence' (p. 62). These accounts also highlight that along the 'summers and winters', the heterarchies of symbolic value between 'AI', 'machine learning', and other categories (e.g., computer vision) have changed. While the general practice of redefining the contents of categories in pursuit of goals and interests fits well with boundary-work (Gieryn, 1999), the changes in these heterarchies and the disappearance of some of these categories (like 'AI' during AI winters) – contra science's authority over non-science that persists despite their changing context – are less congruent with this framework.

This tendency to produce favourable representations of one's technoscientific work that ends up reconstructing it was noted by the 'epistemic living spaces' approach in another empirical context (Felt & Fochler, 2012). Felt and Fochler showed how life scientists' repertoires of communicating their work were affected by mediatisation strategies focused on promising societal applications of the research, and public resistance to genetic modifications in agriculture. This combination becomes problematic for plant gene researchers, who, due to connotations with the controversial technology, cannot make promises about applications which are (understood as) necessary for acquiring funding and presenting one's research in the media. This negative association extends to descriptions of one's work, as Felt and Fochler mention that the public opinion on gene modification in agriculture 'renders "genome researcher" one of the least popular lines of work to mention in a pub or any other informal conversation' (p. 147).

To account for these and other potential interpretive dynamics in ML practitioners' considerations for labelling their work, I draw on Stuart Hall's theory of representation (Hall, 1997), into which I incorporate the concept of boundaries, and his encoding-decoding model of communication (Hall, 1973).

3.3.1 Stuart Hall's codes and encoding-decoding

Hall's theory of representation affords the destabilisation move I needed to explore 'AI' and 'machine learning'. Hall (1997) argues that words do not

inherently mean anything in themselves but take on meaning through culturally established interpretive links to experiential phenomena and other concepts. Hence, Hall breaks away from the semiotic distinction between denotations that directly refer to 'things' and connotations with other meanings. Rather, according to him, the 'directness' of denotations reflects a high level of stability and agreement in a certain culture on the link between a sign and a meaning. These links between signs and meanings, and among meanings, which Hall terms 'codes', are the main building block of his theory of representation.

For instance, the different meanings ascribed to a technoscientific term as well as the degree of 'interpretive flexibility' in interpreting an object (Pinch & Bijker, 1984) can both be seen as codes. The broad concept of a code also allows me to integrate boundaries as a type of code – codes that articulate inclusion and exclusion into categories and distinctions between categories (Gieryn, 1999; Lamont & Molnár, 2002). The exclusion of statistics-based approaches from being 'AI research' in the 80s can be seen as such a code (Forsythe, 2001). Moreover, this theory notes that codes are subject to change and that different cultures can draw on different codes; thus, social location can contribute to different meanings ascribed to different words or different connotations between meanings (Hall, 1997).

The encoding-decoding model conceptualises representation as an active interpretative process, undertaken in moments of signifying meaning through language (encoding) and interpreting that discursive message (decoding) (Hall, 1973). To encode meanings, actors draw on codes and relations of codes, which can include considerations for potential understanding of the message (decodings) and reactions. Notably, this model was initially developed for mass communication, from professionalised producers to audiences. For this, Hall pointed to the existence of 'professional codes', specific to this professional community due to their knowledge and positions in society, through which they create cultural products like news reports or television shows. However, he also noted that professional encoding tends to take into account potential audiences' reactions – though this link disappears in the 1980 version of the text (Hall, 2019, p. 47) – and to reify the dominant, and thus naturalised, shared, and stable significations.

Receivers of the message similarly draw on codes to decode its meanings; as multiple relations between signifiers and signifieds, and connotative links, can exist, these might not be the same codes as used in the encoding of the message, resulting in distorted (to differing degrees) communication. As noted by Hall, this degree of proximity between encoded and decoded meanings has to do with shared culture, social positions, and context of the encoding and the decoding (e.g., personal communication versus mediated). He

thus classifies decodings into three categories based on their relation to the encoded meaning. Dominant or preferred readings align closely with encoded meanings; negotiated decodings are congruent with the broader meaning of the encoded message but also depart from it in certain respects; oppositional readings are contrary to the crucial elements of encoded meanings.

This model is used to examine the meanings ML practitioners encode when they use certain labels to describe their work and how they decode the reactions and implications of these labelling practices (see Chapter 5). I have chosen this theory for its fit with my empirical material and the dynamics described in the literature on AI and by Felt and Fochler (2012). This framework allows me to attend to the interpretive flexibility of 'AI' and how actors themselves navigate it, and to connect the situated dynamics of representation with broader cultural dynamics – which can contribute to the changes in these codes.

3.4 Representations of ethics in ML work

The theoretical foundation described in section 2.2 has also informed my approach to studying how ML practitioners construct the relationship between ethics and their work. First, I outline my approach to ethics as an actor's category and my language-close operationalisation of it that informed the interview guide (see Chapter 4). Then, I outline the development of the conceptual metaphor of constellations I employed to study how ethics fits into practitioners' narratives of their work.

3.4.1 From ethics to 'ethics'

Empirical approaches to ethics conceptualise it as 'an ongoingly emergent facet of the everyday experiential world' (Pink, 2022, p. 46). For instance, Louise Amoore conceptualises algorithms as 'ethicopolitical arrangement[s] of values, assumptions, and propositions about the world' (Amoore, 2020, p. 6). This notion emphasises that ethics cannot be abstracted from the practices and relationships in sociotechnical systems and does not differentiate between ethics and epistemology (Pols, 2015). Ethics is enacted in everything actors and technologies do, even if it is done, in the words of Verbeek (2006), 'by other means'.

However, actors themselves do not always see it that way. For example, Pickersgill (2012) showed that neuroscientists' narratives reflect an ambiguous and somewhat fluid relationship between ethics and science. On the one hand, values and concerns he would consider ethical permeate scientists' accounts of their everyday work. On the other hand, neuroscientists continuously redraw boundaries between science and ethics and demarcate areas free

from ethical considerations. Similarly, previous studies note that ethics in computational cultures is discursively absent, but value judgment is constantly present in (accounts of) practice (Barocas & boyd, 2017; Slota et al., 2023). A stream of studies investigated these demarcations: whether actors demarcated ethics as separate from science or technology development or integrated ethics into them, and how they delineated ethical and non-ethical parts of their practice (Frith et al., 2011; Pickersgill, 2012; Wainwright et al., 2006; Wehrens et al., 2023). In Chapter 6, I follow this line of inquiry into how actors delineate the role of ethics in their work, arguing that these are representations of work that are meaningful in their own right. By doing so, I learn more about ethics and *where* it is visible in their practice.

Moreover, I decided to focus on 'ethics' as a signifier of a specific, culturally constructed concept that would be among the concepts involved in ML practitioners' everyday ethics. Notably, previous studies noted that the word 'ethics' itself has connotations with institutional approaches. For example, Pickersgill makes a distinction between ethics and what he has termed 'regimes of normativity'('webs of official and informal institutions, agents and discourses that prescribe what scientists ought to do'), in order to 'move away from an understanding of "ethics" in science as solely defined by and under the purview of formal regulatory bodies and moral philosophers' (2012, p. 580). Similarly, Heimer argues that ethics on the ground is less likely to be identified as such, and the uniform language of 'ethics' is more characteristic of official systems (2013, pp. 373-374). Both studies point the reader's attention to the connotations of the term 'ethics' with professionalised and 'official' ways of doing ethics. In this study, I explore empirically whether practitioners associate 'ethics' with institutionalised procedures and/or other 'ethics' discourses in ML/AI (Greene et al., 2019; Stark, 2023). In doing so, I am prodding the relationship between the two notions of AI ethics: 'the institutionalized academic field that has significantly shaped the contours of contemporary AI Ethics principles, guidelines, institutions, economies and policies' and 'everyday practices oriented toward achieving diverse visions of the moral good' (Helm & Gerlek, 2026, p. 2).

Consequently, I focus on how ML practitioners understand 'ethics' and relate this concept to (parts of) their work. In the interview guide (see Appendix I), I foreground 'ethics' to ask various questions about how practitioners understand this concept and where in their practice they encounter and do things related to it. I thereby perform the constructivist move of shifting the task of defining the phenomenon of interest from analyst to actor (Gieryn, in Calvert, 2006, p. 200). At the same time, I recognise that these understandings would be narrower than the practitioners' full repertoire of 'everyday practices

oriented toward achieving diverse visions of the moral good' (Helm & Gerlek, 2026, p. 2) or 'regimes of normativity' (Pickersgill, 2012).

Having outlined this approach to ethics (or rather, 'ethics'), I move on to develop the theoretical approach to studying the relationship between 'ethics' and representations of ML work.

3.4.2 Developing the constellations metaphor: spatial metaphors and narrative infrastructures

Trying to elicit the relations between ML work and ethics in dialogue with my data, I arrived at spatial thinking. Gieryn (1999) argued for the productivity of the 'map' metaphor to study representations of science:

> Maps do to nongeographical referents what they do to the earth. Boundaries differentiate this thing from that; borders create spaces with occupants homogeneous and generalized in some respect (though they may vary in other ways). Arrangements of spaces define logical relations among sets of things: nested, overlapping, adjacent, separated. Coordinates place things in multidimensional space, making it possible to know the direction and distance between two things. Landmarks and labels call attention to typicalities or aberrations, reduce ambiguities about the precise location of a boundary, highlight differences between spaces of things: they are reality checks, of sorts. Most important, just as maps of earthly patches get drawn to keep travelers from getting lost, so maps of other worlds—culture, for example—are drawn or talked to help us find our way around. (p 7).

Thus, one way to think about the relations between representations of ML work and ethics would be through their boundaries and maps. While several studies mentioned above drew on Gieryn's concept of boundary-work (Frith et al., 2011; Pickersgill, 2012; Wainwright et al., 2006), they did not utilise the metaphor of the map, and I considered that it would help to better understand the areas of ethical versus non-ethical in ML practitioners' practice.

Another spatial metaphor has been employed by Sigl et al. (2020) to elicit the proximity-distance and centrality-periphery relations to compare positions of different obligations in the scientists' overall 'geographies of responsibility'. What they found is that these geographies were centred on scientific excellence and publishing, which scientists saw as being valued by the scientific institutions, while societal responsibilities were ascribed to the periphery and to the purview of other actors. Moreover, they identified the shared narrative points across the narratives that worked as the narrative infrastructure for the particular geographies of responsibility: a linear model of innovation, in which societal use and relevance come temporally after scientific discoveries, and the notions of what is rewarded by the academic system. I have

borrowed this idea of shared narrative points among scientists in the analysis in paper 1, identifying notions that underscore each of the ethics constellations.

Maps of science (Gieryn, 1999) and geographies of responsibility (Sigl et al., 2020) have been inspirations for developing the constellations metaphor. However, they did not fully 'fit' what I saw in the data. First, as I was interested in the representations of work and of practice, my conceptualisation needed to better connect to the notion of representations as situated and shaped by practice (Suchman, 1995). Previous studies on ethics in science and technology have noted how different contexts, places, and even the material form of ethical toolkits elicit different ways of engaging with ethics (Goñi et al., 2024; Shklovski & Némethy, 2023; Smith et al., 15.05. 2024). Moreover, it needed to attune to the granularity of various parts of work (Frith et al., 2011) and narrative points (Sigl et al., 2020). Both previous studies in the tech ethics context (Barocas & boyd, 2017; Slota et al., 2023) and my empirics suggested that this lens needed to consider not only explicit boundary-drawing or prioritisation but also (in)visibility as interpretative mechanisms that form representations (Suchman, 1995). Finally, it needed to incorporate the connection to narrative infrastructures as meanings that support particular representations (Felt, 2017; Sigl et al., 2020) to account for shared culture and structural conditions. These considerations have underscored developing the conceptual metaphor of constellations.

3.4.3 Constellations

The notion of a constellation is based on stars that *appear* to be grouped together. By using the constellations metaphor, I emphasize the work of interpretation that notes certain objects in the night sky and connects them into meaningful shapes that form maps of the night sky. The night sky is full of disparate objects; the constellations metaphor emphasises how interpretation makes meaningful shapes out of these disparate points, connecting some objects and not others. Importantly, these meaningful shapes are influenced by familiar cultural concepts, from mythical creatures to animals (which is what makes them meaningful). Cultural concepts we use, like 'science' or 'ethics', group disparate meanings into constellations, 'things' we see, and coherent narratives. Notably, we connect some of the many disparate objects seen in the night sky and not others – these connections could have been otherwise. Similarly, the meanings and practices one could see as part of ethics could have been associated with other concepts, like 'quality' or 'rigour'. Moreover, once we see something as a constellation, stars around it that were not included become less noticeable, belonging to the map's background rather than the figure in it. Once we know a constellation, we can find it again in the night

sky more easily, while excluded or ungrouped objects are harder to notice. This reflects the tendency that representations make some aspects of reality more salient and others invisible (Bowker & Star, 1999; Suchman, 1995).

By using the constellations metaphor, I make use of maps' affordances to visualise relationships (Gieryn, 1999; Sigl et al., 2020) but instead of focusing on relationships of inclusion-exclusion (as done through boundary-work) or proximity-distance, I focus on connections that link meanings to each other and to available cultural notions. I borrow the notion of narrative infrastructure (Felt, 2017) to incorporate meanings that equip a stargazer to see some shapes and not others, thus underscoring situated and personal interpretations. Moreover, the fact that stars are not equally visible depending on the observer's location in time and space reflects the situatedness of representations (Suchman, 1995). With this conceptual metaphor, I examine practitioners' representations of ethics in their work and find constellations of ethics based on the shared narrative points, noting the meanings and concerns that these constellations exclude, and trace their construction to discourses and practices that comprise their narrative infrastructures (see Chapter 6).

3.5 Friction in relations of ML work and (biomedical) data

The focus on meanings that comprise ML work has also informed my approach to data and data friction (see Chapter 7). While one of the main analytical benefits of the friction lens is its attention to materiality (Edwards, 2010; Pelizza, 2016), scholars have also used it to highlight tensions between epistemic cultures (Edwards et al., 2011; Hartley & Thylstrup, 2025; Thylstrup & Hartley, 2025). Hence, this lens allows me to extend my analytic attention to the more material and infrastructural aspects of machine learning by considering them in relation to the meanings and practices of ML work.

Several existing approaches de-essentialise data by highlighting their embeddedness in the contexts of their production and use: 'data are cultural artifacts created by people, and their dutiful machines, at a time, in a place, and with the instruments at hand for audiences that are conditioned to receive them' (Loukissas, 2019, pp. 1-2). They conceptualise data as materials produced by the abstraction of phenomena (Kitchin, 2014) or as entities (to be) used for future evidence (Borgman, 2015). As Borgman argues, something *becomes* data when it is drawn on for future insight or argument. This relationality to the knowledge production is also at the heart of Drucker's suggestion for the alternative term 'capta', a Latin word meaning 'taken' rather than the 'given' meaning of data, emphasising that they come from somewhere and are obtained through actions and choices rather than passively absorbed

(Drucker, 2011). Data are not just 'cooked' rather than 'raw' but also embedded in the relations between actors and tools that created it, and a lot of work can be done to strip data of that situatedness in order to move and (re)use them (Leonelli, 2014; Wirth et al., 2025). Machine learning then entangles humans, data, algorithms, and other non-human actors (Amoore, 2020; Thylstrup et al., 2022). Loukissas hence suggests turning our attention from data sets to *data settings* (2019, p. 2).

In Chapter 7, I consider ML work and biomedical data as a case of data settings and notions of ML work as a crucial but not the only central part of these settings. I then consider these settings' 'partiality and bumpiness' (Davies & Díaz-Valderrama, 2026, p. 346) with the register of friction. I develop the friction lens by foregrounding relationality by drawing on infrastructural thinking (Star & Ruhleder, 1994, 1996).

3.5.1 From friction to friction in working relationships

The concept of friction in data studies stems from the metaphor of physical processes. As evocatively described by Edwards et al. (2011, p. 669):

> Friction resists and impedes. At every interface between two surfaces, friction consumes energy, produces heat, and wears down moving parts.

Paul Edwards (2010) introduced the concept of data friction to account for the materiality as well as effort and risk involved in collecting, moving, and transforming data within the global knowledge infrastructures of climate science. Data journeys within these infrastructures are comprised of such interfaces as data move from one place to another, one medium to another, one material form to another. Friction can also occur between input and output of any calculating operations, or among collaborating epistemic communities that work with the same data (Edwards et al., 2011).

The other antecedent of the friction lens in data studies is the work of anthropologist Anna Tsing (2005), who studied the spread of global interconnections across local differences. Employing the metaphor of friction, Tsing argued that global projects like capitalism, science, and politics do not smoothly spread to local contexts. However, according to her, friction does not just resist and impede their movement. Rather, it gives them *grip*: this friction 'where the rubber meets the road' enables movement while shaping where the car goes (p. 6). With this metaphor, Tsing points to the generative effects of friction: the 'awkward' interactions between the global and the local produce particular social realities. By incorporating Tsing's notion of friction into the discussion of data and science frictions, data studies sensitised their analyses to the universal(ising) versus local dynamics as well as generative effects of

friction in computational infrastructures (Bonde et al., 2019; Hartley & Thylstrup, 2025; Thylstrup & Hartley, 2025).

Friction is usually what is not expected and what stakeholders of digital infrastructures intend to limit, and data studies have promoted studying friction together with efforts to reduce it (Edwards et al., 2011; Pelizza, 2016; Ruckenstein & Lehtiniemi, 2026). However, Bates (2017) notes that Edwards' concept of data friction originates from studies of scientific knowledge production, a project mostly seen as worthy and important, which is why impeding it is seen as a loss. She draws on the case of data marketisation in surveillance capitalism dynamics to argue that in some contexts, impeding the collection and use of data might be considered a safeguard and promoted.

My overall research focus on meanings of ML work sensitised my attention to such relative and normative aspects of friction. To account for them in the analysis of the frictions ML practitioners experience when working with biomedical data (Chapter 7), I draw on the conceptual approach developed by Susan Leigh Star and Karen Ruhleder in their analysis of challenges that arose in the adoption of custom-made software in the intended user community of biology researchers (Star & Ruhleder, 1994, 1996). My conceptualisation of friction borrows their relational thinking: Star and Ruhleder posit that infrastructure is 'made' by its relation to the work done in a particular context, meaning both its 'background' rather than focal role and the supporting function it is meant to perform. Similarly, I suggest that something is friction in relation to the work in question, and what is friction in one context would not be in another.

Moreover, Star and Ruhleder (1996) theorise that issues preventing the software from becoming an infrastructure stem from tensions and tangles between the various contexts involved in the seemingly simple activities of learning about the software and how to use it. They do so by drawing on information scholar Gregory Bateson's distinction between the message and the different levels of context that frame it:

> There is a first and second order difference in learning something and learning about learning something; and between the second and third are even more abstract differences between learning to learn, and learning about theories of learning and paradigms of education. (Star & Ruhleder, 1996, p. 117)

As Star and Ruhleder note, these second- (and further-order) meta-contexts are often what is tacit for the involved actors and is researched by social studies of science; they make up epistemic cultures and are the reason knowledge cannot always be abstracted from a community of practice. Similar categories are invoked by Pinch et al. (1996), who examine difficulty as a crucial second-order characteristic of skill. Such contextual knowledge is important for

learning as, for instance, one might give up on learning the skill too early or harm oneself or others if one attempts to perform the skill before being ready.

Star and Ruhleder (1996) demonstrate how discontinuities and tensions in the mixing and matching between the contexts involved in the software design and adoption project produce problems that cannot be solved by resource allocation or information transfer. For example, on the first level, the biology software works on a Unix operating system as this was the developers' preference. The issue with that, on the first level, is that most biologists work with Mac. However, simply getting computers that would support the software and teaching biologists to use it does not solve the issue because the discontinuity involves broader-context issues: it is more convenient to have the same operating system as other researchers, and the biology community in question positioned itself as a Mac user community.

Notably, the analysis I undertake does not fully commit to the Bateson model used by Star and Ruhleder (Star & Ruhleder, 1994, 1996). However, I take inspiration from their notion of context discrepancies arising in interaction that make up the supposed infrastructure. This notion is simultaneously general and yet conducive to granular analysis. This allows me to conceptualise friction as occurring in the particular and holistic working relationships assembled when ML practitioners work with biomedical data or 'data settings' (Loukissas, 2019). Moreover, I draw on the concept of more meta-contextual, second-order features of knowledge (Pinch & Bijker, 1984; Star & Ruhleder, 1996) to conceptualise more tacit notions and assumptions within ML work as elements of these data settings, interactions, and uneasy fits between which friction can generate. This conceptualisation enables me to examine the various particular interactions where friction occurs in the ML practitioners' working relationships with biomedical data.

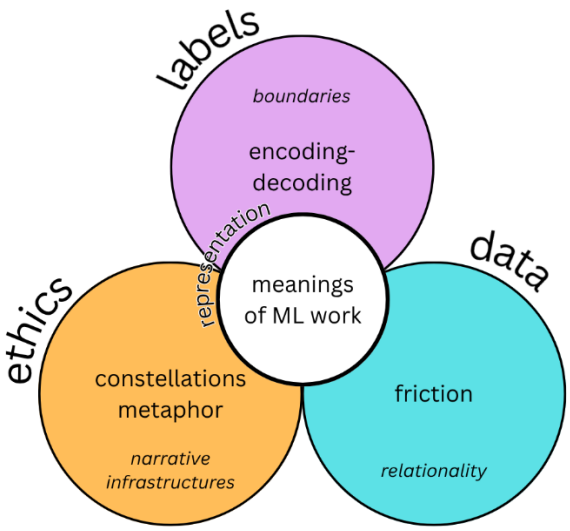
3.6 Summary and the role of theory in this dissertation

This chapter has outlined the eclectic theoretical framework I used in this dissertation. The STS sensibilities and the focus on meanings of machine learning work 'provide[d] the roughest sketch to travel' for the empirical inquiry (Haraway, 1992, p. 295). The interest in the relationship between ethics as a cultural category and meanings of ML work informed the design of the interview study (foregrounding 'ethics', see Chapter 4 and Appendix N). However, the study did not aim to test or commit to using specific theories. As data were generated, coded, and analysed, I put them in dialogue with various theoretical approaches to consider their fit. I did not commit to understanding all themes and patterns in the data with one theory, supposing that a reflexive combination of partial perspectives would produce better knowledge (Haraway, 1988). In this line of thinking, I also allowed myself to develop

hybrid approaches, such as incorporating boundaries into Stuart Hall's notions of codes, combining spatial metaphors with the concept of the narrative infrastructure, and foregrounding relational and infrastructural thinking within the friction lens. Chapter 4 further describes my abductive analytical approach. Overall, my approach to theory can be categorised as 'theoretically interpretive' (Antoft & Salomonsen, 2007), as it is focused on producing new empirical insights and uses theories as lenses to better interpret empirical findings.

This approach resulted in the framework outlined in this chapter, which is grounded in the meanings of ML work, in relation to which it positions and investigates three phenomena encountered in this work: labels, ethics, and (biomedical) data. Figure 3.1 illustrates the links between ML work, these phenomena, and the conceptual tools I use to study their relationships.

Figure 3.1. Topics and theoretical approaches of the dissertation



The figure highlights that while the approaches differ, they all foreground relations to meanings of ML work as key elements of the explanation. For labels and ethics, these relations are those of representation (codes and constellations), which a constructivist approach to meaning denaturalises and sees as produced by active interpretive work, in which actors draw on meanings, concepts, and practices they have experience with. In using Hall's concept of encoding-decoding and the constellation metaphor as theoretical approaches to representation, I also consider how practitioners' representations connect to other meanings and practices. These include discourses and practices that comprise the narrative infrastructures of ethics constellations and the nascently dominant understandings of 'AI'. With data, relations of representation are replaced by friction, a concept that focuses on the interaction between two

surfaces. I consider meanings and assumptions of ML work as elements of the working relationship where they can generate friction with its other elements in the overall relationship. By doing so, I emphasise the meaning-making aspect of friction while considering it part of the overall relationship that also includes materiality and practice. Altogether, these theoretical approaches foreground ML work as embedded in relations with various meanings and practices. Before returning to these discussions in the analysis, I first elaborate on the research design and methods of the dissertation.

Chapter 4
Methodology

This chapter presents the strategies of inquiry pursued in the dissertation. It starts by introducing and situating the overall research design and approach to knowledge production before discussing the data and methods used: the expert interview pre-study that helped me open up the field, the main interview study that formed the empirical basis for the three papers, and the secondary data used in paper 2. Throughout the chapter, I reflect on the methodological considerations, decisions, and contingencies that shaped the research findings.

4.1 Overall research design

To produce knowledge about how machine learning (ML) practitioners understand their work, I adopted a qualitative, interpretative research design centred on semi-structured interviews. Qualitative methods are best suited to explore the world from actors' perspectives, i.e., the meanings they ascribe to the social world and situations around them, and the processes of meaning-making (Yanow & Schwartz-Shea, 2006). Qualitative methods thus privilege attention to local contexts and the complexity of lived realities, and their exploratory nature allows for changes in research goals and actions based on empirical discovery and for (re)conceptualisation of the phenomena under study as part of the analysis rather than a priori determination (Soss, 2021).

4.2 Ontological-epistemological grounding

This project's approach to studying the social reality of machine learning work is ontologically constructivist and epistemologically interpretivist. Constructivism argues that there is no objective real world that we can observe without perceiving it through the socially shaped tools we use to make meaning, and that we necessarily shape it by doing so. To illustrate this, Dwora Yanow uses the example of colour perception: as previous studies have shown, the repertoire of colour terms available to people affects how they see (or do not see) colour, indicating that sensory stimuli are processed through socially constructed categories (Yanow, 2006, p. 10). Thus, the interpretivist epistemology focuses on meanings and their relations as key elements that form social reality, explaining it by eliciting 'the beliefs or meanings that make actions and practices possible' (Bevir, 2006, p. 283).

Social construction through meaning-making includes researchers, as they are not extrinsic to the social reality they study. Interpretative scholars produce knowledge through their engagement with the studied actors and interpretation of their narratives or actions, an interpretation that also necessarily draws on the meanings and categories (among them social-science ones) available to them. This entails reflexivity about the researcher's role in the produced knowledge. For example, a researcher 'not only collects data per a prior expectation but in fact produces the data, in the field, as a function of decisions made in the moment about what to probe or not probe', so 'their words and actions are inextricably part of the data' (Small & Calarco, 2022, p. 147).

This acknowledgement of the social construction of reality does not presuppose complete relativism that completely renounces the production of scientific knowledge beyond an individual perspective. Rather, this inquiry strives to practice a kind of partial objectivity embraced in feminist STS and anchored in Donna Haraway's concept of situated knowledges (Haraway, 1988). Using the metaphor of vision to critique scientific objectivity, Haraway argued against the claim to absolute objectivity and neutrality of scientific viewpoint from 'nowhere and everywhere'. Instead of this impossible disembodied knowledge, she proposed aiming for objectivity as a 'view from somewhere'. This approach notes the active work and particularities of ways of seeing and acknowledges that the knowledge one generates is shaped by one's embodied positioning and perspective. Moreover, this approach is more than acknowledging one's positionality or claiming perspective; it implores practicing critical positioning, reflexivity, and responsibility for how and what the researcher sees. To borrow from another feminist scholar of technology, Diana Forsythe, acknowledging that one has a perspective as an analyst means aspiring to awareness rather than neutrality (2001, p. 70).

This means that in the dissertation project, I strove to understand ML practitioners' meaning-making while acknowledging that my engagement with it is shaped by my social and epistemic grounding and the interactions through which these meanings are elicited and trying to reflexively account for these positionings without erasing them. Crucially, this entails attention to the epistemics and politics of studying one knowledge culture (machine learning) from the viewpoint of another (social science), and awareness of the relationship between technologists and their 'qualitative critics' (Moats & Seaver, 2019, p. 8). Stylistically, this also means adopting an active voice that reminds the reader of my partial perspective and active role in knowledge (co-)production.

The next sections discuss the data and methods I used in more detail as well as the challenges and limitations of this study.

4.3 Expert interview pre-study

In June-August 2023, I conducted six expert interviews as a pre-study. As noted by Bogner and Menz, expert interviews can be used, among other purposes, to familiarise oneself with a new field and gather contextual information to inform the development of the main study's research design and instrument (Bogner and Menz in Flick, 2009, p. 166). With this aim, I conducted expert interviews in addition to desk research and literature review, which commonly precede research design and data collection. The insights generated through expert interviews informed the research questions and methodological choices for the main interview study.

The sampling focused on engaging with expertise in topics that constituted pieces of my research puzzle: ethical issues in data-driven computational research, governance structures, and the machine learning and big data fields. The target number of interviews was five to seven. Recruitment was conducted by identifying potential experts through publications and research organisations' websites and inviting them to participate via email or through personal recommendations from my colleagues. Six contacted experts rejected the invite or stopped replying; in these cases, I looked for experts with similar profiles.

The interviewees gave written consent to participate in the study and to the recording of our conversation. While it is common to report experts' names for transparency, I opted for pseudonymisation, mainly due to time constraints in determining the data processing procedure under the General Data Protection Regulation (GDPR). In the timeline of the study, I did not figure out the procedure that would allow disclosure of participants' identities (with their consent), so I opted for a known approach of processing in the public interest, under Article 6(1)(e) of the General Data Protection Regulation and section 10(1) of the Danish Data Protection Act, which would not allow disclosing experts' names as this would be publication of personal data. This compromise was deemed suitable as the aim of the pre-study was to gain familiarity with the context rather than to generate data to be used for argumentation. This early experience with data processing regulations helped me be aware of the different options and practices for the main study.

The interviewees represented the following areas of expertise:

1. big data research ethics and regulation
2. big data research ethics
3. machine learning (in engineering)
4. data protection in research (in Denmark)
5. data management and mining, machine learning (in computer science)
6. (Danish) data governance and policy

Three interviews took place in person, three online. For each interview, I prepared questions specific to the interviewee and their area of knowledge. I recorded the interviews with a voice recorder, transcribed them using automated transcription software (NVivo Transcription) to generate the initial transcripts, and then manually corrected them against the recordings. I read through the transcripts multiple times and made notes for further thinking and questions to answer for my research design.

As I did not have a background in ethics and governance of science and technology, expert interviews helped me familiarise myself with the state of the art in these areas and figure out how to approach ethics, a notion with multiple ontologies and disciplinary approaches (Hoeyer, 2006). The interviews with the 'technical' experts helped me navigate the landscape of technologies and terms (big data, data science, machine learning, etc.). However, while this pre-study aimed to help me develop a strategy for the 'terminological anxiety' that Seaver (2017) ascribes to the disciplinary transgression that a social-science or humanistic inquiry into technical practice is, this inquiry was not about getting the 'correct' definitions or boundaries but to gain more insight into existing communities of practice and their social realities, including the distinctions between them their perspective (Dourish, 2016).

4.4 Interview main study

Semi-structured interviewing has been the primary method I have employed to generate data for the dissertation, and the 44 interviews I conducted with ML practitioners in February-August 2024 form the empirical basis of papers 1, 2, and 3 (chapters 5-7). Like all methods, interviewing is a particular entry way into phenomena of interest, and a rich body of work in sociology and anthropology has debated its merits and limitations (Jerolmack & Khan, 2014; Lamont & Swidler, 2014; Soss, 2006; Tavory, 2020). Most importantly, semi-structured interviews suited my interest in the meanings ML practitioners ascribe to their work, as they 'elicit explicitly how people understand themselves and their circumstances' (Small & Calarco, 2022, p. 16). As noted by Small and Calarco, interviews produce information on what cannot be observed directly: actors' perceptions, the meanings and significance they assign to these perceptions, and what people personally believe their motivations are (pp. 27-28). Interviews also provide space to examine how different meanings relate to each other (Soss, 2006).

4.4.1 Sampling approach and recruitment

The goal of the sampling strategy was to capture machine learning work as a community of practice (Bowker & Star, 1999) rather than to focus on a singular

title, e.g., data scientist (Avnoon, 2024) or discipline. The inclusion criterion thus was working with machine learning models, while occupational titles, disciplinary affiliations, and data and model types were left open, with a maximum variation sampling approach aimed to represent the range of meanings (Weiss, 1994).

Thus, the aim was to recruit both within and outside academia. However, as one of the research interests included topics of data governance and institutional ethics, there was a preference for practitioners working on topics and data somewhat related to humans, with the notion that they would be more likely to have encountered these institutional frameworks.

To achieve greater diversity in the participants' working contexts (especially academic ones), recruitment focused on ML practitioners in Denmark, the Netherlands, and the United Kingdom. This methodological choice was not due to interest in country comparison but in feasibly accounting for institutional contexts, such as data protection frameworks. In addition to Denmark, the country where I am located and where public funding made this research possible, the Netherlands and the UK were chosen for the feasibility of data collection. My personal experience living and working in Denmark and the Netherlands indicated a high level of English proficiency among residents; moreover, English is common as a workplace language in academia, and many technology teams are highly international. Thus, potential interviewees in these locations were more likely to be comfortable with being interviewed in English. In addition, I had networks that could help with recruitment in these countries.

While national contexts have not been a key interest for the study, the participants' locations are an important part of the situatedness of the results, from the demographics of research participants to the organisations in which they work, the data they work with, and more. Perhaps most explicitly, this situatedness manifests in the role of European and national data protection regulations in practitioners' notions of 'ethics' (paper 2, chapter 6) and friction in accessing health data (paper 3, chapter 7). Procedural ethics approaches also vary broadly across national contexts. For example, Metcalf et al. (2019) showed how institutionalisation of ethics in the US tech industry is shaped by Silicon Valley values, while Avnoon et al. (2024) noted that such formal approaches to ethics are not common in Israeli tech companies. As noted by van Dijk et al. (2021), 'ethics' has been salient in European research and technology governance, including the promotion of ethics as a selling point for European AI in comparison to the US and China. The supposition that 'ethics' discourses would be common in European contexts contributed to the decision to focus on how ML practitioners understand and relate this concept to their work rather than conceptualise ethics as, e.g., moral logics (Avnoon et al.,

2024), discussed as part of the conceptualisation of ethics in Chapter 3 (see also section 4.4.3 below and Appendix I regarding interview guide).

Recruiting approaches included:

1. Searching university websites and LinkedIn (for industry practitioners) for relevant profiles, followed by cold emails/LinkedIn messages inviting them to participate in the study.
2. Attending two large international machine learning conferences and searching for potential interviewees among the participants, followed by an invitation to participate in the study via email or the app that both conferences used to facilitate networking among attendees. Most interviewees were recruited through these two 'cold' approaches.
3. Utilising my social networks: asking people around me if they knew someone working with machine learning and, if so, to tell them about my project and whether they would be comfortable participating or hearing more about it. Twelve interviewees were recruited through my social networks.
4. Snowballing: asking the interviewees to recommend someone they knew within the field. Only two interviewees were recruited by snowballing.

Using the 'cold' approaches as the main recruitment method proved time- and labour-intensive, as many of my emails and messages remained unanswered. Using conferences as a site of data collection thus helped, since, as some of the interviewees mentioned, it was easier to find time for interviews while they were away from their labs and daily lives. Machine learning conferences also encourage networking (including through the social networking features of the conference apps), so contacting attendees for a chat fit the broader social context, even if the type of conversation differed. Moreover, the interviews collected in connection with conferences showed how this was not only a recruitment option and place of interview but a specific 'window' shaping the dialogue: participation in the conference provided some shared experience and sense of connection, and I could use the conference to ask about research they liked or disliked, dynamics between research communities, ethical statements, etc. It is important to note that I did not conduct a proper observation of the conferences (as I spent most of the time arranging or conducting interviews). Lastly, using my own and the interviewees' networks helped pragmatically with recruitment, as people might more easily agree to a request coming through an acquaintance, as well as with contacting practitioners I would not find otherwise, as 'cold' approaches favour visible profiles (e.g., on LinkedIn or maintained personal webpages).

4.4.2 Sample

As mentioned, this recruitment approach resulted in 44 interviews (see Table 4.1 for an overview) with a diverse group of ML practitioners. I reflect on the limitations related to this sample below. 43 interviewees worked directly with machine learning models, and one was a statistician working on a collaborative project with two ML practitioners I interviewed. This interview was quite helpful in inspecting the boundaries of the communities and divisions of labour. This was also the only case in which I knowingly interviewed people who worked on the same project (though they also talked about work experiences outside the shared project).

Most of the interviewees worked in research (academic and/or industry), analytics, or research and development, and several interviewees worked on actively deployed systems. They also worked with different methods and types of models, from 'classic' machine learning like support vector machines (SVMs) and decision trees, to neural networks and deep learning, including generative models. The data types and modalities (e.g., time series, video) they worked with were also quite diverse. The interviewees varied in terms of professional identities and affiliations with fields and disciplines. Some positioned themselves within the machine learning community; others self-affiliated with other fields (e.g., natural language processing) while working with and/or on machine learning methods.

Table 4.1. Description of interview participants

Total	44
Gender	Men (28), women (16)
Country of employment	Denmark (18), the Netherlands (17), the United Kingdom (9)
Academic employment	PhD student, incl. industrial (12), postdoc (4), assistant professor (8), associate professor (2), professor (2)
Non-academic employment	Data scientist (6), research and development (8), business analyst (1), software engineer (1)
Places of non-academic employment	Medium or large company (7), small company (7), startup (1), independent foundation (1)
Current specialisation	Bioinformatics, bioengineering, biostatistics, biosignals (9), natural language processing (8), computer vision, including medical imaging (5), speech recognition (3), emotion recognition (2), privacy (3), fairness (3), reinforcement learning (1), causality (2), other (8)

The overview in Table 4.1 is necessarily simplified. Most crucially, it limits each practitioner to a singular category (e.g., in specialisation), although many worked on multiple projects that refer to different specialisations at the same

time; several participants had multiple roles (e.g., academic and industry). Thus, I share the considerations for this visualisation.

Specialisation categories are also made up to provide an indication of the types of data used. Overall, I aimed to assign one person to a singular category while avoiding categories with only one case when possible. Moreover, there were principles for how interviewees who could be assigned to multiple categories were placed in one category. Practitioners with an academic affiliation are categorised through them rather than their industrial roles. In choosing specialisation categories, I went with how they described current projects; if they had multiple projects, I focused on the ones we discussed most extensively. When there was a choice between more 'application' and more 'methods' categories, I went with 'application', as they are less granular (thus avoiding categories of one). The dreaded 'other' category (Bowker & Star, 1999) in specialisation includes practitioners whose work, based on their narrative, did not include a strong affiliation with a topic, is not uncommon with data scientists (Avnoon, 2021), or varied according to projects and clients. It also includes practitioners whose work I struggled to place in a category that would not be a category of one or describe without revealing too much of their identity.

For reasons of feasibility and access, I managed to conduct more interviews with Denmark- and the Netherlands-based practitioners than with UK-based practitioners, even though the latter is the largest country and AI cluster (Zhang et al., 2022). During data collection, I did not notice significant references to differences based on national contexts in the data, which contributed to the decision to finish data collection despite this imbalance. However, this does contribute to my treating this sample as a whole rather than as three country cases. Moreover, while the interviewees' current work took place in the three selected countries, many of them have previously studied and worked in other countries, representing the international and mobile workforce in academia and tech. The data thus also reflect the work interviewees did in other countries.

While I wanted to include diverse voices, the final sample also reflects the demographic tendencies in data science and AI, where people of colour, women, and queer persons are underrepresented (Zhang et al., 2022). Here, it is important to acknowledge that I did not inquire specifically about my interviewees' demographic characteristics and identities, such as gender, ethnicity, citizenship, or age, as it seemed unnecessarily invasive considering my research interest in constructions of work and could potentially alter the atmosphere of the interview. Some interviewees brought up their identities on their own. However, this choice means treating these categories and identities in an essentialist way, e.g., inferring from information I found online or self-

presentations and treating gender as binary. Though I would honour participants' identified pronouns if they mentioned them, it would have been better to ask about pronouns for more careful reporting. I contemplated using the singular 'they', suggested for referring to an author whose gender is unknown to the writer (American Psychological Association, 2020). However, this made it more difficult to distinguish between the interviewees as a whole versus individual interviewees in writing, and introduced concerns about potentially erasing women in a field where they are already underrepresented (Young et al., 2023). Thus, I settled on using the inferred gendered category and pronouns, recognising this limitation and reminding myself to find better solutions in the future. I left information about race and ethnicity out of the analysis for reasons of protecting anonymity and the contingencies of what sociologist Miri Song refers to as 'visible minority status' (Song, 2020).

The makeup of the final sample, for example, the lower representation of industry practitioners working on active ML systems compared to researchers, might also be a limitation of the recruitment methods and the nature of participation. As noted by Seaver (2017), there are more and less accessible 'windows' into studies of algorithms, and focusing on smaller companies or more available contexts can still produce important insights. Academics were easier to find, and they might have felt more obliged to help a PhD student. Perhaps most important is that the outcomes of their work are supposed to be openly published, so they have fewer concerns about proprietary information and non-disclosure agreements than their industry colleagues. The same probably applies to the underrepresentation of Big Tech participants: despite active recruitment attempts, only one interviewee worked at a Big Tech company, and two had previous experience working or interning there. Seaver's (2017) notion that access is not all-or-nothing but textured and constantly negotiated in the field also rings true for this study, as some interviewees were willing to speak on some parts of their work but not others. For example, one interviewee who worked in academia and had a collaboration with a tech company agreed to be interviewed only in their academic capacity. The dominance of researchers and academia in the final sample inevitably has shaped the themes emerging through repetition and salience (Ryan & Bernard, 2003), from the importance of conferences in paper 2 to the focus on 'AI' vs. 'machine learning' and not 'data science' (which is an industry title) in paper 1 (see chapter 5). Moreover, there might be something particular among ML practitioners who agreed to participate in the project, e.g., finding this topic worthy of exploration and an hour or so of their time. In fact, several interviewees worked on topics related to ethics and regulation (e.g., fairness, interpretability).

Twelve interviews took place via video call, but the majority were conducted in person, e.g., at the interviewees' offices or in cafes. The interviews

usually lasted around an hour, the shortest 40 minutes, and the longest 1 hour 40 minutes. While I did not find that online interviews elicited lower-quality data, it felt easier to establish a rapport and the desired tone in the face-to-face interviews, and I committed to the latter whenever possible. Some interviewees showed me their labs or their office before or after the interview; a few used their whiteboards or computers to provide examples. These engagements 'around the interview' helped establish rapport and added context and texture to the interaction.

4.4.3 Interview guide

The interview guide (see Appendix I) was developed to elicit meaning-making on how practitioners understood and experienced machine learning work, and ethics and governance in relation to it. The guide is centred around five main conversational points.

Background and self-identification/presentation. Focus on producing contextual information about the interviewees' professional background, current work, and positioning within the broader ML/AI landscape. This included questions about their affiliation with fields and disciplines and how practitioners would describe their work in different contexts and their relationship with different labels.

Research projects. Focus on eliciting a deeper, more practice-oriented inquiry into the practitioner's projects: topics and aims, data used, involvement of other actors, and distribution of labour among them, including the practitioner's particular role and tasks.

Valuation of research and impact. Focus on the practitioners' repertoires of valuation of ML research and practice, as well as their notions of impact to get into broader values and normativities ascribed to work that might or might not be connected to the emic 'ethics' category.

Ethics. Focus on how the practitioners themselves understood 'ethics' in the context of ML and how it connects to their own work. I drew on ordinary language interviewing (Schaffer, 2006) and cultural domain analysis (Bernard, 2017) techniques to develop questions like:

- Could you tell me what comes to mind if you think about 'ethics' in your [area of research/machine learning/data science, etc.]?

I would then ask how practitioners experienced those in their own work, e.g.:

- Could you tell me something about the last time you saw, read, or heard something about ethics?

These questions are anchored in practitioners' own understanding of 'ethics' while connecting what can be a vague, somewhat esoteric concept to their own work.

Governance. Focus on the practitioners' experiences with regulations and procedures related to institutional ethics (e.g., ethical approval of research projects) and data governance (e.g., GDPR compliance) and the meanings they ascribe to them.

As is common in qualitative research, topics and foci shifted somewhat during data collection. The topics that made up the three articles became salient, while other topics moved out of the spotlight largely due to reflections on the early(-ish) interviews, questions, and probes that seemed to produce rich narratives (Lareau, 2021). For example, the topic of paper 1, which is not immediately connected to ethics or governance, developed from a background interest and question. I view these changes as mostly productive and results of interaction with the field, which also changes the research and the researcher. However, conducting the interviews was not without contingencies and challenges, which affected the resulting data and findings. As noted by Seaver (2017), an interview is not that far from a contemporary technologist's experience. Many of the interviewees had regular meetings and presentations, gave talks, and provided commentary for the media. My position as an interviewer – as a PhD student but also a 'non-technical person' – shaped these dynamics as well. For instance, I noted that often it was challenging to get the interviewees to talk about their own experience rather than the state-of-the-art or standard practice in the field in general (e.g., regarding dealing with bias). These tendencies might be attributed not only to my way of asking questions but also to the 'expert' position the interviewees assumed in these moments of interaction. It was my task to ensure the palpability of the (co-)produced interview narratives, capturing details and rich descriptions over generalised answers (Small & Calarco, 2022). For this, I also needed to negotiate the level of knowledge about ML practice I wanted to present. While the outsider, uninformed position is often helpful to elicit what practitioners take for granted, I noticed that it often led to them stepping into a more teaching or science communication role, which is conducive to generalisation and sparing me the 'boring technical details' (which I was often interested in!). Thus, letting my interviewees know that I am familiar with at least some of the vocabulary (e.g., I knew what an 'embedding' was) helped ease communication and generate more experience-close discussions. However, I also needed to be careful not to overcorrect with this to maintain the epistemic benefits of the outsider role – as well as genuinely understand what the interviewees were talking about. Moments when I navigated these contingencies less optimally produced less rich and useful data.

Previously, I mentioned that practitioners' participation might have been shaped by interest in or acknowledgement of the importance of ethics and governance on their part. However, I am less worried about this affecting the narratives as a version of social desirability. First, my interest in ethics had less to do with assessing their practices and more with how they interpret what ethics entails. Even if they would tell me what they thought I want to hear (which sometimes they did, pointing me to topics or parts of their work that they thought would be most relevant to the interview or my research interest), this still reflected what they understand 'ethics' *to be* (regardless of what they think *about it*). More empirically, some interviewees openly shared that their work did not involve ethical considerations or that they 'just want[ed] the method to work', indicating the legitimacy of this position in relation to ethics in their social reality.

4.4.4 Ethical considerations

A project about ethics has ethical considerations of its own. Formal ethical approval of the interview study was granted by the research ethics committee of Aarhus BSS, Aarhus University (#2023-119-K, see Appendix II), and this section reflects on what attending to research ethics meant in practice.

One ethical consideration of this project is the burden 'cold' recruitment can impose on the studied population, i.e., ML practitioners who agreed to participate as well as the numerous 'potentials' who received an email or a LinkedIn message and two email follow-ups but did not participate. Some agreed to participate after a follow-up or two, which is why I continued to send them. Whether this inconvenience is worth the produced results is up to this community and to the reader – I did my best to make sure it is.

When I contacted potential interviewees, I described the project as interested in the culture of machine learning work and community, including the role ethics and governance play there. Although this could have limited recruitment, mentioning the supposedly more sensitive topics was in the interest of transparency with the participants and their comfort. I also communicated that the project focused on culture and meaning-making rather than on evaluating individual practitioners' conduct. The participants were given the consent and information forms, which we went over if they had questions. The interviewees did not receive compensation.

Throughout the project, I followed university practices of data protection by storing recordings and signed consent forms on secure servers and pseudonymising the transcripts as soon as possible. Furthermore, since the machine learning community (especially research communities or communities in smaller countries such as Denmark and the Netherlands) is relatively small, I needed to balance transparency and contextualisation of the presented

evidence with protecting the interviewees' anonymity in reporting and sharing interview quotes. For this reason, the tables describing the interviewees in this chapter and the empirical papers portray categories rather than individuals. I have previously elaborated on my recording and reporting of demographic characteristics. For related reasons, I avoided giving the interviewees pseudonyms, as they would either crudely match inferred identities (and reveal them) or distort the representation of the field in potentially problematic ways. Thus, when I describe the content of the interview narratives or share quotes, these descriptions are accompanied by the characterisations of the interviewees' professional contexts and specialisations. I thereby balance contextualisation with protecting anonymity by limiting the information provided when it is not analytically relevant. For example, I do not share practitioners' countries of employment or offer more general descriptions of their specialisation or title when it is not important to the interpretation.

Lastly, previous studies in STS have articulated the significance of the titles we as researchers use to describe research participants, e.g., when ascribing knowledge and expertise (Bennett et al., 2025). Following Bennett et al., I mostly refer to the participants as practitioners, 'in the pragmatic and literal sense of the term as "those who practice a profession or an art"' (Epstein 1995 in Bennett et al., 2025, p. 3287). Sometimes, I refer to researchers and scientists when focusing on specific activities and the academic system.

4.4.5 Transcription

The interviews were recorded with a voice recorder, and I took notes during and immediately after interviews (once, the recorder turned off mid-interview, so for the second part, there are only notes). The voice recordings were transcribed by me or by one of three student assistants, who received transcription instructions from me. First, the recordings were transcribed automatically, using the university software based on the Whisper speech-to-text model, and then corrected by one of us. I also listened to every recording the student assistants transcribed to check the transcriptions, edit (including pseudonymisation) as needed, and to familiarise myself with the data.

4.4.6 Coding

In coding the interviews, I employed an approach that fits with the abductive thematic method: having theoretically informed interests but being open to new things, as well as putting the themes emerging from the data in dialogue with theory as part of the analysis (Thompson, 2022). First, using NVivo software, all interviews were coded openly, producing granular codes in the grounded-theory fashion (Charmaz, 2006). To be able to return the data with a different focus and question, I also created broad codes for research interests

and interview questions (e.g., what comes to mind when you think about ethics?) as suggested by the flexible coding approach (Deterding & Waters, 2021). After the first coding round, accompanied by memo-writing, I noted the very broad themes of interest and sufficient salience that would make up three avenues of analysis: ethics (paper 2), labels (paper 1), and data (paper 3). Then, I began iterative focused coding, refining existing codes by comparing data with data and codes, and codes to each other, identifying the most relevant codes, and relating codes to each other hierarchically (Charmaz, 2006), which aided the development of themes and further theorisation. First, I conducted focused coding with an interest in ethics (see paper 2, chapter 6). Next, I went back into the data and initial codes to zoom in on labels, self-descriptions, and identities (paper 1, chapter 5). Lastly, I examined the data with an interest in ML practitioners' relationship to data (paper 3, chapter 7). Both times, I returned to the data with the new question, read through all interviews, looked at the codes generated in the first round, and then created a new folder for focused coding in order not to affect the rest of the analysis, adding the codes connected to the broad topic from the initial list to the relevant folder. See the three lists of codes in Appendices V-VII.

4.5 Secondary data: conference ethical procedures

Paper 2 also draws on secondary data: webpages featuring submission instructions and explanations, ethical codes, and guidelines of the following conferences: the Conference on Neural Information Processing Systems (NeurIPS), the International Conference on Learning Representations (ICLR), the International Conference on Machine Learning (ICML), the Conference on Computer Vision and Pattern Recognition (CVPR), and the Responsible NLP Checklist adopted in natural language processing conferences. These conferences were mentioned by the interviewees as relevant to their work (I included conferences mentioned by more than one interviewee). See the full list in Appendix IV.

This data source was added during the analysis of the interview data (April 2025). The topic of conference ethical statements was brought up by the interviewees early in the data collection. From then on, I asked researchers (including industry interviewees who publish some of their work) about their experience with statements and checklists. However, it was during focused coding that I noted similarities between how the interviewees talked about ethics generally and about the content of ethical statements. Moreover, several interviewees made narrative links between ethics and statements, e.g., mentioning the statements when I asked about ethics more broadly. Considering the salience of the statements' practice in the interview narratives, I decided to seek out previous publications on the topic (Ashurst et al., 2022; Liu et al.,

2022) and to explore the discourse of the conferences in comparison with the interview narratives. For this, I collected the conference materials from their respective websites for the 2024 and 2025 editions. Using both years seemed appropriate, since collection took place in spring 2025, and the interviews were conducted in spring-summer 2024, when the preparations for the 2025 iteration of some of the conferences could have already started.

The analysis of the conference materials focused on the themes present in the interviews. Thematic analysis often includes comparison between datasets, though usually groups of participants according to significant categories. Here, the datasets came from different types of data, but the comparison followed the same questions of whether themes were present in both datasets, and where the themes were differently manifested (Guest et al., 2012, pp. 162-163). Specifically, I compared conference requirements with the themes of ethics as applications, ethics as data, and ethics as a topic.

4.6 Analytical strategy

The coding process provided the basis for further abductive analysis (Thompson, 2022). The initial analysis was mostly empirically driven, grounded in the data. As patterns started to emerge through coding, the research (sub-)questions were adjusted to fit the emerging themes. At that point, I returned to within-case analysis of the interviews, gathering information about each interviewee's narrative to answer the research (sub-)question for their case and collecting all relevant information from the interview, thus summarising each interviewee's relation to the themes. Based on those summaries, I conducted cross-case analyses and noted relevant factors for group comparison (e.g., position on the basic–applied spectrum).

From that moment, these emerging themes and patterns were put in dialogue with theories to find the most theoretically insightful explanation for observed phenomena (Tavory & Timmermans, 2014). Through iterative comparison between theoretical approaches and the themes and patterns within the data, I arrived at the frameworks that best fit the data and produced theoretically significant insights, tinkering with the existing theories to work with my data (for example, adding boundaries as a type of semiotic code in paper 1) or developing new conceptual tools (the constellations metaphor, paper 2) when I felt that existing theories did not fully account for significant aspects of the empirical insights. Writing and re-writing the 'analysis' sections of the three papers and the cognitive validity of the produced descriptions were the respective methods and evaluation approaches for arriving at theoretical fit. This iterative process resulted in the development of the eclectic theoretical toolkit described in the previous chapter. The next chapters discuss the results of this process.

Chapter 5
Paper 1 – The Wrong Shortcut: 'AI' and representations of machine learning work

Abstract

Recent discussions about the hype surrounding artificial intelligence have sparked new debates about how language describing technology shapes its understanding and proliferation in society. This paper provides new insight into these dynamics by exploring how language around AI is navigated within computational cultures. Drawing on interviews with machine learning practitioners in academia and industry, the paper demonstrates how technologists' relationship with the terms 'artificial intelligence' and 'machine learning' is intertwined with their interactions with the general public, distinctions between the members of this professional field, and the popularisation of generative models, such as ChatGPT. Employing Stuart Hall's encoding-decoding model of communication, the analysis illustrates how machine learning practitioners take notice of the broader cultural understandings and connotations of 'AI' and other labels and draw on them to communicate their work to non-specialists. Used as a public-facing label, 'AI' brings diverse forms of machine learning work under one broad representational umbrella and draws the boundary between technologists and non-specialists. However, the analysis challenges this distinction by showing that some of these outward representations seep into how practitioners themselves conceive of their work and the AI landscape. By unpacking the intricate interpretative dynamics behind labelling practices, this paper emphasises the role of representation and communication in shaping the meaning of machine learning and artificial intelligence.

Introduction

> We usually joke within the community, saying that when you are writing a grant application or talking to a journalist, you say it's AI. When you're teaching, you're doing machine learning. And when you're finally sitting and implementing, it's linear regression.

This joke[1] came up in three separate interviews I conducted with machine learning (ML) practitioners in 2024 as we discussed how they describe their work: a topic that gave me 'terminological anxiety' (Seaver, 2017) as the lines between technologies, professional roles, and identities seemed blurry. The joke indicated shared culture and meaning-making, suggesting there were

more complex dynamics behind these labels that required deeper exploration. In this paper, I demonstrate what the labels 'artificial intelligence' and 'machine learning' reveal about computing cultures and their relationship with wider society.

What can we learn about AI by investigating 'AI'[2]? Coined in 1956 by John McCarthy and colleagues (McCarthy et al., 2006), the term 'artificial intelligence' has since come in and out of scientific and popular discourse. Recently, science and technology studies scholar Lucy Suchman (2023) has problematised what she calls the 'uncontroversial thingness of AI', arguing that the naturalisation of the fuzzy meaning of this word, which simultaneously presupposes a concrete, stable, and agentic referent, prevents a more critical examination of the sociotechnical systems it names. Overall, a significant body of research has studied discursive representations of AI, from its portrayal in media (Brause et al., 2023) to anthropomorphising language (Inie et al., 2026) and other linguistic devices in AI discourses (Kajava & Sawhney, 2023). However, technologists' own meaning-making on language around AI has not been explored in depth so far, even though prior works noted the ambivalent relationship technologists have with the notion of artificial intelligence (Gallagher et al., 2024; Joyce et al., 2021). Drawing on interview data, this paper addresses this gap by focusing on the following question: How do practitioners working on ML models consider labels such as 'AI' and 'machine learning' as representations of this work?

By focusing on representation, I destabilise the connection between labels and their referents (Hall, 1997), which fits with the idea of social construction of technology (Pinch & Bijker, 1984), but is also particularly necessary in the context of AI. Besides its current interpretative flexibility (Suchman, 2023), historical accounts emphasised how the understandings of artificial intelligence, machine learning, and related computational concepts have shifted throughout time (Dobson, 2023; Galanos, 2023). This changeability is reflected in the 'AI effect' folk theory: AI is redefined whenever it achieves a new goal (McCorduck, 1979 and others, in Galanos & Stewart, 2024, p. 28). Most notably, the neural-network approaches in the 2010s shifted from a peripheral to a central position in the AI agenda, becoming synonymous with the current understanding of AI (Galanos, 2023). Moreover, researchers' communication with funders and the media, and the discursive credibility contests between different competing approaches, have played a significant role in shaping the notions they employed (Cardon et al., 2018; Galanos, 2023).

Taking a constructivist approach thus equips me to view technologists' word choices not as an objective description or a presentation of self through manipulation of stable symbols but as an interpretative practice that (re)produces the meanings of the notions it uses. Using Hall's (1973) encoding-

decoding model of communication also allows me to situate these labelling practices within the broader cultural context and examine how the current public representations of ChatGPT and other generative ML/AI applications affect how those specialising in related technologies think and talk about their work.

In what follows, I develop the paper's theoretical framework, weaving Hall's theory together with insights from science and technology studies on technoscientific (buzz)words and symbolic boundaries. Next, I describe the interview study that forms the empirical basis of this paper, before moving on to an analysis of the dominant and negotiated codes organising practitioners' relationship with 'AI', and the relationship between the outward, public-facing representation of ML work and the representational dynamics under the AI umbrella.

Theoretical framework

Stuart Hall's approach to representation and the encoding-decoding model

Hall (1973) challenges the idea that words or other signifiers denote, i.e., directly refer to objects and phenomena, arguing that all signs are fixed in cultural convention only and are therefore subject to change. The denotations' supposed objectivity (i.e., naturalisation) thus indicates that these meanings are highly stable and agreed-upon compared to connotations: associations with other concepts and meanings based on convention. This notion sensitises the analysis to destabilising 'AI' and other terms' meanings as part of the broader 'interpretative flexibility' in social construction of technology (Pinch & Bijker, 1984) and allows me to focus on the dynamic connections between material phenomena, meanings, and signs.

The patterns of these links, termed 'codes', are the main analytical unit of Hall's theory (Hall, 1973). The codes can be linguistic, connecting concepts and words, but they also apply to culture more broadly. For example, the 'uncontroversial thingness of AI' noted by Suchman (2023) can be seen as an example of such a code. Within Hall's broad framework, I also integrate the notion of symbolic boundaries articulating relations of inclusion in and exclusion from categories and groups (Gieryn, 1999; Lamont & Molnár, 2002). Boundaries would be codes indicating the rules of inclusion/exclusion and similarity/difference: for example, why statistics-based approaches were often rejected as 'not AI' in the 1980s (Forsythe, 2001).

To produce and communicate meaning, actors continuously engage in active interpretative work of representation; they weave networks of codes when signifying meaning through discourse, which Hall terms 'encoding', and interpreting these messages, 'decoding' (Hall, 1973). Initially developed to analyse

television discourse, Hall's conceptualisation portrays a professional encoding and a 'lay' decoding. The former uses internal, 'professional codes,' but also necessarily involves drawing on the broader cultural meanings, including the (imagined) audiences' reactions. The joke at the beginning of the paper illustrates a set of meanings involved in an encoding: the audiences/ contexts, the terms, and the fit between them. In that sense, the encoding is also shaped by the potential decoding. Moreover, Hall argues that professional encoding, often inadvertently, reproduces the dominant (which also means, the more naturalised) ways of signifying situations and phenomena it portrays. The codes involved in encoding and decoding need some congruency for the message to be understood, but encoded and decoded meanings are not necessarily equivalent. So, Hall classifies decodings into dominant or preferred readings (strongly corresponding with the encoded meanings), negotiated decodings (taking in the broader signification of the encoded message but departing from it in some aspects), and oppositional readings (significantly contrary to the encoded meaning).

While this model has mostly been debated and applied in studies of mass media, it has also extended to studying science communication (see Raz et al., 2019), and I will show that this conceptualisation can be a productive lens for analysing how actors in technoscience use labels when communicating their work.

Encoding technoscientific work into labels

Representation dynamics are intertwined with the genealogies of machine learning and artificial intelligence; Dobson characterises them as both having material histories and as 'discursive constructions that depend as much on the vicissitudes of cultural dispositions, the changing uses and different connotations of these terms within government agencies, corporate marketing, and the media as they do on the structured logics internal to algorithms' (2023, p. 63). Other studies (Cardon et al., 2018; Galanos, 2023) noted that scientists' discourse took such connotations into account, for instance, avoiding association with approaches or technologies in which audiences have been disillusioned; these practices have contributed to the changing understanding of AI.

While ML/AI specialists' meaning-making on encoding their work in labels specifically is underresearched (see Gallagher et al., 2024 on communication more broadly), such work in the context of nanotechnology is of interest. Granqvist et al. (2013) examined how company executives used the 'nano' label to represent their firms, considering how the stakeholders would interpret it: whether the stakeholders themselves would categorise the firm as 'nano', and whether the labels would connote being cutting-edge, or hype and overpromising. The anticipation of such negotiated decodings of 'nano' made

executives hedge their association with the label, using it situationally or indirectly. This study emphasises that executives navigate potential decodings but are not in control of how this representation will be perceived. Among scientists, Bassett (2011) noted a '"disdain" [...] for the word nanotechnology, although they all used it to describe their work' (p. 148-149). Scientists problematised 'nano' as a buzzword, lacking substance and connoting strategic overselling, and worried that it portrayed their work as technology rather than science, which invokes societal concerns. This finding highlights a tension between the professional code and the outward encoding-decoding of their work with external audiences.

This theoretical framework denaturalises the meaning of labels and situates them within representation practices, in which actors can anticipate but not control their interpretation in different contexts. This framework will help me analyse machine learning practitioners' meaning-making about labelling their work with 'AI' and its others.

Methodology

This paper is based on a 2024 interview study focusing on the culture and ethics of machine learning work, within which the topic of labels and representation emerged, yielding rich material (e.g., the opening joke). The study aimed to capture the diversity of work on ML models, which can take various forms and be carried out in different contexts. To balance diversity and feasibility, recruitment was limited to Denmark, the Netherlands, and the United Kingdom. Combining cold invitations based on desk research and utilising my social networks, recruitment resulted in 44 interviews (see Table 1). Most interviewees worked in research (academic and/or industry) and development, but some also deployed and maintained active algorithmic systems. Participants worked with different methods, from statistical machine learning models to deep learning, and varied in the topics and types of data they worked on. The gender composition of the study reflects the imbalance in this field.

The interviews included questions about the practitioners' views on different terms related to their area of work, such as machine learning, AI, and big data; their professional self-identification and how they position their work in the broader landscape; how they would describe their work and specialisation to others; and the dynamics and development in their field, including public opinions and hype. 32 interviews took place in person and 12 online, with durations from about 40 minutes to 1 hour and 40 minutes.

Table 1. Overview of interviewed machine learning practitioners[a].

Total	44
Gender	Men (28), women (16)
Country of employment	Denmark (18), the Netherlands (17), the United Kingdom (9)
Academic employment	PhD student (12), postdoc (4), assistant professor (8), associate professor (2), professor (2)
Non-academic employment	Data scientist (6), research and development (8), business analyst (1), software Engineer (1)
Places of on-academic employment	Medium or large company (7), small company (7), startup (1), independent foundation (1)
Specialisation	Natural language processing (8), computer vision (5), biomedical engineering, bioinformatics & biostatistics (9) speech recognition (3), emotion recognition (2), privacy (3), fairness (3), reinforcement learning (1), causality (2), other (8)

a. Table 1 provides a simplified view by allocating each participant to one category to avoid double-counting. But some had multiple roles (e.g., academic and industry) or specialisations.

Transcribed interviews were coded in NVivo with a combination of granular open codes (Charmaz, 2006) and broad codes in line with the flexible coding approach (Deterding & Waters, 2021). Focused coding for this analysis then concentrated on a set of codes related to the topics mentioned above, creating themes and categories that were put in dialogue with literature, resulting in the use of the described framework.

Results

This analysis elucidates the codes structuring the interviewees' relationship with 'AI'. When conducting the interviews, I was somewhat surprised by how ready (at least some of) my interviewees were to discuss 'AI' not only as a scientific concept but also as a label with cultural connotations and communicative effects. While a few of them avoided 'AI' completely, and a few used it very non-problematically, many of them had opinions on and experiences regarding the term to share. The analysis explores these meanings and connects them to the broader cultural circulation of ChatGPT, large language models, and other generative AI products. First, I focus on the dominant code of 'AI' as a 'public-facing' signifier (Joyce et al., 2021) and its negotiated readings, before shifting to consider the extent to which they correspond with internal representation dynamics.

The dominant code: 'AI' as outward communication

The most salient way in which the interviewees characterised 'AI' was as a useful term for communicating with people outside the field. Many of the interviewees shared that they used 'AI' to present their work to non-specialists; some said that they did so while not being fully satisfied with its definition or fit with their work. Moreover, they often narrated similar explanations as to why 'AI' was useful for such communication, illustrated by this natural language processing (NLP) researcher[3]:

> If I talk outside of an academic context, I say whatever… seems to be best known, and that is often artificial intelligence rather than either machine learning or natural language processing. And I highly suspect that everybody interprets that differently as well, but then that's kind of more accessible. <…> by artificial intelligence, thanks to some companies, people tend to mean chatbots these days, so that's a shortcut, and it's a wrong shortcut, but I often use it.
>
> *– Wrong shortcut, because that's not what you work with?*
>
> Yeah, because whether it should be called artificial intelligence is questionable. But… I do identify myself as that when that helps establish a connection, or rather, explain to people faster roughly what I'm doing, even though it's not quite that.

This quote shows how the researcher considers her audience's potential interpretations and reactions when selecting a way to describe her work. The meanings she ascribes to 'AI' and how she anticipates this labelling will be decoded do not match. However, her narrative shows that the 'wrong shortcut' delivers on the broader encoding of such communications that emphasises recognition and connection. Other interviewees described similar encoding-decoding processes, explaining that 'AI' is more familiar to the general public than 'machine learning' or other narrower, more specialised terms that practitioners themselves considered more accurate or fitting. For example, an R&D specialist working in a large tech company explained:

> Everybody knows AI, so it's less difficult. I try not to use too much technical jargon. So, I always say, with a little bit of sarcasm: 'Oh, I'm an AI researcher.' So, I'm building a machine to talk, helping a machine to read text out. I'm using very generic terms.

Notably, this emphasis on avoiding 'technical' terms places 'AI' in the general rather than specialised category. The quote also features the other salient representation strategy: highlighting the tasks or applications of the work ('helping the machine to read text out', 'helping people sleep', and even 'reading people's minds'). This PhD student, for example, emphasised that the application of her work is what she brings up in conversation:

> I would definitely start talking about sleep, because it's a topic people can understand and relate to much more easily than abstract machine learning stuff.

Compared to machine learning, considered too abstract for a lay recipient to connect to, sleep is relatable. While using different signifiers, this is the same broader encoding that privileges recognition and connection, which resembles how, according to previous studies, scientists perceive science communication to the public and media (Davies, 2008; Felt & Fochler, 2012).

Moreover, the idea that the more technical facets of the work are difficult for non-specialists to understand and appreciate is also salient in how some of the interviewees encode their self-descriptions. For example, a PhD student explained:

> Especially people who aren't particularly mathematical or something like that, they just tune out. They're just: okay, cool, something with AI. Sure, whatever.

Such supposedly unengaging representations might be the comparison against which 'AI' shines. Notably, the similarly broad 'data science', an established title in the industry, was not attributed the same communicative features. Conversely, the concision of 'AI' as a signifier is also effective when attention is not captured. This makes 'AI' a safe start, with a possibility to go into more details or quickly move on, as a more nuanced explanation of the work might not be appreciated. For example, one academic researcher described that she used to 'lose [her] audience' due to getting into the details and using jargon when describing her work. However, over time, she developed strategies for not doing so:

> I first try to get a feel of where they're coming from… Like, what is the background? Is it non-technical, technical, AI, non-AI? But if it's completely non-technical, I use the word AI. I just say: 'Oh, I do AI.' <…> And then if it's a person who is from medicine, I talk to them about this. <…> So, I try to give them examples of something <…> they can connect to.

This quote illustrates how conversational discourse constructs and classifies audiences through the use of more or less specialised terms (Kitzinger & Mandelbaum, 2013). Thus, 'AI' marks and reproduces the boundary between insiders and outsiders to the field, referred to as the general public, lay persons, or even 'non-technical' people.

For example, a computer vision researcher explained that she would put 'AI' in grant applications to signal what the proposal was about for readers from other fields. Considering the audiences' potential decodings and resulting communicative effects goes into the practitioners' encoding, which entails categorising their audiences and drawing boundaries.

The representation practices described above draw on broader cultural understandings of 'AI', a connection that becomes more apparent in the moment of change. Taking place 15-20 months after ChatGPT's release in November 2022, the interview narratives often invoked this application. The next section explores the association of AI with ChatGPT (or similar products) as a nascently dominant code and its implications for technologists' encoding of their work.

ChatGPT as a dominant referent of 'AI'

Many of the interviewees explicitly attributed a change in the encoding-decoding of their work in communication with general audiences to the broad cultural circulation of ChatGPT uptake and media coverage. First, this circulation made the term 'AI' more recognisable, as illustrated by this PhD student:

> Up to a year ago, I always told people I did machine learning. Today, I would very much associate myself with machine learning, but AI has become a popular word in the media. <…> So, sometimes I say I'm an AI researcher just because people know what it's about. <…> I come from a small village in the Netherlands, and I guess, if I asked random people in the village if they know machine learning, probably not so much.
>
> *– Yeah, but with AI, maybe.*
>
> With AI, quite. Since ChatGPT, yeah.

In addition to the people's increased familiarity with the 'AI' term, the interviewees noted that their interlocutors associated the word with ChatGPT and similar products. This dominant connotation, or even denotation, has two implications for practitioners' communication. As a concrete application, ChatGPT roughly denotes their work and makes it more relatable. But as this academic researcher shares, it also brings in various connotations with these products:

> I say I research AI. The first thing I get nowadays <…> is: 'Oh, so ChatGPT'. Because it is the best-known AI model, and it is used by almost everyone. <…> Otherwise, the second most frequent question I get is: "Oh, so you'll take away all the jobs". So, people understand that what we are doing is interesting, important, but they're also scared of it. So, that's the two things I get. Prior to ChatGPT, I used to get more scared people. Now they are interested because of ChatGPT. Because they realise that it can do amazing things. But again, they're saying: <…> okay, if these large vision models can give you very nice art, or ChatGPT can answer all the questions you have, then what happens to the teachers? What happens to the artists?

This quote connects people's association of his work with ChatGPT through 'AI', and notes the communicative effects of this link: general excitement about AI and interest in the researcher's work. Several other interviewees described similar experiences. However, this narrative also highlights the more negotiated interpretations of 'AI research'. Concerns about job losses, for example, still stay within the dominant interpretation of AI as incredible technology with potential for major impact (both positive and negative). And while the interviewees generally portrayed others' reactions to representations of their work as either interested (helped by 'AI') or indifferent to this technical work, some of them also mentioned concerns and fears as reactions to it. These findings showcase how broader discourses seep into researchers' and practitioners' lives.

Not only their interlocutors, but also practitioners themselves invoked meanings that can be interpreted as negotiated/oppositional readings of the dominant code, coming from their positions as both encoders and decoders in the broader circulation of the term. While less salient, they are important for understanding the limitations of the interviewees' identification with 'AI'.

Negotiated codes: too much of a buzzword?

These negotiated readings draw on the same crucial features of 'AI' as a communicative device, but invoke other, more problematic, connotations. First, some interviewees problematised the fuzziness of the term's denotation, a critique connected to the dismissal of buzzwords as meaningless (Bensaude Vincent, 2014). Secondly, the term's use to attract attention was also associated with strategic self-promotion and (over)selling one's work. In some cases, the idea that the term and its interpretations bring individual researchers and the field in general more attention was more accepted and viewed positively. For example, one academic researcher explained that although he identifies more with machine learning,

> it does not mean that AI as a term is all bad. It has gotten us a lot of attention. Who does not want it in academia, so it is okay…

Furthermore, he viewed effective communication as a legitimate and even responsible academic practice:

> I also think it's my responsibility to engage with different stakeholders. And if using a different term helps me engage more, why not?

However, some other interviewees were uneasy about this promotional connotation. For example, a PhD student shared:

> I just get a kind of ick when people use AI. <…> it always seems like… consultants' language for machine learning. But I think that's also kind of a very

> North Jutland, Danish way of not really wanting to be in newspapers or marketing for your research, right? Maybe that notion will be a pitfall for future funding applications, but I kind of try to stay away from it.

This quote illustrates a variety of cultural codes contributing to this 'ick': from boundaries between scientists and non-scientists (consultants) to the idea of the medialisation of one's research being against the Danish national code of not promoting oneself (see Horst, 2013). Similarly, I could not help but think about the gendered aspect of claiming expertise and value when another interviewee, who worked in a small AI company, explained that she did not tell people that she works on AI as she finds it 'cocky'. When I further asked whether the company positions itself in such a way, she specified:

> When we're at conventions and stuff, then we are the AI company that you need to go to. Then we're as cocky as we can get. But in general, we're not.

Both quotes emphasise how certain contexts (funding applications, conventions) are more associated with self-promotion and 'AI' as a communicative device. However, these promotional connotations then trace back to the term.

Another problematic connotation related to both the fuzzy denotation and promotional qualities of 'AI' is its stretching by those 'saying that they are doing artificial intelligence just for marketing purposes', as one academic researcher put it. This boundary-drawing is not just about (self-)promotion but also about what does or does not count as AI. Despite its denotative breadth and interpretative flexibility, there are boundaries to 'AI', and several interviewees mentioned examples of self-proclaimed AI work they would not categorise as such. But when I then probed about judging or contesting someone's labelling, the interviewees distanced themselves, explaining that words did not matter to them or that each person could call themselves whatever they wanted. However, these connotations may lead practitioners to disassociate themselves from the term, as illustrated by this PhD student:

> From the perspective of being a PhD in AI, I kind of try on purpose not to do that. Because I feel like AI has become such a buzzword that if I'm in the pub and tell people 'hey, I work on AI', I feel like everyone says they are working on AI nowadays.

This quote connects 'AI' being a buzzword to the stretching of its meaning so broadly that it does not represent the interviewees' 'real' work on AI.

Lastly, some of the interviewees worried about the grand expectations that 'AI' connotes. As a practitioner working in a consultancy explained, the hype around AI can contribute to viewing it as '*a* hammer that can solve everything' rather than focusing on existing problems and tailoring solutions to them, and lead to expectations that cannot be met technically. On the other hand, 'AI'

also connotes concerns about existential risks or similar dystopian narratives. As an academic computer vision researcher describes:

> I think AI already brings up ethical questions… from the outset. It kind of flags some idea of AI taking over the world and stuff already, but machine learning is a much duller description. Which is probably good.

Compared to 'machine learning', which foregrounds a narrower technological framing, 'AI' has broader connotations that practitioners cannot control but have to navigate. Overall, these negotiated codes showcase the other side of the same coin of 'AI' as a broad term that generates excitement and engagement, and contribute to the more ambivalent relationship practitioners have with 'AI' as a representation of their work.

Practitioners' narratives present 'AI' as bringing different kinds of ML work under one broad umbrella in public-facing communication. In that, the interview narratives articulate a boundary between the outward and internal representations of their professional field. The rest of the analysis considers how porous that boundary is by considering whether the dominant codes identified above – 'AI' as a signifier of ML work and ChatGPT and generative or large language models as its main referent – seep into how practitioners conceive of their work among themselves.

Dominant vs. professional codes

The interviewees described 'AI' as something for the benefit of the general public and less relevant for technologists themselves, and a few interviewees even expressed that this term should not be used among specialists. But the use of 'AI' in the interviews was not confined to accounts of self-description to outsiders. Many interviewees also invoked it when talking about large language models, research topics, and fields, and referred to their work as such in passing. This also happened with interviewees who criticised the term and described it as a communication device; only a few abstained from using the term overall. To some extent, this could be because conversing with me about their work was, in that moment, enacted as communication with a 'non-technical' person.

However, the continuous use of the term in (outward) communication might also domesticate it for practitioners themselves. The field is characterised by intensive communication to various mixed audiences (Gallagher et al., 2024). Whether for promotional purposes or to make the work understandable to scientists from other disciplines, practitioners might use the term in such communication, as illustrated by a PhD student working on ML for sleep, who explained:

> In the paper that I have, I've used AI three times specifically. One's in the title, one's in the abstract, one's in the introduction. Then I use machine learning the whole way through. And it's to… because it's a buzzword, so people can find the articles. That's the only reason that I have the words AI.

As practitioners represent their work as 'AI', they reproduce the connotative link between the label and their own work, if only three words at a time. Moreover, this representation is often not up to them: 'AI' can be signified in degree names, job titles, calls for proposals, etc. For example, one of the interviewees described how in her previous job, titles were changed from 'data scientist' to 'AI specialist', and that some of the employees did not appreciate the new title. Entire subfields, such as 'explainable AI' (Dwivedi et al., 2023), emerged with 'AI' in their names. And of course, discussions about artificial *general* intelligence (AGI) (Graham, 2022) centre around this concept. Thus, whether practitioners want this or not, 'AI' seeps into the representations of their work.

The situation with ChatGPT and generative models as the dominant denotations of 'AI' is more complex. Reminiscent of the 'AI effect', this code focuses on a particular type of model while rendering others less visible as AI. Although interviewees varied in their opinions on generative models, it is notable that the interview data included almost no frustration with this representation or narratives arguing for non-generative machine learning's inclusion into and recognition as AI. Even researchers and practitioners who did not work on generative models accepted this representation, as exemplified by the sentiment expressed by several interviewees that the uptake of ChatGPT gave their colleagues and friends a better understanding of their work. Only a couple of interviewees even addressed this distance. One of them was a PhD student working on privacy-preserving machine learning. To my question on whether he identifies with 'AI', he replied:

> Yes, but I do view it as something very different from ChatGPT or something. Yeah, it's a very different subject, but I am very much working with AI.

Notably, this acknowledgement was not followed by further rhetorical justifications that privacy-preserving machine learning counts as AI. While the dominant code might partly explain the interviewees' lack of recognition issues, it also relies on another code: the uncontentiously accepted relationship between machine learning and AI. Consider how it is articulated in this scientific software engineer's reflection:

> I would say machine learning is definitely part of AI. What else does AI involve? Like that's a bit blurry, I think.

Although the content and boundaries of 'AI' are not fully clear, machine learning's position is core and undebated, even though the interviewee herself acknowledged it has not always been so:

> I studied AI just before that real machine learning boom, so we also had a lot of more classical logic-type things.

Mostly, machine learning was described as part of AI, although some practitioners almost equated the two, like this PhD student specialising in reinforcement learning:

> Right now, there is no other real working AI outside of machine learning and deep learning <...>. So, it's clear what you're talking about. If a new paradigm comes up that is not machine learning for AI, then maybe I would get a bit more like, 'oh what are you talking about?' But for now, I'm fine.

Both quotes address the changing meaning of 'AI' as well as machine learning's dominance in it at the moment. Thus, this code entails that if someone works with ML, they automatically do AI, and machine learning does not necessarily have a 'generative' or 'language' connotation.

Thus, the public's dominant association of 'AI' with generative products does not seem to threaten practitioners' claim to the category, as the professional code puts ML at the forefront. In that, the boundary between the internal and outward representations holds. However, this dynamic becomes more complicated when considering the relationship between two fields/categories connected to large language models, presented as the dominant denotation of 'AI': machine learning and natural language processing.

While many forms of ML work gather under the 'AI' umbrella in outward communication, the distinctions between them matter among professionals. For example, the interviewee for whom 'AI' worked as a 'wrong shortcut' asked for my definition of ML researchers at the beginning of our conversation, explaining her question so:

> Depending on how you ask it, I would say or not say that I identify with machine learning research. If you asked in a context in which it would be clear it's about the methods one uses...Yes, sure, I do machine learning, so does pretty much everybody these days. In my community, which is the natural language processing community, there is also the...I guess the distinction of the focus of this work, which then roughly breaks into the conferences that people tend to publish at. <...> So, <...> if I were at an [NLP] conference, would I say that I'm a machine learning researcher? Probably not.

This quote illustrates that different codes are at play among specialists: doing machine learning does not make one an ML researcher. Further elaborating

on the difference, she says: 'I also want to know about language, and that's not part of the machine learning agenda.'

The code she articulates to draw the boundary between machine learning and NLP researchers, including herself, is a distinction between methods and applications. A separation between computational techniques and substantive areas of life and knowledge in which they are applied (domains) has been described as naturalised in contemporary computational cultures (Ribes et al., 2019). This distinction seems to translate to fields: many of the interviewees invoked this code, referring to the 'general' or 'methods' ML field versus those formed around certain topics, language being one of them. However, some of the further exchanges with NLP researchers indicated that the way the relationship between these categories is or has been changing, not least due to the popularisation of large language models within and outside the field. A few of the study participants who identified with NLP were interviewed during major 'general' ML conferences. Their presence in these spaces complicated the harsh boundaries drawn by the researcher quoted above. When I asked one of them, a PhD student working on machine translation, whether NLP was a niche, marginal topic in this space, he explained:

> I would have said yes two years ago, but the thing is that after ChatGPT, I would say now it has taken the main stage, and it was a weird transition. <…> the usual setting in which an NLP researcher was doing research was basically, you have all these machine learning labs. And when you say machine learning, you're taking for granted that they're doing computer vision or… something else. And then there were the NLP labs that, in many cases, were affiliated with humanities departments. <…> And then I would say now, at these kinds of conferences, you're likely to find a kind of an equal balance of topics. And it's also interesting that a lot of people who never did NLP are suddenly experts in NLP, right? Because it's the hyped topic.

This quote illustrates the shift of NLP's position in the machine learning spaces, from being far from the dominant denotation (that was 'computer vision or…something else') to being a popular topic and even taking the 'main stage'. Importantly, this shift is more than labelling and involves a change in epistemic practices: NLP research drawing more on the machine learning components than linguistic theories, and ML scholars engaging with language modelling. It is the latter that is the focus here. Indeed, several interviewees, both among NLP and 'general' ML researchers, mentioned that large language models were the hyped or popular topic at the time (spring-summer 2024). The notion of the 'hyped topic' reflects how some of the researchers characterised the development of the general ML field: as another PhD student described, 'there's so much to do, AI so popular that you sort of follow the hype, get a few papers out, and then move on.' This means there is always the hyped

topic of the moment, which is included and centered in the machine learning field (for a time): a professional code as another version of the 'AI effect'(McCorduck, 1979 and others, in Galanos & Stewart, 2024, p. 28). The hyped topic of the moment, being connected to ChatGPT, thus shows stronger correspondence between the outward and internal dominant denotation. While it seems not to threaten ML practitioners who do not work on the hyped topic, it affords the language 'domain' more space in the method-identifying field and a stronger connotation with 'machine learning' research.

Discussion

Recent critiques problematise the naturalisation of computational terminologies (Lee & Ribes, 2025; Suchman, 2023). In this paper, I highlight the intricate interpretative work of representation behind the use of two such terms, 'AI' and 'machine learning'.

First, I elucidate the narrative explanations behind the often-mentioned but underexplained tendency for 'AI' to be a public-facing term for machine learning practitioners: 'AI' is portrayed by the interviewees as an imperfect but useful signifier fitting the broader constructions of communication with non-specialists. As such, 'AI' marks the boundary between the technologists and 'non-technical people'. However, the interview narratives also contained ambivalence around the label in the more contrary interpretations of the term's buzzword qualities. Moreover, I have observed these meanings' connection to the broader AI representations, such as the nascent dominance of ChatGPT and other generative products as referents of 'AI'. These findings contribute to emerging literature on the communication performed by ML practitioners (Gallagher et al., 2024) and technologists' relations with the broader public (Su et al., 2021), as well as the (re)production of the figure of AI through certain practices (Suchman, 2023). Through its use for the benefit of non-specialists as one of these practices, the broad fuzzy meaning of the term persists and even gets more naturalised among technologists themselves.

Furthermore, I complicate the idea of an emerging single AI field (Berman et al., 2025) by pointing to the complexity of internal representation dynamics behind what in outward communication appears as a singular figure of AI. Intertwined with epistemic and paradigmatic changes, internal and outward representational shifts contribute to the changing meaning of AI (Galanos, 2023). The study thus provides empirical evidence regarding the 'AI effect' (McCorduck 1979, in Galanos & Stewart, 2024) and emphasises the role of representation and communication in it.

This paper cannot conclude without recognising the limitations of the study. Crucially, interviews are also a form of communication with their own encoding-decoding processes. And as the constructions of audiences were so

salient in practitioners' narratives, my own figure as a decoder might shape the results from both sides. Thematically, the interviews did not probe for the use of (buzz)words for acquiring funding and other self-promotional activities, as the interviewees' framing of 'AI' did not centre around them. Moreover, this analysis has focused on AI and ML, leaving 'data science' in the background; this was partly due to the prevalence of academics among participants, while data science is an industry title. Conceptually, Hall's (1973, 1997) theories fit well with the interview data and equipped me to highlight both practitioners' meaning-making about communication and its connections to broader cultural codes. However, while this theoretical framework views meanings as mutable, it is less conducive to explaining the mechanisms of their change. Finally, this focus on language does not address the role of the distinctively material aspects of the social construction of technology (Pinch & Bijker, 1984): for example, the resource-hungry production of deep learning models. To truly understand the (re)production of the changing figure of AI, further work needs to consolidate the symbolic and material aspects.

Both theory and empirical research on AI highlight that its meanings are prone to change. Ongoing developments and discourses, from concerns about existential risk to the spread of generative tools-assisted 'vibe coding', can alter the future connotations of 'AI'. How would this affect the representation patterns? The insights from previously hyped fields like nanotechnology and the negotiated codes examined in this paper suggest some of the potential avenues: from the narrowing of the term's denotation to differentiate the core community, as happened with 'nano' (Grodal, 2017), to its avoidance due to societally unacceptable connotations, as in the biotechnology case (Felt & Fochler, 2012). Future research should also investigate past and ongoing (the next 'AI' breakthrough might already be developing under another name) label-shifting in research papers, job titles, etc., thus more systematically exploring the labelling patterns along the summers and winters of 'AI' (Galanos, 2023).

In conclusion, let us return to the joke with which this paper opened. While the first part about 'AI' as a public-facing signifier aligned with the research findings, the punchline took a different direction in terms of the dynamics under the 'AI' umbrella. More importantly, the analysis has demonstrated that the dynamics in which labels are involved are more complex than symbolic impression management. These open new questions for how interpretive flexibility of technology (Pinch & Bijker, 1984) and naturalisation of categories by communities of practice (Bowker & Star, 1999) are intertwined with communication and technological hype (Gallagher et al., 2024).

Notes

1. Previously, this joke appeared in a blog post by a researcher and engineer, Joe Davison (2018).
2. 'AI', with quotation marks, refers to the term/ the signifier.
3. Here and throughout, quotes are lightly edited for readability without affecting their meaning.<...> indicates that text was cut for focus and space limitations.

References

Bassett, D. R. (2011). Notions of identity, society, and rhetoric in a speech code of science among scientists and engineers working in nanotechnology. *Science Communication*, *34*(1), 115-159. https://doi.org/10.1177/1075547011417891

Bensaude Vincent, B. (2014). The politics of buzzwords at the interface of technoscience, market and society: The case of 'public engagement in science'. *Public Understanding of Science*, *23*(3), 238-253. https://doi.org/10.1177/0963662513515371

Berman, G., Williams, K., & Cohen, E. (2025). The benefits of being between (many) fields: Mapping the high-dimensional space of AI research. *Big Data & Society*, *12*(1). https://doi.org/10.1177/20539517241306355

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press. https://doi.org/https://doi.org/10.7551/mitpress/6352.001.0001

Brause, S. R., Zeng, J., Schäfer, M. S., & Katzenbach, C. (2023). Media representations of artificial intelligence: Surveying the field. In S. Lindgren (Ed.), *Handbook of Critical Studies of Artificial Intelligence* (pp. 277-288). https://doi.org/10.4337/9781803928562.00030

Cardon, D., Cointet, J.-P., Mazières, A., & Carey-Libbrecht, L. (2018). Neurons spike back. *Réseaux*, *211*(5), 173-220. https://doi.org/https://doi.org/10.3917/res.211.0173

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE.

Davies, S. R. (2008). Constructing communication:Talking to scientists about talking to the public. *Science Communication*, *29*(4), 413-434. https://doi.org/10.1177/1075547008316222

Davison, J. (2018). No, machine learning is not just glorified statistics. *Medium-Toward-Data-Science. com*. https://medium.com/data-science/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3

Deterding, N. M., & Waters, M. C. (2021). Flexible coding of in-depth interviews: A twenty-first-century approach. *Sociological Methods & Research*, *50*(2), 708-739. https://doi.org/10.1177/0049124118799377

Dobson, J. E. (2023). *The birth of computer vision*. University of Minnesota Press. https://doi.org/https://doi.org/10.5749/9781452968889

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., & Morgan, G. (2023). Explainable AI (XAI): Core ideas, techniques,

and solutions. *ACM computing surveys*, *55*(9), 1-33. https://doi.org/https://doi.org/10.1145/3561048
Felt, U., & Fochler, M. (2012). Re-ordering epistemic living spaces: On the tacit governance effects of the public communication of science. In S. Rödder, M. Franzen, & P. Weingart (Eds.), *The Sciences' Media Connection –Public Communication and its Repercussions* (pp. 133-154). Springer Netherlands. https://doi.org/10.1007/978-94-007-2085-5_7
Forsythe, D. E. (2001). *Studying those who study us: An anthropologist in the world of artificial intelligence*. Stanford University Press. https://doi.org/10.1515/9781503619371
Galanos, V. (2023, 21.11.2023). *To have done with the metaphor of summers and winters: Can AI and Internet history cure hype?* http://dx.doi.org/10.2139/ssrn.4640305
Galanos, V., & Stewart, J. K. (2024). Navigating AI beyond hypes, horrors and hopes: Historical and contemporary perspectives. In A. Ponce del Castillo (Ed.), *Artificial intelligence, labour and society* (pp. 27-46). ETUI. https://www.etui.org/sites/default/files/2024-03/Chapter2_Navigating%20AI%20beyond%20hypes%2C%20horrors%20and%20hopes%20historical%20and%20contemporary%20perspectives.pdf
Gallagher, J. R., Avgoustopoulos, R. E., Hamilton, A., & Seilkhanova, T. (2024). The challenges and writing practices of communicating artificial intelligence and machine learning in an era of hype. *AI & Society*, *40*, 2895-2908. https://doi.org/10.1007/s00146-024-02049-0
Gieryn, T. F. (1999). *Cultural boundaries of science: Credibility on the line*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226824420
Graham, R. (2022). Discourse analysis of academic debate of ethics for AGI. *AI & Society*, *37*(4), 1519-1532. https://doi.org/10.1007/s00146-021-01228-7
Grodal, S. (2017). Field expansion and contraction: How communities shape social and symbolic boundaries. *Administrative Science Quarterly*, *63*(4), 783-818. https://doi.org/10.1177/0001839217744555
Hall, S. (1973). Encoding and decoding in the television discourse. In D. Morley (Ed.), *Essential Essays, Volume 1: Foundations of Cultural Studies* (pp. 257–276). Duke University Press. https://doi-org.ez.statsbiblioteket.dk/10.2307/j.ctv11cw7c7.16
Hall, S. (1997). The work of representation. In *Representation: cultural representations and signifying practices*. Sage in association with the Open University. https://us2.sagepub.com/sites/default/files/upm-binaries/55352_Hall_ch_1.pdf
Horst, M. (2013). A field of expertise, the organization, or science itself? Scientists' perception of representing research in public communication. *Science Communication*, *35*(6), 758-779. https://doi.org/10.1177/1075547013487513
Inie, N., Zukerman, P., & Bender, E. M. (2026). De-anthropomorphizing "AI": From wishful mnemonics to accurate nomenclature. *First Monday*, *31*(2). https://doi.org/10.5210/fm.v31i2.14366

Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S. G., Noble, S. U., & Shestakofsky, B. (2021). Toward a sociology of artificial intelligence: A call for research on inequalities and structural change. *Socius: Sociological Research for a Dynamic World*, *7*. https://doi.org/10.1177/2378023121999581

Kajava, K., & Sawhney, N. (2023). Language of algorithms: Agency, metaphors, and deliberations in AI discourses. In S. Lindgren (Ed.), *Handbook of Critical Studies of Artificial Intelligence* (pp. 224-236). Edward Elgar Publishing. https://doi.org/10.4337/9781803928562.00025

Kitzinger, C., & Mandelbaum, J. (2013). Word selection and social identities in talk-in-interaction. *Communication Monographs*, *80*(2), 176-198. https://doi.org/10.1080/03637751.2013.776171

Lamont, M., & Molnár, V. (2002). The study of boundaries in the social sciences. *Annual Review of Sociology*, *28*(1), 167-195. https://doi.org/https://doi.org/10.1146/annurev.soc.28.110601.141107

Lee, F., & Ribes, D. (2025). Computational universalism, or, Attending to relationalities at scale. *Social Studies of Science*, *56*(3). https://doi.org/10.1177/03063127251345089

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12-12. https://doi.org/https://doi.org/10.1609/aimag.v27i4.1904

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, *14*(3), 399-441. https://doi.org/https://doi.org/10.1177/030631284014003004

Raz, A., Pontarotti, G., & Weitzman, J. B. (2019). Epigenetic metaphors: An interdisciplinary translation of encoding and decoding. *New Genetics and Society*, *38*(3), 264-288. https://doi.org/10.1080/14636778.2019.1601009

Ribes, D., Hoffman, A. S., Slota, S. C., & Bowker, G. C. (2019). The logic of domains. *Social Studies of Science*, *49*(3), 281-309. https://doi.org/10.1177/0306312719849709

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2). https://doi.org/10.1177/2053951717738104

Su, N. M., Lazar, A., & Irani, L. (2021). Critical affects: Tech work emotions amidst the techlash. *Proceedings of the ACM on human-computer interaction*, *5*(CSCW1), 1-27. https://doi.org/10.1145/3449253

Suchman, L. (2023). The uncontroversial 'thingness' of AI. *Big Data & Society*, *10*(2). https://doi.org/10.1177/20539517231206794

Chapter 6
Paper 2 – Where is ethics in machine learning? Constellations of ethics in representations of machine learning work

Abstract

Studies on artificial intelligence and machine learning (ML) have researched how technologists construct their work and how they understand ethics. This paper foregrounds the relationship between the two by exploring how ethics – the way technologists themselves understand it – fits into their representations of ML work. Drawing on interviews with Europe-based ML practitioners and the ethical requirements of computing conferences, it develops a conceptual metaphor of constellations to analyse the constructions of ethics in ML work emerging from the data. Like constellations in the night sky, emic notions of ethics are comprised of (narrative) points brought together through interpretation. The analysis identifies three such constellations: 1) ethics of applications, 2) ethics of data, and 3) ethics as a research topic. These form distinct representations of ethics in ML work related to different locations within ML practice: for instance, the ethics of applications refers to the broad scope of ML work based on its range of downstream (mis)uses, whereas the ethics of data concerns more granular practices of data work. The constellations metaphor also allows the analysis to explore how these representations are both situated in practice and underscored by broader narratives and institutionalised procedures.

Introduction

Social studies of technology have explored how actors within the machine learning (ML) and artificial intelligence (AI) field[1] conceive of their work, arguing that these constructions have implications for the sociotechnical systems they build. In her ethnographic research in labs developing early AI models that were supposed to imitate a human expert's decision-making, Forsythe (2001) showed that building systems through writing code, which is what researchers considered to be the 'work' of artificial intelligence, represented only a small portion of their activities. She further argued that overlooking less technical tasks, such as interviewing experts to collect data, contributes to the narrow functionality of the developed systems. More recently, Bennett et al. (2025) argued that the notion of a linear 'pipeline' of AI development obscures the iterative nature and intricacy of data work, which leads to the underappreciation of both data annotators' labour and the role of data in AI systems. And Widder and Nafus (2023) demonstrated how the imaginary of AI development

as a supply chain consisting of isolated modules limits practitioners' sense of responsibility for potential harms.

In this paper, I bring this attention to representations of work to studies of technology ethics as an empirical field (Frahm & Schiølin, 2024). This burgeoning literature has explored how data scientists, ML/AI researchers, and engineers understand the ethics of their work, showcasing how their moral logics, understanding of responsibilities for harm, and attitudes towards the work of ethics are embedded in the professional context (Avnoon et al., 2024; Widder et al., 2022). Some of these studies noted that data and AI professionals often present their work as devoid of 'ethics', whereas observation emphasizes that it is value-laden (Barocas & boyd, 2017; Slota et al., 2023). These findings actualise the questions of what both technologists and those researching them mean by 'ethics'. Moreover, from the representations of work perspective, this invisibility might be indicative not of the presence or absence of ethics, but of situated representations shaped by specific positions in networked relations, cultural understandings, and interests (Suchman, 1995). Here, I put these cultural dynamics of representation centre stage as I investigate 'ethics' as an actor's category by considering how ethics – the way technologists themselves understand it – fits into their representations of machine learning work.

To do so, I draw on 44 interviews with ML practitioners based in Denmark, the Netherlands, and the United Kingdom and the ethical requirements of conferences in machine learning and related fields of natural language processing and computer vision. To examine constructions of ethics in these narratives, I take inspiration from scholars who used the analogy of maps to emphasize the (in)visibilities and inclusion-exclusion relationships that representations construct (Gieryn, 1999; Suchman, 1995). This leads to a focus on this question: Where and how is ethics featured in practitioners' representations of ML work? By using a 'where' question, I consider which elements of their work practitioners connect to ethics, for whom or when it is supposed to be relevant to varying degrees, and the form it takes in these narratives.

Drawing upon my findings, I introduce the metaphor of constellations to better capture the distinctive characteristics of representations of ethics that emerge in the interviews. Constellations of stars in the night sky exemplify how interpretation creates meaningful figures out of disparate objects that *appear* grouped together based on the available ways of perception and concepts. This metaphor helps me think about how ML practitioners form representations of ethics in their work based on various meanings and practices. Constellations connect some objects in the night sky and leave out others. Using this metaphor to analyse narratives of work brings attention to tacit inclusion-exclusion dynamics involved in making sense of one's world, rather than

explicit contestation implied by boundary-work (Gieryn, 1999). Moreover, it takes into account the situatedness of such representations (as constellations' appearance differs depending on one's position in space and time) as well as emphasizes the role of shared cultural concepts and categories in interpreting them. These features make the constellation metaphor a productive tool to theorise about ethics as an emic concept and offer a new perspective on studies of AI ethics in practice.

By tracing the cultural tools based on which practitioners make out certain constellations of ethics, I also consider how these representations relate to formalised, institutional approaches to the responsible development and use of data and AI. In recent years, many of these efforts have focused on ethics (Greene et al., 2019; Kerr et al., 2020), leading to an observation that AI has experienced the 'ethical turn' (Andrejevic, 2020). Moreover, in Europe, data and AI have come under the remit of legal governance with regulation such as the General Data Protection Regulation (GDPR) and the developing AI Act (for overview, see Krarup & Horst, 2023). How these institutional conditions affect practitioners' relationships with ethics is an important question, as studies have noted that formalised approaches tend to narrow the meaning of ethics to the scope of their regulation and signal that ethical issues have been solved (Heimer, 2013; Hoeyer & Tutton, 2005).

In what follows, I develop the paper's theoretical framework, elaborating on spatial concepts in science and technology studies (STS) approaches and research on ethics and related concepts, such as responsible research and innovation (RRI). Building on this, I introduce the constellations metaphor. After a description of data and methods, the results section explores the three constellations of ethics identified in practitioners' narratives and how they are informed by broader discourses and institutional procedures. The concluding discussion considers the implications of the findings for discussions of responsible AI and reflects on the study's approach and the constellations metaphor.

Theoretical framework: from maps to constellations

Spatial thinking in STS and empirical studies of ethics

Gieryn (1999) argued that the analytical benefit of maps as a conceptual metaphor for representations is that they elicit the relationships of inclusion-exclusion and proximity-distance drawn by discourses (such as the boundary between science and non-science). Although they do not explicitly draw on the mapping metaphor, some STS studies on ethics explore similar relations and boundaries that outline 'the ethical domain' in technoscientific contexts. Wainwright et al. (2006) found that stem cell researchers included ethics in their constructions of proper science rather than draw boundaries between

them. Frith et al. (2011) juxtaposed the ethical areas of a fertility clinic's practice and areas seen as non-ethical, lacking ethical issues because they were solved by institutional procedures. Wehrens et al. (2023) found that stakeholders in data-driven healthcare tended to discursively construct ethics as a demarcated and minor sphere to be addressed separately rather than be integrated into technology development.

Sigl et al. (2020) invoked a geographic metaphor of proximity and distance to analyse how life scientists envision their responsibilities. Theorising scientists' narratives as 'geographies of responsibility', they explored the positions of different responsibilities in them. These geographies placed scientific excellence at the core and societal responsibilities (e.g., public communication and considering the societal relevance of research) farther from the scientists' remit. Furthermore, drawing on Felt's conceptual work (Felt, 2017), Sigl et al. identify the shared reference points in the scientists' narratives, the linear model of innovation, and institutional conditions as the narrative infrastructure supporting these representations. As a narrative infrastructure, these shared notions underscore more situated conceptions of responsibilities.

While these studies use spatial concepts to study representations, a rich body of research highlights that ethics is shaped by sociomaterial sites and contexts. In her ethnographic work, Shilton (2013) noted that certain practices act as 'value levers', bringing to the surface sociopolitical questions that were usually tacit in software development practice. Shklovski and Némethy (2023) distinguish between two types of sites where engineers engage with ethics: 'nodes of certainty', which are settled and familiar procedures such as checklists, facilitate everyday action, and 'spaces for doubt' and reflection. Goñi et al. (2024) problematise the tendency of ethics toolkits to adopt reading guides or checklists, arguing that these formats are more conducive to narrowing and routinising ethical thinking than to continuous reflection.

These discussions highlight the importance of considering understandings of ethics as situated, a notion championed by feminist STS. Building on Haraway's concept of situated knowledge, Suchman (1995, 2002) argued that representations of work and human-technology interactions are shaped by local perspectives, practices, and power dynamics. These representations make certain things (types of work, connections, influences) salient and obscure others, making them invisible. Widder and Nafus (2023) draw on Suchman's work to examine how responsibility is approached in software development. They find that the industry conceives of the software development process as a supply chain from general-purpose technology to end-use software, comprised of isolated modules, and that ethical issues are viewed differently across this supply chain. Issues that are notable downstream are invisible upstream, and vice versa. At the same time, the boundaries drawn by modularity separate

technologists from ethical issues arising from decisions made in other modules. As modularity is both an imaginary of the ideal work process, which can be seen as a narrative infrastructure, and the basis for organisation of labour that creates situated perspectives, this study emphasises that both aspects shape actors' relationship with ethics.

The metaphor of constellations

Consolidating these insights, I propose using a different kind of map as a conceptual metaphor for examining ML practitioners' understandings of ethics: not maps of countries (Gieryn, 1999), but maps of constellations in the night sky. Constellations are sets of stars that *appear* to be grouped together. The night sky is full of disparate objects, and the constellations metaphor emphasises that interpretation makes meaningful shapes out of these disparate points, connecting some objects and not others. Similarly, the narrative approach (Felt, 2017; Sigl et al., 2020) works with narrative points, i.e., meanings, experiences. Cultural concepts we use, like 'science' or 'ethics', group disparate meanings into constellations, 'things' we see, and coherent narratives. These constellations thus help us make sense of the world around us, and we use them to navigate our everyday practice. These constellations also draw on broader cultural meanings, similar to how mythological creatures or familiar shapes inspire the constellations we draw from stars – they function as narrative infrastructures. However, these 'things', their names, and what they include and exclude could be something else: what is seen as part of 'ethics' could have be attributed to 'quality' or 'rigor'. Moreover, STS continuously reminds us that categories and representations amplify some things and make others invisible, especially as they become naturalised (Bowker & Star, 1999; Suchman, 1995). Similarly, not all stars shine equally bright, but those that are not included in constellations also fade into the background by this exclusion, into an 'other' category (Bowker & Star, 1999). Moreover, the night sky and the constellations look different from different vantage points, depending on our location in space and time.

This conceptual metaphor sensitises my analytical gaze to consider the constellations that form ML practitioners' perceptions of ethics: what ideas and practices are included in them, and where in their professional location they see them. These representations are situated, yet also (somewhat) shared as they draw on shared concepts as their narrative infrastructures.

Methodology

The main source of data for this paper is 44 semi-structured interviews with ML practitioners conducted from February to August 2024. Ethical clearance for this study was obtained from the faculty's ethics committee (#2023-119-

K). During the analysis stage of the interview study, I added a secondary data source: submission requirements relevant to the interviewees' conferences, which include ethical considerations. These conference requirements often came up in the interviews, so I set out to further explore their role in practitioners' perceptions of ethics.

Interview study

The research design of the interview study aimed to capture the diverse types and contexts of machine learning work, representing a community of practice (Bowker & Star, 1999) spread across field and institutional lines. Recruitment, combining cold invitations based on desk research and drawing on social networks, was limited to Denmark, the Netherlands, and the United Kingdom. This was not to compare national contexts but rather to balance the aim to achieve greater diversity in terms of interviewees' working contexts (especially in academia) with feasibility. This approach resulted in 44 interviews (see Table 1).

Table 1. Description of interview participants

Total	44
Gender	Men (28), women (16)
Country of employment	Denmark (18), the Netherlands (17), the United Kingdom (9)
Academic employment	PhD student, incl. industrial (12), postdoc (4), assistant professor (8), associate professor (2), professor (2)
Non-academic employment	Data scientist (6), research and development (8), business analyst (1), software engineer (1)
Places of non-academic employment	Medium or large company (7), small company (7), startup (1), independent foundation (1)
Current specialisation	Bioinformatics, bioengineering, biostatistics, biosignals (8) natural language processing (8), computer vision, including medical imaging (5), speech recognition (3), emotion recognition (2), privacy (3), fairness (3), reinforcement learning (2), causality (2), other[a] (8)

a. Includes interviewees whose work is spread across topics or not strongly associated with particular topics, which is not uncommon with data scientists (Avnoon, 2021).

The majority of interviewees worked in research and development in academia and/or industry, but some also deployed and maintained active systems. They worked with different machine learning methods, from deep learning to statistical machine learning, and with different types of data (e.g., signals, images). As the table illustrates, the interviewees worked in a variety of fields. The 28 academic interviewees came from 13 different universities; the non-academic interviewees worked in different organizations. The

interviewees came from different countries, and many had worked in other countries before, representing the international and highly mobile workforce of the field. The gender imbalance among the participants reflects that of the field.

Table 1 provides a simplified overview by allocating each participant to a single category to avoid double-counting. However, some participants had multiple roles (e.g., academic and industry) or specialisations. All interviewees with an academic role counted as academics.

The interviews (12 online, the rest in-person) lasted from 40 minutes to 1 hour 40 minutes, most around an hour. The interview guide (Appendix I) focused on practitioners' backgrounds, projects, the data and methods they worked with, and how they think about the value and impact of ML work before probing for notions of ethics. I asked the interviewees to share what comes to mind when I ask them to think about ethics in their area of work/ML/AI/data science and then asked about those notions in relation to their own work. Then, I probed to elicit examples of different things in their work that they associated with ethics. For example, I asked them to tell me about the last time they read, heard, or discussed something related to ethics. The interviews also covered the interviewees' experiences with regulations and institutional procedures, such as ethical approvals and data protection.

After the interviews were transcribed, they were coded in NVivo with a combination of granular open codes (Charmaz, 2006) and broad codes in line with the flexible coding approach (Deterding & Waters, 2021). Next, focused coding concentrated on the interviewees' understandings of ethics. These codes were compared to the rest of the narratives and initial memos outlining what ethics entailed for each participant. Here, the focus on a 'where' question and spatial thinking emerged as analytically productive to think about three representations of ethics in the data. Putting these findings in dialogue with literatures that use spatial concepts and metaphors resulted in the development of the constellations concept. The analysis is illustrated with quotes from participants.[2]

Secondary data

During focused coding (April 2025), I noticed that several interviewees referred to 'ethical' or 'impact' statements written for conference submissions (the main publication outlet in computing) when they described ethics. To consider how this procedure relates to the interviewees' notions of ethics, I collected and analysed the materials concerning the statements (submission instructions and explanations, ethical codes and guidelines) for the conferences mentioned by and relevant to the interviewees: the Conference on Neural Information Processing Systems (NeurIPS), the International Conference

on Learning Representations (ICLR), and the International Conference on Machine Learning (ICML), the Conference on Computer Vision and Pattern Recognition (CVPR), and the Responsible NLP Checklist adopted in natural language processing conferences. The materials were collected in April 2025; as the interviews were conducted in spring-summer 2024, I collected both years' materials for conferences whose websites had different pages and information for 2024 and 2025 (as submissions for 2025 started as early as autumn 2024, so preparation for 2025 could have been underway). Appendix IV provides a full overview of collected resources and reference codes to resources mentioned in the text.

Findings

This analysis explores how ethics fits into the interviewees' portrayals of their work. From the interviews, it was evident that ethics is not recognised as equally relevant to all types of ML work or all stages of the ML project. When asked what comes to mind when thinking about ethics in ML or to describe ethics in their own work, the interviewees often pointed to certain aspects of research and development or specific practices as the 'ethical part'. Many interviewees portrayed ethics as not being a prominent aspect of their work and offered various explanations for this.

By analysing patterns of where ethics is salient versus absent (explicitly or implicitly by being unmentioned) in portrayals of ML work and the meaning of ethics when evoked, I identify three main understandings: ethics of applications, ethics of data, and ethics as a research topic. To better understand them, I employ the metaphor of constellations. The next sections explore each constellation, identify their narrative infrastructures, and point to what is overlooked by these representations.

Constellation 1: applications as ethics of the whole topic

How the models are used, or machine learning *applications*, was a key theme in practitioners' narratives about the ethics of machine learning. Applications refer to problems and tasks that computational methods can be tasked with solving, often associated with particular "domains" of world activity and knowledge, such as health or finance (Ribes et al., 2019). When the interviewees invoked the notions of ethics and applications together, they gave examples of applications they considered unethical, applications they would not personally work on, or research they would not want to see published. For example, several computer vision researchers mentioned that they avoid topics associated with surveillance, such as re-identification of people or anomaly detection in video data.

However, as representations of ethics in one's own work, applications were most relevant not to practitioners who immediately deployed active machine learning systems but to researchers. For example, an information retrieval researcher explained that while ethics is always a part of one's work, 'the difficulty in my research is mostly the application'. Given that he mostly worked with theory and simulations, what he meant by application could not be implementing ML systems in practice. As he explained, the ranking methods he developed (and published as a research output) could be used by web companies to build recommendation systems, and 'that doesn't have to be unethical, but it can definitely be unethical, and we don't really know what they do, right?' These potential applications of his methods thus represent the ethics of his own work. Similarly, for a geometric deep learning scientist, potential future applications reflect on how he sees the ethics of his present ML work:

> Facial key point detection <…> that's an interesting problem in itself. But if I see what happens after that – it's profiling, or emotion detection <…> I cannot really imagine anything but bad scenarios or commercial elements, like filters for Instagram or avatars. And these applications I find stupid <…>. So, I want to stay away from that and focus on the AI for science or medical imaging domains, because I think that's ethically a somewhat safe area. And if you make an impact there, I imagine that can be useful for humanity in the long run.

The range of potential applications of ML methods thus makes up an 'ethically safe' area for the researcher to work in. This distinct meaning of ethics of applications – the potential *downstream* applications – when it comes to what practitioners see as ethics in their work, is what characterises the constellation.

Widder and Nafus (2023) found that actors' perceptions of the potential (mis)use of AI technologies depended on the technology's place in the AI development supply chain. The supply chain runs from basic to applied technology: from general-purpose methods upstream, to partial tools and systems in the middle, and end-use software downstream. The closer technology is to end-use, the narrower the range of potential (mis)uses (p. 2). However, my interviewees' concerns about the misuse of their work did not fully resemble this spectrum. Indeed, the perceptions of ethics in their work among practitioners developing and implementing end-use algorithms were less focused on applications, as they were settled by the specialisation of their company or team. At the other end of the spectrum, several interviewees invoked the foundational nature of their research to explain why it did not involve ethical issues while acknowledging that ML/AI and technology in general can be used for different purposes. So, at the ends of the spectrum, concern for applications and misuses is less salient.

It is in the middle of this spectrum, in research oriented towards applications but not implementation, that future applications can be seen as a range: more concrete than all-encompassing possibilities of basic methods yet not limited to the one intended application. Here, consideration for downstream misuse represents the ethical aspect of one's machine learning work, as exemplified by an emotion recognition researcher's concern:

> Some of <…> the applications we are working with <…> will probably be applied in clinical settings. We'll rethink some of these models if we even want to open-source them <…> A lot of research today works not within the context, close to a clinical setting, but just working to open-source a model that can look at emotions from faces. And that very soon becomes surveillance technology and biometrics and stuff. Question is … okay, you are interested in doing that <…> but how do you address the ethical implications of these things?

Notably, the ethical implications of the model here refer to its potential misuse of the model for surveillance or biometrics rather than its intended clinical application. The former are the meanings and concerns the constellation foregrounds, and the focus on them makes the latter fade into the background.

This quote also points to actions and practices in ML work associated with doing ethics as applications, i.e., publication of results and (not) open-sourcing the model. Overall, this constellation positions doing ethics at the 'bookends' of ML projects. Before starting a project, practitioners determine the intended application or application area by choosing the topic, employer, or task. Interviewees across the spectrum commonly associated choosing work with ethics, which is consistent with previous studies (Avnoon et al., 2024; Slota et al., 2023). Here, the interviewees also invoked ethics as applications in connection to the end-product and publication of results: release of artifacts (data and models), licensing, reflections on the ethical implications of the research in papers, and the peer review process where work on problematic applications can be flagged and potentially rejected.

However, based on interview data and the analysis of the ethical statements' procedure, I argue that statements of ethical implications are not just a practice associated with ethics within this constellation; they supply notions that shape this representation of ethics in researchers' work. The connection to ethical statements also reveals another point of this representation of ethics.

Ethics statements as narrative infrastructures

Ethics statements (also known as statements of 'broader impact' or 'potential negative societal impacts') as part of submission requirements were first introduced at the NeurIPS conference in 2020 (Liu et al., 2022) and spread to

other major ML conferences and related fields, which often refer to NeurIPS instructions as exemplary. The interviewees often associated this practice with ML ethics, bringing it up not only as an example of ethics-related activities but also as an indicator of the field's awareness of ethical issues. For example, a professor mentioned them in response to a question about whether ethics is discussed at conferences:

> It's funny you're asking, because for the machine learning conferences, when you submit a paper, you also have to make an ethics statement. So, you have to say: okay, I did this algorithm, what kind of ethical implications? <…> So, you have to think a little bit about: what can it be used for? Like if you do better image classification, then it can be used for surveillance. If you do better generative modelling, then it can be used for fake news. Stuff like that. It's kind of a pressure from the community, on the community, to think about these things.

This description of what researchers are invited to write in ethical statements closely resembles the narrative points in this constellation: applications and the focus on downstream (mis)use as the main ethical consideration. Moreover, a previous study of ethical statements from one of the leading conferences, NeurIPS, highlighted the statements' focus on inappropriate applications: incorrect use and misuse by ill-intentioned users (Liu et al., 2022).

As I compared the interview narratives with the instructions for ethical statements provided to submitting authors by the conferences, I also found that the ACL Checklist Guidelines (ACL), CVPR Ethics Guidelines (CVPR-24a; CVPR-25a), and the NeurIPS Paper Checklist Guidelines (NeurIPS-25b) encourage authors to reflect on and acknowledge potential misuse. The 2025 NeurIPS Paper Checklist Guidelines (NeurIPS-25b) include these suggestions for considering broader impacts of submitted research:

> Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
>
> We expect many papers to be foundational research and not tied to particular applications, let alone deployments. However, if you see a direct path to any negative applications, you should point it out. For example, if you improve the quality of generative models, you might point out that your approach can be used to generate Deepfakes for disinformation. On the other hand, if you develop a generic algorithm for optimising neural networks, you do not need to mention that this could enable people to train models that generate Deepfakes faster.

The depiction of the potential negative impacts in this excerpt resembles the above-described representation of ethics: from focusing on misuse

possibilities to the exclusion of foundational research that attributes this consideration to the same type of ML work (applications-oriented research) as the group of interviewees who highlighted misuse as the ethical issue in their own work.

ICLR and ICML give less detailed instructions for ethics/impact statements and feature misuse less. The ICLR 2024-2025 Author Guides mention a variety of topics, including 'potentially harmful insights, methodologies and applications' (ICLR-24, ICLR-25a). The ICML Calls for Papers suggest authors state 'ethical aspects and future societal consequences' (ICML-24b, ICML-25a). However, the post from the Program Chairs introducing the practice in 2024 also gives examples of potential negative uses of ML methods and applications and distinguishes between submissions that do/do not touch directly upon potential negative applications and thus can use a provided generic statement (ICML-24a). Notably, conference ethics policies and submission instructions include broader considerations; even the statements of potential negative impacts description quoted above mention other themes like privacy and security. However, potential (mis)uses are a big theme in the descriptions of impacts/ethics statements, especially when the instructions are more detailed. Together with the interviewees' descriptions of ethics statements and the prevalence of misuse concerns among researchers who attend these conferences, this suggests that the constellation is underscored by the discourse of ethics statements.

Moreover, the connection to ethical statements brings out another distinguishing point of this constellation: its focus is not on individual projects or specific details within them or on ML technology in general but rather on certain *topics or subfields* of ML. Both the NeurIPS instructions and the interviewee quoted above discuss ethical implications of *types* of models: generative models can be used for deepfakes or fake news, classification models for surveillance. While the conference instructions generally suggested researchers discuss the potential (mis)uses and impacts of their approaches, they also gave examples of connections between types of models and potential issues. Moreover, according to Liu et al. (2022), NeurIPS impact statements often posited that models within the same topic or task have the same implications, thus attributing concerns about misuse to research fields or model classes rather than individual projects. This locus of ethics as applications was most explicitly articulated by a computer vision researcher:

> For this potential negative impact in the paper <...>, there are different types. There's the bias and the small ethical considerations. I think there are also ethical considerations of the whole topic you're doing. For instance, there's a topic called person re-identification, which I never want to work on because the main point of this is so you can re-ID people across surveillance cameras. So,

> there's an initial obvious thing: which problem are you doing. So, you at least pick a project that can be used for good in some way. But then the smaller – well, they're not smaller, but the considerations within that task – end up being an afterthought.

This quote illustrates how, in the context of impact statements, ethics are considered for the 'whole topic' rather than the decisions within the project. Therefore, more granular issues within the project fade into the background and are less noticeable. Evaluating the ethics of applications based on the merit of the whole topic also explains why medical or clinical applications are invoked in the quotes above as non-problematic, 'ethically safe'.

This distinct representation of ethics as applications is defined by the combination of its narrative points. The imagined downstream (mis)use as the main ethical consideration in one's work is intertwined with the 'whole topic' locus of ethics and its relevance in the middle of the basic-applied spectrum. This representation of ethics leaves other meanings and practices (like smaller considerations of intended applications) in the background. This constellation draws on the categories of methods and applications, the distinction between the two (Malazita & and Resetar, 2019; Ribes et al., 2019), and ethics/impact statements written for conferences as parts of its narrative infrastructure (Felt, 2017; Sigl et al., 2020). Ethical statements draw together the main points characterising this constellation in their instructions, constructing what ethics in machine learning is about. As researchers articulate ethical implications for their work when they submit to conferences, this representation is reproduced and shapes how practitioners understand ethics within and outside this context.

This representation of ethics acknowledges the risks of machine learning technology, portraying it as a double-edged sword rather than an unambiguously positive tool of progress. However, it entails that while the most problematic topics associated with negative applications (such as re-identification) are to be avoided as 'the ethics of the whole topic', the mere existence of potential negative uses does not delegitimise most topics. Rather, these potential misuses are to be articulated as the ethical implications of ML work in the ethical statements. As such, this constellation exemplifies the tendency that Moore (2019) has characterised as 'AI for not bad' rather than 'AI for good'.

Constellation 2: data work as ethics in practice

Data were also a salient theme in how the interviewees understood ethics to fit into machine learning development in general as well as in their own work. Similarly to applications, the ethics of data was not considered relevant to everyone doing machine learning; the types of data (not) involved in ML work were invoked to explain how big a part ethics played in it. For instance, a

business analyst working in a pharmaceutical company emphasised that working 'with real-time, real patient data' meant having to be extra careful. A PhD student contrasted her previous job in content moderation with her PhD research focused on developing methods. In the previous job,

> [ethics] was very very very talked about and important all the time, because of the type of work that we were doing. But <...> the only thing that I will think of in my day-to-day PhD life, other than high-level discussions about what LLM should do, is maybe stuff about equal opportunities and gender ratios in academia and stuff like that.

The reason for the relative absence of ethics in PhD research was that it was 'high-level and not using any real data'. Both the application and data notions of ethics contribute to this representation as foundational research that does not use any real data, the only considerations about ethics she encounters relate not to her project or topic but to science and technology in general. Several interviewees stated that their work with publicly available benchmarks does not entail ethical issues related to data.

However, what characterises this constellation when ethics as data is relevant to practitioners is its narrative connection to practice. The ethics of data was not talked about as a consideration to keep in mind or be aware of; the interviewees invoked it in connection to specific experiences and actions in their work, indicating it was something they addressed. An industrial NLP researcher mentioned a project where they tested large language models' responses to problematic user instructions:

> So, one very clear thing for me is, especially because working with annotators, how important they are and how important it is to support them. <...> For instance, in the safety project <...> it's not really easy to write these really dark examples. So, it was very important to provide mental health support, and there were rules that they should not spend multiple hours a day on this. <...> So, when I think about ethics, I think ethically collecting data is really part of that.

For some of them, ethics was an obvious part of data generation or processing. A few interviewees drew parallels between my study where they were the subjects and their own research. As a computer vision researcher explained, 'because I have to deal with human data, I take care of whatever you take care of'. A PhD student researching emotion recognition, who was in the middle of producing a dataset for his project, described that:

> When you are doing data collection, there are a lot of problems and ethical considerations. I wasn't aware of that. When you do data collection, the power dynamic suddenly changes. You have a lot of power over your participant.

He explained that data collection with physiological sensors might feel strange, but the power dynamic could force participants to continue even if they were uncomfortable. To mitigate that, he needed to reassure participants that they could stop at any time and make the process easier with small talk and breaks. He also expressed that participants need to feel that their contributions are valued. So, when one participant's measurements were difficult to calibrate, meaning the data would probably be unusable, he decided to proceed as normal so as not to make them feel that something was wrong with them. In addition to such accounts of 'human subjects' research ethics (Metcalf & Crawford, 2016), interviewees connected their other data-related practices, such as gaining access, preprocessing, and storing, to ethics.

This focus on practice and something practitioners actively did is also reflected in who brought up ethics as data: not just interviewees whose work involved human or sensitive data, but those whose jobs involved a lot of data handling. Those practitioners developed new datasets and personally carried out data acquisition and wrangling rather than using pre-processed, anonymised, and otherwise 'ready' data from collaborators, subcontractors, or clients. Thus, ethics in this constellation is seen in ML practitioners' data-related *practices*, as illustrated by a data scientist in adtech:

> Firstly, we care about not using underage users' data, because it just feels quite wrong. And another big one is figuring out information that will identify a person from the post. <…> But the general topic of whether it is ethical to use this type of data or in what ways we are using it, which could be right or wrong, is not really touched upon. So, it's a more granular level of, you know, specifically that post: is it identifiable, is it not? Can we use it from that user? Can we not use it from that user? <…> And once you aggregate the data, it's not as ethically wrong because you talk about the group rather than a specific person. And that's one of the big things, this personally identifiable information.

The locus of ethics as data is the granular decisions of navigating whether data is identifiable or not and arriving at an ethically acceptable state through data wrangling. Meanwhile, the more 'general topic' questions are in the background as they are not incorporated into this constellation of ethics. Indeed, most interviewees' descriptions of ethics concerning data in their own work were granular. Interviewees rarely raised broader questions about data, e.g. what kinds of data are more or less acceptable to use. An example of these less granular questions was the doubt one interviewee (the computer vision researcher who discussed the ethics of the 'whole topic') raised about whether using data collected from the internet (a type/bigger category) was ethical.

The quote also illustrates the connection of this representation of ethics to institutional frameworks, such as personally identifiable information. Many of the interviewees working with personal data of different kinds explained

that gaining access, storing, and otherwise ensuring compliant treatment of data takes time and work. While not synonymous with the representation of ethics as data work, various institutional frameworks inform this constellation.

Institutional procedures as the narrative infrastructure

Various institutional requirements contribute to the data work in ML projects. Acquiring and using *some* types of data means dealing with rules and regulations; reconciling them with the epistemic needs of the project is necessary for being able to get the job done, as illustrated by a data scientist from a large healthcare company. When I asked about examples of something related to ethics in his own work, he replied:

> I think that all the conversations we have with quality assurance, for example, are ethics in a way. So, this happens frequently, right? <…> So, when we had a data sourcing part <…>, specifically this pseudonymisation phase. There were all these discussions on what we mean by pseudonymisation and also making sure that we cannot get the identity of the person based on the data that we receive. <…> for us, the only thing that we were interested in was the images. So, we solved the issues by saying we are interested only in the images and <…> the doctors' feedback on the images, so if there is a disease, at which stage, and so on.

As negotiating what data can be used legitimately is associated with ethics, ethics is seen as a salient part of the technologist's day-to-day work. Institutional requirements provide moments where the normativities related to generating and processing data are articulated: in acquiring access to data through ethical approvals or various bodies that perform the role of data stewards, or in consultations with compliance specialists in organisations (e.g., quality assurance, privacy, and security officers), as described in the quote. Also, conferences ask submitting authors to indicate whether their research used human data, participants, and/or annotators, and if so, to mention ethical approvals and consent and discuss data practices in more detail (ACL; CVPR-24b; CVPR25-b; ICLR-24; ICLR-25a; NeurIPS-25b). NeurIPS and ACL checklists also ask questions about compensation for workers involved in producing data (ACL; NeurIPS-25b). In that sense, datasets that fall outside these procedures, like public benchmarks, can be considered free of ethical issues in terms of the lack of associated data work (including compliance work) that represents ethics in this constellation. Moreover, all these articulations focus on the details of each project, corresponding to the granular locus of ethics as data.

Therefore, the various institutional procedures shape this constellation as its narrative infrastructure; their requirements focus on the details of projects and shape what handling different types of data necessitates, providing categories in which this work is done (e.g., personally identifiable, pseudonymisation). Depending on how this labour is organised, it can be a more or less significant part of ML practitioners' work and their representations of ethics. This connection to institutionalised, formal procedures might be the reason why data work here is recognised as ethics work, a finding that differs from previous studies (Barocas & boyd, 2017; Slota et al., 2023), which noted that data work is rarely openly acknowledged as such, even when it is value-laden and done with care. It echoes the finding from Samuel et al. (2021) that health AI researchers' understanding of research ethics is synonymous with data governance, but ethics in this constellation also evokes notions of care towards data and data subjects (Zakharova & Jarke, 2024), as described by the quoted PhD student.

Previous studies note that institutionalised procedures can create a feeling that ethical considerations or responsibilities are taken care of, rendering ethics in practice invisible (Frith et al., 2011; Samuel et al., 2021; Wainwright et al., 2006). While this constellation makes the ethics of data visible to the interviewees, its granular, practice-oriented locus and reliance on established procedures exemplify a more feasible representation of ethics, something that can be 'solved', even if it takes a lot of time and work. In that aspect, data practices are 'nodes of certainty' rather than 'spaces of doubt' (Shklovski & Némethy, 2023). Thus, in addition to excluding data outside the purview of institutional requirements (though a few interviewees problematised those), the focus of ethics of data on smaller decisions and practices means that bigger questions about data generation and use also stay in the background, as they are not included in this constellation.

Constellation 3: ethics as the main research topic

While the first constellation focused on the ethics *of* the topic of ML work, ethics can also be *the* topic of machine learning research and development. One aspect of the 'ethical turn' in ML is the rise of research about it, also within the ML field (Zhang et al., 2022). Issues like bias, fairness, privacy, and explainability have developed into recognisable research (sub)fields, together forming a stream of work described as Responsible AI, Trustworthy AI, etc., and associated with spaces like the ACM Conference on Fairness, Accountability, and Transparency (FAccT).

Several interviewees (mainly, but not exclusively, researchers) worked on these topics, either as their main focus or within some of the multiple projects they were involved in. For them, the work *on* ethics-related topics often

represented ethics in their work. For others, not pursuing a topic associated with ethics was a narrative explanation for ethics not playing a notable part in their work. This connection between the topical focus of the work and the presence of ethics was expressly articulated by a couple of interviewees whose previous jobs or projects were related to ethics as a topic. When asked about ethics in their work, they said that ethics was a salient part of their professional lives in their previous jobs compared to now. A data scientist in adtech had previously worked in a non-profit organisation that focused on, among other topics, societal aspects of AI. She noted that in terms of taking or hearing something about ethics,

> I think it's been a long time. It's probably when I was working at the [organization] because that was the most relevant at that point. And maybe the last thing is not a paper, but I followed a presentation that was made by some people from the university. I think it was about a risky AI system, the type of risk that we could have with AI systems in general. It's been a long time.

Here, something about ethics means ethics-focused, and she portrays that version of ethics as absent from her current work due to topical irrelevance. Moreover, academic interviewees distinguished between specialised, ethics-focused papers and more typical papers that develop methods based on benchmark datasets, which do not address ethics issues like bias or explainability. For instance, a previously quoted computer vision researcher (who discussed the ethics of the 'whole topic') noted that

> It's good to do a paper that's trying to address the biases, but then this is a whole different paper <...>. because it's almost too much work to try, and it deserves its own publication on the topic rather than being like a small thing in a paper that you try to correct for.

In addition to invoking conceptions of the academic publication process and its pressures, the discussion of papers reflects the 'all or nothing' feature of ethics as a topic. Due to this constellation, ethics as a topic is only seen as a notable part of one's ML work if it is *the* topic, the specialisation of a project, a paper, or the researcher's subfield in general.

The specialisation narrative point also appears in how the interviewees described (not) hearing, talking, or learning about ethics. Reflecting the intensive publication and dissemination that the field is known for (Gallagher et al., 2024), keeping up with scientific and technological developments by going to conferences and events, reading papers, and following social media discussions was part of both researchers' and practitioners' jobs. In describing these practices, interviewees emphasised the focus on content relevant to their specialisation or concrete projects. As such, engaging with ethics-related literature and developments was not portrayed as an essential element of ML work,

unless this was the work's specialisation (e.g., focusing on fairness or privacy). For instance, some of the interviewees described attending or following only directly relevant presentations or posters, which did not include 'ethics-coded' topics. A few interviewees, however, described deliberately seeking ethics information, i.e., reading publications from FAccT, following experts like Timnit Gebru, or developments of the AI Act as a matter of personal interest.

Despite that, several interviewees, including those who did not work on ethics-coded topics, mentioned media items such as news or posts and shared cases of ML/AI ethical controversies when I asked them to share the last time they heard or read something about ethics and ML, a further indication that it was not absent from practitioners' experiences. How these encounters with ethics-related content fit into practitioners' representations of their work was unclear, until a PhD student working on ML for sleep replied that he only hears, reads, or talks about ethics 'when I'm forced, or if it pops up in my feed.' Several other practitioners portrayed ethics-associated content as something that just 'comes up' or something they 'happen to see' in traditional or social media. The fact that these experiences were brought up as examples of encounters with ethics signifies that the interviewees associated them with ethics to some extent. However, this narrative connection emerges only with probing, and even then, they are framed as non-intentional, indicating they are not a part of an ML practitioner's work in the way that reading papers and news about their immediate specialisation is. While it seems that these encounters are not that rare, they are in the background of practitioners' representations of ethics in their work, as they fall outside the constellation of ethics as a topic.

Beyond specialisation, I suggest that the reason is that these encounters with ethics take place in contexts and roles outside ML work. For example, the PhD student who stated that he only interacted with ethics content when 'forced' or when it 'popped up' in his feed mentioned multiple topics concerning AI and machine learning that deserve more discussion. He explicitly addressed two roles he has in relation to them:

> Also, do people consent to the data? That's a pretty big problem that hasn't been figured out because Facebook is publicly available, and you can just scrape Facebook, people didn't consent to it. Is it good, is it bad? I don't know, and for the most part … from the researcher part, I don't really care. I just want to make cool stuff. But I care that Facebook doesn't have my data, and I don't want them to misuse it.

This quote emphasizes the role of *social* location: what is salient for him as the data subject is less of an issue for a researcher. If practitioners see themselves engaging with these topics as social media users, data subjects, and social milieu, these encounters will not be part of ethics *in their work*. This separation

between the two roles echoes the described tendency of computational cultures to split the technical and social identity and thus distance sociopolitical concerns from technical work (Malazita & and Resetar, 2019).

In a few cases, the interviewees made a narrative link between personal and professional positions when discussing ethics. For instance, a software engineer brought up her sister being an artist when problematising for-profit large language models trained on internet data, noting that her sister's work could have also been used without consent or compensation. She linked this to her concerns about licensing issues within ML development, pointing out that some benchmark datasets have similarly been scraped off the internet and emphasising that a part of her organisation's approach that she feels strongly about is proper licensing. However, the interviewees' discussions of controversies concerning AI or other issues 'out there', which they probably encountered through (social) media, were mainly quite separate from narratives about their own work and ethics within it.

This constellation centres around the idea of ethics as the main focus of ML work. It is reflected in how technologists specialising in related topics see ethics in their work as work *on* ethics, and explains ethics' absence for others. In that, the specialisation logic functions as this constellation's narrative infrastructure, as practitioners who did not work on ethics as a topic either described unintentional exposure to ethics or framed their interest as a personal one, without relating it to their professional role. However, the many encounters with ethics through content that stay in the background also suggest that social location and roles contribute to representations of ethics in ML work.

Mapping ethics in ML practitioners' work

Overall, the three constellations of ethics (see Table 2 for an overview) organise how practitioners conceive of ethics as part of their work (including when they see that part as minimal). They outline a kind of map of ethics in ML work, based on ethics' salience in certain aspects and locations.

Table 2. Three constellations of ethics

Point	Applications	Data	Topic
Criterion of salience of ethics in one's work	potential applications	data used	ethics-related research topic (e.g., fairness, privacy)
Main representation of ethics in one's work	downstream (mis)use of a type of model	navigating data collection, wrangling, and data protection in practice	work on an ethics-related topic
Salient in …	applications-oriented research (middle of basic-applied spectrum)	extensive work on personal data collection or wrangling	specialised projects or subfields
What fades into the background	intended applications; current decisions relevant to future (mis)use	data outside governance purview; higher-scale questions re data use	encounters with sociopolitical aspects of ML through (social) media, personal relations, etc.
Elements of narrative infrastructure	Separation of methods and applications; ethical statements procedure	various institutional procedures	specialisation narrative; separation of technical and social identities

This overview also points to meanings and practices that are overlooked by these representations of ethics. In addition to those excluded within individual constellations, this map highlights that, among the interviewees, modelling (training and evaluation) was mostly not connected to ethics. This could be related to the abstraction of computing methods, which are connected to the real world by data and applications (Ribes et al., 2019). While absent from the representations of ethics, this technical work is likely to have tacit values and normativities (Slota et al., 2023).

Discussion

This paper set out to explore what we can learn about ethics as an actor's concept by studying how and where it appears in practitioners' narratives of ML work. To examine these representations, I employed the metaphor of constellations. Like constellations, the emic notions of ethics explored in this paper take distinct shapes, formed by the (narrative) points that comprise them. The metaphor has helped me illustrate these points and links between them in the constellations of ethics as applications, data, and a research topic, demonstrating their differences not only in the considerations they put forward but also in the locations in ML practice where they are relevant. It also facilitated thinking about how the notions of ethics include certain meanings and

practices while leaving others out, thus rendering them in the background of the map: maybe not fully invisible but less noticeable next to the recognisable figure.

Constellations also illustrate that representations of ethics are both situated (intertwined with certain practices and types of ML work) and underscored by broader narratives. The shared points within the interview narratives of practitioners working across three countries, alongside the links to findings from previous studies (Liu et al., 2022; Samuel et al., 2021; Widder & Nafus, 2023), highlight the shared cultures and stable narratives that constitute narrative infrastructures (Felt, 2017) within this global community of practice. Moreover, the study's findings demonstrate that institutional, formalised procedures contribute to practitioners' understandings of ethics, also fostering similarities and differences between situated representations. Conferences configure narratives for a global community they bring together, while research ethics and data governance regulations are both spread across geographic contexts and vary among them. At the same time, the findings highlight how one's notion of ethics in their work is shaped by local contexts: for example, the distribution of labour regarding data work. The interviewees' narratives highlight that practitioners' relationships with ethics can change from project to project, based on the project's basic versus applied orientation, topic, the data used, organisation of labour, etc.

This study's findings also contribute to discussions about ethics in practice in ML/AI development. The feminist approach to accountability developed by Suchman (2002) is based on making networked relations within sociotechnical systems visible and identifying one's position and responsibility within them. The practice of reflecting on potential downstream (mis)uses of ML models, promoted by conferences, draws such connections between methods and potential future applications. By doing so, it characterises a bigger part of the basic-applied spectrum as having impacts and ethical considerations. Still, it draws on narratives of technological neutrality, ascribing ethical valence to applications rather than methods and leaving out basic research. Moreover, the representations of ethics of applications highlight that articulating the connections and potential harms does not necessarily mean that they are addressed. While avoiding applications can be seen as a form of refusal from within, it seems to apply to a limited set of explicitly problematic purposes. For others, potential negative applications of published research and open-source code were seen as something that can be acknowledged but not really prevented. Without active refusal movements and mitigation strategies, statements of potential negative impacts can embody the 'non-disruptive reflection' which is often the extent of efforts in engagement with societal implications (Sigl et al., 2020, p. 1570). For deeper integration of these considerations

into research and development, the narrative of technological neutrality (or double-edged sword-ness) needs to be grappled with. Moreover, these approaches need to reconcile with values regarding reproducibility and open-sourcing of code and data (Widder et al., 2022).

Slota et al. (2023) note that AI ethics tends to be invisible work and argue for explicitly recognising and valorising it. The conclusion that institutional conditions make ethics a more visible part of ML work can be seen as a positive development from this perspective. However, the limitation of my approach and the constellation metaphor, the way it is employed here, is that it does not attend to the unequal (e)valuation of different kinds of work. While visibility and worth often go hand in hand, in the context of hierarchies noted by previous studies, the association of ethics with institutional requirements might be more ambivalent. Widder (2024) showed that association with admin tasks places the work of ethics lower in the hierarchy of tasks within ML work. Thus, narrative connections to compliance or data work, another underappreciated part of AI development (Bennett et al., 2025; Slota, Fleischmann, et al., 2020), might also hinder the position of ethics. These concerns add to the call for interventions with the hierarchies of (e)valuation within the field and specifically valorising the intricacies of data work as a way towards responsible AI (Bennett et al., 2025). Moreover, the connection to compliance and the granular focus of ethics as data, representing 'nodes of certainty' in practitioners' work, actualise the need for spaces of doubt in which more structural and political issues could be addressed (Shklovski & Némethy, 2023). Otherwise, these interventions might repeat the mistakes of formal ethics in other sectors: 'defin[ing] problems narrowly and then divert[ing] scarce attention and resources to compliance work, making dissent difficult' (Heimer, 2013).

Widder and Nafus (2023) suggest that widely publicised media controversies or regulations create shared social locations for such bigger questions, outside of practitioners' immediate positions in organisations and value chains. The findings on practitioners' encounters with ethics as a topic add more nuance to the arguments for acknowledging the role of locations outside the workplace for addressing societal aspects of technoscience. First, they emphasise that these locations are social and can also entail different roles, identities, and relations to others. This paper suggests that in addition to encouraging AI ethics work coming from lived experience (Widder, 2024), it might be productive not only to engage with scientists and engineers as technologists but also to bring out their other social roles and identities: data subjects, citizens, and relations thereof. At the same time, the paper opens the question of whether socially situated engagements with ethics translate into the professional realm, especially considering the separation of the technical and social

identity in the computer science epistemic culture (Malazita & and Resetar, 2019).

Notably, this paper focuses on narratives and the meanings, concerns, and practices that practitioners associate with 'ethics'. The 'ethics' category and discourses ground the partial perspective on the topic produced by this study (Suchman, 1995). As such, its limitation is that it does not reflect meanings and practices that are not connoted by this concept: those falling under a different actor category or too invisible to be voiced in interview narratives, such as activities excluded from the 'work' of AI that Forsythe (2001) noted through ethnographic observation.

With this focus, the paper highlights the existence of multiple ontologies of ethics that can be studied. Previous work has juxtaposed ethics as the formal pursuit of specific value commitments versus an empirical field and category (Frahm & Schiølin, 2024) and 'formal ethics' versus 'ethics on the ground' (Heimer, 2013). In addition to them, we might need to distinguish between meanings and practices connected to the language of 'ethics' (Hoeyer & Tutton, 2005) versus tacit normativities of work not bound by a certain concept but aiming for 'good' work in different ways (Pols, 2015). These two distinctions might overlap, but they are qualitatively different. Actors' categories can also draw on the language of ethics, especially as various formal ethics efforts make that discourse more present in everyday practice. Moreover, the latter distinction emphasises not only the specific language but the tendency of categories and representations, including actors' portrayals of their own work, to amplify some aspects of practice while making others invisible (Bowker & Star, 1999; Suchman, 1995). This paper has highlighted that practitioners' notions of 'ethics' are quite distinct and shaped by both practices and broader narratives, suggesting that they are narrower than the values and normativities involved in their work. As representations of technoscience can be reproduced through feedback loops between meanings and practice (Bennett et al., 2025), understanding the relationship between what easily comes to mind among practitioners and what remains tacit might be important. Connecting and translating between these ontologies necessitates careful consideration in both research and intervention, especially considering that language contributes to the persistent gap between technologists and social science and STS scholars in responsible research and technology efforts (Barocas & boyd, 2017; Moats & Seaver, 2019).

Notes

1. Historical accounts emphasise that these terms and the relationship between them have a complex history (Cardon et al., 2018; Olazaran, 1996). Currently, the debate about the actuality and performative functions of these terminologies

is itself productive. However, here, I bracket the constructions of machine learning and artificial intelligence to direct my analytical sensibilities elsewhere. For this study, the overlap between these notions is significant enough (with current constructions of artificial intelligence entering machine learning as a dominant technique) to not warrant a distinction.

2. Here and throughout, quotes are lightly edited for readability without affecting their meaning. <…> indicates that text was shortened for focus and space limitations.

References

Andrejevic, M. (2020). Data civics: A response to the "ethical turn". *Television & New Media*, *21*(6), 562-567. https://doi.org/10.1177/1527476420919693

Avnoon, N. (2021). Data scientists' identity work: Omnivorous symbolic boundaries in skills acquisition. *Work, Employment and Society*, *35*(2), 332-349. https://doi.org/10.1177/0950017020977306

Avnoon, N., Kotliar, D. M., & Rivnai-Bahir, S. (2024). Contextualizing the ethics of algorithms: A socio-professional approach. *New Media & Society*, *26*(10), 5962-5982. https://doi.org/10.1177/14614448221145728

Barocas, S., & boyd, d. (2017). Engaging the ethics of data science in practice. *Communications of the ACM*, *60*(11), 23-25. https://doi.org/https://doi.org/10.1145/3144172

Bennett, S. J., Catanzariti, B., & Tollon, F. (2025). "Everybody knows what a pothole is": Representations of work and intelligence in AI practice and governance. *AI & Society*, *40*, 3283–3294. https://doi.org/https://doi.org/10.1007/s00146-024-02162-0

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press. https://doi.org/https://doi.org/10.7551/mitpress/6352.001.0001

Cardon, D., Cointet, J.-P., Mazières, A., & Carey-Libbrecht, L. (2018). Neurons spike back. *Réseaux*, *211*(5), 173-220. https://doi.org/https://doi.org/10.3917/res.211.0173

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE.

Deterding, N. M., & Waters, M. C. (2021). Flexible coding of in-depth interviews: A twenty-first-century approach. *Sociological Methods & Research*, *50*(2), 708-739. https://doi.org/10.1177/0049124118799377

Felt, U. (2017). "Response-able practices" or "new bureaucracies of virtue": The challenges of making RRI work in academic environments. In L. Asveld, R. van Dam-Mieras, T. Swierstra, S. Lavrijssen, K. Linse, & J. van den Hoven (Eds.), *Responsible Innovation 3: A European Agenda?* (pp. 49-68). Springer International Publishing. https://doi.org/10.1007/978-3-319-64834-7_4

Forsythe, D. E. (2001). *Studying those who study us: An anthropologist in the world of artificial intelligence*. Stanford University Press. https://doi.org/10.1515/9781503619371

Frahm, N., & Schiølin, K. (2024). The rise of tech ethics: Approaches, critique, and future pathways. *Science and Engineering Ethics*, *30*(5), 45. https://doi.org/10.1007/s11948-024-00510-3

Frith, L., Jacoby, A., & Gabbay, M. (2011). Ethical boundary-work in the infertility clinic. *Sociology of Health & Illness*, *33*(4), 570-585. https://doi.org/https://doi.org/10.1111/j.1467-9566.2010.01308.x

Gallagher, J. R., Avgoustopoulos, R. E., Hamilton, A., & Seilkhanova, T. (2024). The challenges and writing practices of communicating artificial intelligence and machine learning in an era of hype. *AI & Society*, *40*, 2895-2908. https://doi.org/10.1007/s00146-024-02049-0

Gieryn, T. F. (1999). *Cultural boundaries of science: Credibility on the line*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226824420

Goñi, J. I., Eugenia, R., Jesus, P. M., Martín, I., & and Millán, M. J. (2024). Tooling with ethics in technology: A scoping review of responsible research and innovation tools. *Journal of Responsible Innovation*, *11*(1). https://doi.org/10.1080/23299460.2024.2360228

Greene, D., Hoffmann, A. L., & 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*. https://doi.org/https://doi.org/10.24251/HICSS.2019.258

Heimer, C. A. (2013). 'Wicked'ethics: Compliance work and the practice of ethics in HIV research. *Social Science & Medicine*, *98*, 371-378. https://doi.org/https://doi.org/10.1016/j.socscimed.2012.10.030

Hoeyer, K. L., & Tutton, R. (2005). 'Ethics was here': Studying the language-games of ethics in the case of UK Biobank. *Critical Public Health*, *15*(4), 385-397. https://doi.org/10.1080/09581590500523533

Kerr, A., Barry, M., & Kelleher, J. D. (2020). Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720915939

Krarup, T., & Horst, M. (2023). European artificial intelligence policy as digital single market making. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231153811

Liu, D., Nanayakkara, P., Sakha, S. A., Abuhamad, G., Blodgett, S. L., Diakopoulos, N., Hullman, J. R., & Eliassi-Rad, T. (2022). Examining responsibility and deliberation in AI impact statements and ethics reviews. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 424–435. https://doi.org/https://doi.org/10.1145/3514094.3534155

Malazita, J. W., & and Resetar, K. (2019). Infrastructures of abstraction: How computer science education produces anti-political subjects. *Digital Creativity*, *30*(4), 300-312. https://doi.org/10.1080/14626268.2019.1682616

Metcalf, J., & Crawford, K. (2016). Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society*, *3*(1). https://doi.org/https://doi.org/10.1177/2053951716650211

Moats, D., & Seaver, N. (2019). “You Social Scientists Love Mind Games”: Experimenting in the “divide” between data science and critical algorithm studies. *Big Data & Society*, *6*(1). https://doi.org/10.1177/2053951719833404
Moore, J. (2019). AI for Not Bad. *Frontiers in Big Data*, *2*. https://doi.org/10.3389/fdata.2019.00032
Olazaran, M. (1996). A sociological study of the official history of the Perceptrons controversy. *Social Studies of Science*, *26*(3), 611-659. https://doi.org/10.1177/030631296026003005
Pols, J. (2015). Towards an empirical ethics in care: Relations with technologies in health care. *Medicine, Health Care and Philosophy*, *18*(1), 81-90. https://doi.org/https://doi.org/10.1007/s11019-014-9582-9
Ribes, D., Hoffman, A. S., Slota, S. C., & Bowker, G. C. (2019). The logic of domains. *Social Studies of Science*, *49*(3), 281-309. https://doi.org/10.1177/0306312719849709
Samuel, G., Chubb, J., & Derrick, G. (2021). Boundaries between research ethics and ethical research use in artificial intelligence health research. *Journal of Empirical Research on Human Research Ethics*, *16*(3), 325-337. https://doi.org/10.1177/15562646211002744
Shklovski, I., & Némethy, C. (2023). Nodes of certainty and spaces for doubt in AI ethics for engineers. *Information, Communication & Society*, *26*(1), 37-53. https://doi.org/10.1080/1369118X.2021.2014547
Sigl, L., Felt, U., & Fochler, M. (2020). “I am primarily paid for publishing…”: The narrative framing of societal responsibilities in academic life science research. *Science and Engineering Ethics*, *26*(3), 1569-1593. https://doi.org/10.1007/s11948-020-00191-8
Slota, S. C., Fleischmann, K. R., Greenberg, S., Verma, N., Cummings, B., Li, L., & Shenefiel, C. (2020). Good systems, bad data?: Interpretations of AI hype and failures. *Proceedings of the Association for Information Science and Technology*, *57*(1), e275. https://doi.org/https://doi.org/10.1002/pra2.275
Slota, S. C., Fleischmann, K. R., Greenberg, S., Verma, N., Cummings, B., Li, L., & Shenefiel, C. (2023). Locating the work of artificial intelligence ethics. *Journal of the Association for Information Science and Technology*, *74*(3), 311-322. https://doi.org/https://doi.org/10.1002/asi.24638
Suchman, L. (1995). Making work visible. *Communications of the ACM*, *38*(9), 56-64. https://doi.org/https://dl.acm.org/doi/pdf/10.1145/223248.223263
Suchman, L. (2002). Located accountabilities in technology production. *Scandinavian Journal of Information Systems*, *14*(2), 7. https://doi.org/http://aisel.aisnet.org/sjis/vol14/iss2/7
Wainwright, S. P., Williams, C., Michael, M., Farsides, B., & Cribb, A. (2006). Ethical boundary-work in the embryonic stem cell laboratory. *Sociology of Health & Illness*, *28*(6), 732-748. https://doi.org/https://doi.org/10.1111/j.1467-9566.2006.00539.x
Wehrens, R., Stevens, M., Kostenzer, J., Weggelaar, A. M., & de Bont, A. (2023). Ethics as discursive work: The role of ethical framing in the promissory future

of data-driven healthcare technologies. *Science, Technology, & Human Values*, *48*(3), 606-634. https://doi.org/10.1177/01622439211053661
Widder, D. G. (2024). Epistemic power in AI ethics labor: Legitimizing located complaints. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1295-1304. https://doi.org/https://doi.org/10.1145/3630106.3658973
Widder, D. G., & Nafus, D. (2023). Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231177620
Widder, D. G., Nafus, D., Dabbish, L., & Herbsleb, J. (2022). Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2035–2046. https://doi.org/10.1145/3531146.3533779
Zakharova, I., & Jarke, J. (2024). Care-ful data studies: Or, what do we see, when we look at datafied societies through the lens of care? *Information, Communication & Society*, *27*(4), 651-664. https://doi.org/10.1080/1369118X.2024.2316758
Zhang, D., Maslej, N., Brynjolfsson, E., Etchemendy, J., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Sellitto, M., & Sakhaee, E. (2022). *The AI Index 2022 annual report*. S. I. f. H.-C. A. AI Index Steering Committee, Stanford University. https://hai.stanford.edu/assets/files/2022-ai-index-report_master.pdf

Chapter 7
Paper 3 – Where machine learning work meets biomedical data: friction in data settings

Abstract

Machine learning is increasingly used in biomedical research and healthcare. Based on interviews with Europe-based ML practitioners, this paper explores the challenges they experience in working with biomedical data through the lens of friction. By approaching friction with the relational thinking embraced by infrastructure scholarship, the paper demonstrates that friction generates between the various elements of the situated working relationships constructed in biomedical ML projects. These data settings include data governance procedures, infrastructures supporting clinical data sharing, data features, established practices, tacit assumptions of machine learning work, and more. The analysis reveals that data access is not merely an initial barrier to working with data but rather an ongoing aspect of the working relationship that shapes ML epistemic practices. The paper also highlights that friction is relative to the understanding of ML work and its interaction with biomedical knowledge.

Introduction

Both public and corporate discourses champion the intensive sourcing of data and its (re)use in computational modelling with machine learning (ML) or artificial intelligence (AI) methods as solutions to various societal and organisational problems. This includes the healthcare sector, where artificial intelligence-driven discovery and diagnostics, as well as automation of cumbersome clinical processes, are seen as potential solutions to human disease and limited healthcare resources.

However, studies have shown that the realities of these computational solutions are more complicated (Henriksen & Bechmann, 2020; Sambasivan et al., 2021). Data 'do not flow like oil' (Borgman, 2015). Rather, producing data requires hard work, even if that work is invisible (Hoeyer, 2023), and making data suitable for reuse in new contexts requires significant work of, first, decontextualisation and then recontextualisation for the new use context (Leonelli, 2013). In data science projects, datasets are rarely ready for modelling and require significant intervention that can take most of the project's time (Muller et al., 2019). Moreover, in high-stakes areas like healthcare, data practices in ML/AI projects can result in what Sambasivan et al. (2021) term data

cascades: negative effects at later stages of the projects that can result in project abandonment as well as harm to the user populations.

One conceptual register to analyse the 'partiality and bumpiness' of computational knowledge infrastructures is friction (Davies & Díaz-Valderrama, 2026, p. 345). Based on the physical metaphor, the concept of friction has been used to analyse the processes of interaction and transformation within data journeys that slow down and introduce risks into knowledge infrastructures (Edwards, 2010), or the 'sticky moments' between universalising projects like capitalism and science and local contexts (Tsing, 2005). Using this concept, Bonde et al. (2019) have shown that reusing clinical data for analytics and automation is more difficult than anticipated, and that these frictions can be cumbersome but also produce helpful insights. Understanding the nature of such frictions, as well as their detrimental versus generative qualities, is crucial in order to avoid data cascades in high-stakes contexts like healthcare.

In this paper, I draw on this conceptual lens to better understand ML practitioners' experiences of working with biomedical data. The paper is informed by an interview study in which biomedical data were often described as challenging to work with. Previous studies explore the data practices that make health AI doable (Henriksen & Bechmann, 2020), how ML specialists attend to data quality in public health projects (Thakkar et al., 2022), and the role data plays in the separation of the pristine and abstract world of machine learning and the messy world of clinical practice (Zilberstein, 2024). The paper advances this body of knowledge by addressing the following research question: What generates the frictions ML practitioners experience in working with biomedical data?

I combine the friction lens as adopted and used in critical data studies with relational and contextual thinking used in infrastructure studies (Star & Ruhleder, 1996). This framework emphasises that relations and practices in context make up an infrastructure and differentiates between the work and the infrastructure supporting this work. As Borgman (2015) argues, something becomes data when it is drawn on for future insight or argument. In the paper, I extend this reasoning to data friction. This helps me think about data not as discrete entities that flow around but as embedded in the many local contexts and relationships of their production and use (Loukissas, 2019), including relations produced when data, algorithms, and humans are entangled in machine learning practice (Amoore, 2020; Thylstrup et al., 2022).

Based on the analysis that puts this theoretical lens in dialogue with 25 interviews with ML practitioners based in Denmark, the Netherlands, and the UK, I argue that friction arises between uneasily fitting practices and contexts within situated working relationships. Thus, ML practitioners experience working with biomedical data as challenging, not due to the particular features

of these data or the collision between computational and 'domain' epistemic cultures, but because of the friction in relations with data that include different practices, technologies, notions, and assumptions. These insights advance our understanding of the role of data in machine learning and the conceptual discussion about friction.

The paper proceeds as follows. First, I explore the related but distinct conceptualisations of friction in data studies and extend this work by adding the notion of relationality from infrastructure studies (Star & Ruhleder, 1996). Next, I elaborate on the data and methods used in the study before moving to the findings regarding the friction in ML practitioners' relations with data in the dimensions of accessing and knowing. The final section reflects on the implications of the findings and the use of the friction lens in the study, with its affordances and limitations.

Friction in computational knowledge infrastructures

Paul Edwards (2010) introduced the concept of data friction as a part of the global knowledge infrastructures of climate science: when data move from one place or actor to another, or from one medium (e.g., audio recording to textual transcript) or material form to another (e.g., punch card to magnetic tape), friction slows data movement and introduces risks of losing or distorting the data. Edwards et al. (2011) further suggested 'science friction' as a conceptualisation of tensions that occur when two different epistemic cultures strive to use the same data and highlighted the repair work that allows such data reuses to proceed despite friction. Critical data studies have integrated this notion of friction, focused on data journeys, with friction as conceptualised by anthropologist Anna Tsing (2005). In her study of global connections across local differences, Tsing argued that global capitalism, scientific knowledge, and politics spread not only despite friction with local contexts but because of it. Thus, incorporating Tsing's notion of friction has afforded data studies to showcase the generative effects of friction (Bonde et al., 2019).

ML/AI presents a case where the notions of friction as between epistemic cultures (Edwards et al., 2011) and between general and particular (Tsing, 2005) converge. Computational relationalities operate by distinguishing general methods and particular domains of knowledge and practice where the abstract methods can be applied, infused with local context (Ribes et al., 2019). As such, it has been argued that these approaches universalise the world as constituted by a bunch of domains, all amenable to computational analysis and reorganisation (Lee & Ribes, 2025; Slota, Hoffman, et al., 2020). At the same time, this generalised ambition is built on the particular epistemic culture of ML/AI. The tension between this culture and 'domain' contexts has been explored in the case of friction arising between the ML classification

systems and journalistic principles in data annotation for automating news recommendation (Hartley & Thylstrup, 2025; Thylstrup & Hartley, 2025). Showcasing this 'domain friction', Hartley and Thylstrup argue that, against the grain of the visions of smooth application of ML/AI methods in specific domains, this adaptation happens through friction, resulting in new, hybrid approaches.

These studies illustrate the observation by Aula (2019) that the friction concept shines best against the normative presupposition of smooth data flows – or seamless domain application. Such assumptions and normativities are often shown (if not specifically highlighted) in the friction literature. For instance, Bates (2017) argues that data frictions can actually be thought of as safeguards rather than blockages in the case of surveillance capitalism. Minna Ruckenstein (2023) uses friction in a way that foregrounds affect; she explores frictions as annoyance and other emotions Finnish users experience when social media algorithms do not 'get' them, offering incompatible or annoying recommendations. Further, Ruckenstein and Lehtiniemi note that '[f]or technology professionals, friction is typically something that needs to be reduced' (Ruckenstein & Lehtiniemi, 2026, p. 281) and argue that friction needs to be studied together with the efforts to reduce it. Pelizza (2016) proposed thinking about frictions as controversies in which multiple configurations struggle to replace each other to produce the highest-quality data.

Analyses of friction need to better incorporate these tacit normativities and expectations, while not essentialising them. To do so in the context of ML work on biomedical data, I take inspiration from the relational thinking embraced by infrastructure scholars, particularly the analysis of tensions between infrastructural elements by Star and Ruhleder (1996). The relationality principle posits thinking in connections rather than 'things' and argues that something becomes an infrastructure in relation to certain work practices. This has two important implications for the present study. First, one person's infrastructure is another's work and vice versa; the notion that infrastructures support work is relative to the work in question, but infrastructure also shapes the work (Star & Ruhleder, 1996). Second, 'computers, people, and tasks together make or break a functioning infrastructure' (p. 118): the overall assemblage of relations, not individual elements in themselves, shapes when an infrastructure is.

Star and Ruhleder (1996) apply this thinking to explain the challenges arising in the seemingly straightforward processes of signing on to and starting to use a biology software in a research community. To do so, they bring in the distinction between message, context, and metacontext from linguistics and information studies. While the task of implementing the software is the 'message', learning how to do it invokes broader contexts and meta-contexts:

learning *about* learning. Such more meta-contextual and tacit meanings are often mentioned as arguments for why knowledge cannot be simply transferred as information and requires socialisation into an expert community: for instance, knowing how difficult a task is supposed to be is crucial meta- and second-order information about a skill that would help a learner not give up too soon (Pinch et al., 1996). Drawing on information studies, Star and Ruhleder (1996) note that messages and different levels of context related to them do not always align, producing discontinuities. The tensions between seemingly practical matters and their meta-contexts, and between different contexts involved in software use, demonstrate the complex relationship comprising the potential infrastructure. For example, the software itself is easy to use, but it runs on Unix, while biology has been a Mac-using community. The issue there is not only a matter of working with a new operating system but also of broader-context matters such as the value of having the same computing system as one's colleagues and the field's collective identity. In addition, the software's continuous-update function was meant to replace regularly scheduled updates, but many biologists saw the deadlines the schedule imposed as productive tools of organising their workflows, and sometimes researchers wanted to wait on sharing information. Star and Ruhleder noted many such issues preventing the software from becoming an infrastructure for biologists. Some of these examples can be seen as tensions between technology designers' and users' epistemic cultures, or standard system properties versus local practices, but their origins (as well as context levels) differ and thus were theorised as tangles between contexts.

I use this broad notion of context discrepancies to create a holistic understanding of friction between various contexts and practices that make up ML practitioners' working relationship with biomedical data. These elements of the knowledge infrastructures can include tensions as conceptualised in the friction literature (between places or material forms, epistemic cultures, or the global versus the local) but are not tied to them. For instance, some issues causing data cascades in high-stakes AI (Sambasivan et al., 2021) can be understood in terms of such frictions. Real-time data involved in these projects is messy and requires a lot of data work and attention. As ML professionals are often used to working on much cleaner benchmarking datasets, the unaccounted discontinuity between data features and their approaches generates friction and can cause unexpected problems down the line.

Methodology

The paper draws on an interview study focused on the culture and ethics of ML work. In the spring and summer of 2024, I conducted 44 interviews with ML practitioners in academia and industry working in Denmark, the

Netherlands, and the United Kingdom. This paper focuses on a subset of data collected in that study: 25 interviews with practitioners, 19 of whom at the time of the interviews worked with, and six of whom had previously worked with biomedical data and discussed these projects in the interviews.

At the time of the interviews, most of the practitioners worked in research or research and development, primarily in academia but also in industry (see Table 1). Some specialised within health-related contexts; others worked with biomedical data in one-off projects (e.g., a PhD project, after which they moved to another field) or in one of the avenues of their research. The participants worked with various ML methods and diverse types of biological and medical data, such as genomic sequences, clinical records, medical images, and signals. This included generating new data as well as reusing data from open sources (e.g., biobanks) and health organisations. Some projects were in collaboration with medical researchers and clinicians, with these 'domain experts' performing different roles from task definition to data collection and annotation (Diaz & Smith, 2024).

Table 1. Descriptions of the interviewees and their current[a] work

Total	25
Gender	Men (19), women (6)
Country of employment	Denmark (12), the Netherlands (8), the United Kingdom (5)
Academic employment	PhD student, incl. industrial (6), postdoc (4), assistant professor (4), associate professor (1), professor (2)
Non-academic employment	Data scientist (2), research and development (4), business analyst (1), software engineer (1)
Places of non-academic employment	Medium or large company (4), small company (2), startup (1), independent foundation (1)

a. Six of them worked on medical data in their previous employment.

Five interviews were conducted online, the rest in person, lasting from forty minutes to an hour and forty minutes. The interview guide facilitated narratives of practitioners' career journeys, past and current projects, collaborations, the data they worked with, and experiences with ethics and data governance. After all interviews in the broader study (44) were transcribed, I began open coding (Charmaz, 2006) from autumn 2024 to the beginning of 2025. As this yielded many codes about the data, this broad theme was selected as one avenue for deeper exploration, as suggested in the flexible coding approach that supports using a corpus of qualitative data for multiple substudies and returning to the data at different points in time (Deterding & Waters, 2021). In May 2026, I returned to the data for focused coding, rereading all

interviews, working with previously generated open codes, refining the coding scheme, and putting emerging themes in dialogue with literature. The salience of the biological and medical data was noted at that moment, resulting in a focus on this subset of the interviews in this analysis.

Findings: friction in ML work with biomedical data

As mentioned, many of the interviewees characterised biomedical and health data as challenging to work with for different reasons. Here, I discuss the tensions brought up by practitioners across two dimensions: accessing and knowing biomedical data. Within each dimension, I show that friction is generated between elements of the overall work relationship, including naturalised modes and assumptions of ML work. These relationships determine what is friction and what is simply work.

Accessing biomedical data

Accessing different types of data can be more or less difficult, e.g., due to blockages (or safeguards, depending on perspective) to their movement and (re)use (Bates et al., 2016) related to data governance and ownership. For instance, the interviewees noted that proprietary or sensitive data is less likely to be shared as part of published research, which has implications for the reproducibility of methods developed using such datasets. Notably, biological and medical data can also be released and shared as open data, as is the case with genomic data facilitated by biobanks. Many biomedical benchmark data, e.g., medical imaging benchmarks, have also received attention in critical data(set) studies (Jiménez-Sánchez et al., 2024; Luitse et al., 2025). When describing friction in accessing biomedical data, the interviewees mostly referred to (re)using data generated and owned by public medical organisations.

In projects that (re)use health data, access entails oversight procedures that can involve actors like ethical committees, data stewards, etc. Moreover, data sharing and processing procedures need to be negotiated and drawn up among the collaborating parties and translated into digital infrastructures. These processes can be part of the work that ML practitioners have to do, such as applying for ethical approval or data access, or meetings with legal or privacy specialists. However, these processes can generate friction in ML projects beyond simply the work of applying for access or drawing agreements.

First, several interviewees noted that these processes can take from months to a year or more, which generates friction against the temporalities of contemporary academic research (Felt, 2025). A senior bioengineering researcher said:

> It really means that you have a PhD student, or a project starting, and then … two years after, you get approval to use a dataset, but then you're already done with your project, the PhD student is in the middle of starting something new.

Three interviewees participated in a project that entailed pooling data from multiple hospitals. A PhD student working on the privacy-preserving algorithms for the project described the effect of the complexities of this multi-actor collaboration on the project's timeline:

> And basically, the past four years, we spent a lot of time getting all of the legal agreements in place. And we're still in the process of setting up servers and that sort of thing. I finished algorithms ages ago. We could just press a button, assuming it's installed today, and then we have a model, but we haven't yet. I think officially we have access to the data, but the server isn't installed yet. So, we have data on a computer here, but not on the computer where it needs to be.

This quote draws attention to the material and infrastructural aspects of compliance, highlighting that access is more than a binary switch that turns on and allows working with the data. Rather, it is a continuous dimension of one's relation to data – access*ing* – that can afford certain ways of interacting with the data but might exclude others. For example, in this project, the current access arrangement does not afford training the quoted PhD student's models on these data, as they are not stored where they need to be for that interaction. However, the same access arrangement allowed another team member, a statistics PhD student, to work on data wrangling.

In addition to whether or when, friction can generate in *how* practitioners access the data. This was illustrated by an R&D specialist working in a medical AI company. For a project in collaboration with a public healthcare organisation, he and his colleagues were granted access to a cloud server set up and controlled by the healthcare organisation's IT specialists. While both parties agreed on this storage option, the server's restricted internet access ended up being 'the main annoyance in actually working with this data'. While the practitioner was unsure whether this restricted internet connection was necessary for GDPR compliance or could have been avoided, he explained that it made using existing tools and libraries, an important part of development, more difficult:

> You know, when we develop software, we want to download … open source dependencies, packages, other libraries that somebody else built, and use that to build our software on top of. That's like a major thing. That's what we do, right? We wouldn't get anywhere if we couldn't reuse stuff. And that we could not do there, right? So we could not pull things from the internet down.

Instead, they had to run boxed code that they developed and tested within their own system on a mock-up of the server environment. This, in turn,

became a 'whole step of back and forth over this … building, pushing, pulling that fairly large package of code', because the development and testing worked iteratively:

> It's like you had to build it, pull it over, run it, and then realise: Oh no, we missed this part. Oh, now I want to do this thing that I can't do. Then you have to build it again.

The friction generated in this accessing relationship is between the mode of work afforded by the server restrictions and the established practices of working with data in ML, which involve using open-source libraries and tools (from the internet) and iterative rebuilding and testing of the model on the data. Similarly, an academic researcher who often works on healthcare projects noted that accessing arrangements with hospital data do not provide them with all the functionality ML research requires:

> Okay, hospitals give you access to their data. Now, you cannot keep on accessing their servers and doing stuff there. You want to do some things on, say, very large models. Their servers are not powerful enough. So, you want to do it in your GPU. But you cannot get the data properly.

This quote illustrates the friction between the local access arrangement and what it means to 'get the data properly' (using Graphics Processing Units). These frictions can generate particular ways of working with data that aim to reduce them (Pelizza, 2016). For instance, this researcher and his team ended up creating synthetic data as a proxy for the hospital data that could not be accessed 'properly'. For a PhD student working with data from a sleep clinic, who was allowed to put these data on a local hard drive but not the cloud, this restriction meant engaging with data with lower computational power:

> I have one GPU that I'm using, and it's definitely very little compared to what people do in machine learning. But for the non-medical projects where I use public data sets, I can use the bigger compute servers that we have, because then I can just take the data and put it there. But for the medical, for the in-house data, I have to be restricted. <…> So it kind of is what it is, because you know that these things are difficult with medical data and that you cannot just put it on like Amazon clusters or whatever.

This accessing relationship serves as one of the distinctions between the 'medical' part of her project, which must accommodate the restrictions of medical data, and the ML-focused part, where using public data allows her to produce state-of-the-art machine learning algorithms that require more computing resources. Notably, the access arrangements and resulting practices in these three cases vary: for instance, the R&D specialist and his colleagues worked on one of these cloud servers (though serviced by Microsoft Azure, not

Amazon), thus having access to the 'big' AI infrastructure (van der Vlist et al., 2024). However, in all three cases, the local accessing relations create friction with the naturalised elements of ML work, such as iterative development or the use of open tools and cloud computing.

When describing these challenges, the interviewees, like the PhD student above, acknowledged the restrictions and privacy or security concerns associated with biomedical data. In doing so, they invoked the data type and domain as narrative explanations for these frictions. For most interviewees, these instances of friction were seen as slowing down and/or limiting their work. In the words of Bonde et al. (2019), friction in this accessing dimension was seen as cumbersome rather than productive. One academic researcher, however, invoked data constraints as one of the aspects that make working in the healthcare context interesting for her, as she wants to build technology that would work in the real world:

> Healthcare is a domain where you are faced with a challenge, because all of this data is super sensitive. You can't just go around and put it on some cloud. <…> Let's say journals, clinical journals. Clinicians make notes. Ideally, they would love it summarized, right? But the best summarization model now is probably one of these big models … [But] you can't even upload your clinical notes to the cloud. So, you still have to take a step back away from one of these amazing technologies and go back to something smaller. Something you can see, and feel, and trust in. It's not behind a black box, for instance. So just being in the clinical setting helps me bridge this need for me to see my research this way.

This framing of the issue makes the tension between best practices in machine learning (big models, cloud storage) and clinical data storage restrictions more generative, or even the name of the game of doing applied machine learning.

These narratives illustrate the friction practitioners experience in accessing biomedical data, and specifically data from health organisations. This friction is not just a function of the data blockages or safeguards (Bates, 2017) that restrict their movement; it is generated by the uneasy fit between the affordances of the situated access arrangements and naturalised practices of ML research and development. This friction and attempts to reduce it create specific epistemic practices that allow the more general application of ML/AI methods (Hartley & Thylstrup, 2025; Thylstrup & Hartley, 2025), such as synthetic data or 'low-resource' machine learning using lower computational power. In doing so, it contributes to the divergence between (the notions of) 'general' ML and ML in the healthcare domain.

Knowing (how to work with) biomedical data

Working with biomedical data entails direct (in case of collaboration) or indirect (through data) interaction between different epistemic cultures, which is known to generate 'science friction' (Edwards et al., 2011) or 'domain friction' (Hartley & Thylstrup, 2025) of adapting ML/AI logics to the bio- and healthcare domain.

Interestingly, direct collaboration with clinicians or biomedical researchers was not described as generating friction. Some participants described that understanding each other and staying on the same page required time and communication; a couple mentioned that the produced models were not always to domain specialists' liking. However, the interviewees did not refer to collaborations as being difficult, limiting the results, or slowing their project down. In contrast, several interviewees described learning about the domain and data by exposure, through continuous communication with collaborators. For instance, in response to my question about how much he needs to learn about medicine in his projects, an academic ML researcher explained:

> I guess you need some basic understanding of the data and where it comes from. And also need to understand that this is patient data. And that means that we can expect kind of noisy data and all sorts of artifacts. And there's really that diversity in data samples that you can encounter that is incredible. Compared maybe to computer vision, which is a bit more standardised, a bit of higher quality, and a lot more data. But when it comes to clinical expertise, I guess it comes with experience. If you collaborate with someone, you at some point start speaking the same language. You know what they mean if they use some sort of clinical terms. In my case, I already have a good starting point because I studied biomedical engineering. <…> And my PhD students are also helping with annotations. So, they know the clinical problem. They always get some sort of brush-up course or some crash course on, you know, the medical problem. So, yeah, there's definitely some time investment in understanding each other.

This quote distinguishes between the two kinds of knowledge one needs for this kind of work: 'understanding of the data' and its features (e.g., noise, heterogeneity) versus 'clinical expertise' or domain knowledge. One or both types of knowledge were mentioned as part of adaptation to a topic for a new project or job. Like this interviewee, practitioners often connected the latter to understanding and speaking the language of domain researchers and practitioners. One PhD student, who said she had learned 'definitely more than what [she] expected coming out' about cancer biology, referred to medical specialists as potential audiences of her papers: 'being in this applied world where you kind of want to take your research and actually have an impact with it, you need to know something about your biological system'. While acquiring domain

knowledge and communicating with experts require time and effort, these narratives portray it as work rather than friction. Even as a couple of interviewees noted that they might avoid future projects that would require learning about a whole new topic, this reasoning assumes that investing time in domain knowledge is the name of the game.

However, the nature and features of the biomedical data were often invoked as challenging and/or limiting for machine learning research and development. Studies noted that 'real' data, including in high-stakes domains, is messier than the standardised datasets often used in ML research and training (Muller et al., 2019; Sambasivan et al., 2021). The interviewees similarly noted that biomedical data, e.g., medical images described above, had specific features, like noise and heterogeneity, that presented challenges. Some interviewees, like the researcher quoted above, also contrasted these features against the benchmark datasets used in methods-oriented fields. For example, one practitioner had recently shifted from foundational computer vision research to industry research in a healthcare company. Working now with medical images, he was still learning to make ML doable in this new context, discovering new specificities, even though 'the backbone is still the same':

> So what makes it challenging is when you develop something for fundamental vision, it's on natural images. <…> In natural images, you would have spatial bias. And then you see a thing, you would expect a cloud on top, you would expect some grass on the bottom, maybe some cow on top of the grass, or maybe some wolf. <…> And… you incorporate things that work for natural images to your network. But in medical imaging, <…> there's no spatial bias. There's no context. There are no co-occurring objects. There is no colour distribution maybe even. So, this makes a difference. So, maybe… some overcomplication that makes it work well on ImageNet or natural image domain is not necessary in medical imaging. Or many times, hyperparameters do not translate at all. So, you need to discover them again. This also slows it.

STS literature has long emphasised that standard elements of knowledge infrastructures require tinkering in places of use (Fujimura, 1987; Knorr, 1979), and adapting neural networks for particular use contexts can involve substantial reconfiguration (Engdahl, 2024). This quote illustrates such domain adaptation as friction (Thylstrup & Hartley, 2025). However, it also highlights that these tensions are not just about issues that are often framed as data quality or degree of needed intervention versus readiness for analysis (Muller et al., 2019). In this case, the friction originates due to the differences between features of medical imaging data and of natural images (assumed in the general approach), which turn out to be essential for modelling. As this is not expected or explicitly addressed by the practitioner, what causes this friction is also the more meta-, second-order notions (Pinch et al., 1996; Star &

Ruhleder, 1996), such as the assumption that these differences in data features would not be important, and that hyperparameters would translate.

Another example of a friction-generating relationship that goes beyond data *features* came from a recently finished postdoc who worked on an interdisciplinary project dedicated to identifying underdiagnosis of mental health issues by considering potential mental illness markers (such as following suicide) in the population data. He described that the project turned out to be much more difficult than he and his colleagues had 'naively' (his term) expected:

> … because once you get into the weeds of this data, there are all kinds of very strange questions like … when exactly is the diagnostic label actually entered into the system? And it turns out that this is waaay more complicated than you might ever think, because only people in secondary care facilities can do this in certain circumstances, but then somebody else will also do it for a different reason in this special case. And so there are all kinds of very complicated stories behind basically each of the data points that I have in this table. And it turns out that these are actually very important to the analysis that I do. I mean, you get into questions like, how is it labelled whether something is a suicide or not? I mean, a person is found dead in the apartment. It's not immediately clear, is there an investigation or not? <…> So there are all of these kind of weird, very detailed questions, but they actually turn out to be very important for the technical analysis. And that is something that just the domain person can answer.

As in the previous case, the friction arises not only because the practitioner needs to learn about the data labels, which takes a lot of time. Rather, the friction is generated against the second-order notions of, e.g., how difficult the task is, how straightforward the labels are supposed to be, and how relevant these 'domain' details are for the 'technical analysis'. These notions are challenged within the working relationship constructed in this project, which includes getting 'into the weeds' of the data and collaborating with philosophers and epidemiologists, who provided detailed interpretations of the variables and asked their own questions that, as the interviewee described, 'me as a technical person, maybe I just wouldn't have thought about'. This friction results in new epistemic practices. The interviewee further explained that this project showed him that 'a lot of things we put in tables and then analyse in a certain way are … socially made-up constructs', making him more ambivalent about algorithmic decision-making. This friction produces a new relationship with the specific data as well as with data more broadly.

These two cases demonstrate that friction challenges and highlights meta-assumptions of ML work, such as the distinction between what is part of knowing data (and thus part of ML work), and what is domain knowledge. These notions and expectations are an important part of friction-generating

relationships. The last vignette illustrates how similar features of the data can interact with different definitions of the ML work tasks, potentially generating different frictions.

A PhD student mentioned in the discussion of computer restrictions on medical data use works with data from sleep clinics, labelled by experts for different stages of sleep (e.g., REM). Describing this data, she explained:

> Sleep staging is not a process that has 100 percent [agreement] between experts. So, you also notice that the distribution of scored sleep stages on the group level can be slightly different in two clinics, and it's not because the people in another clinic probably have different sleeping patterns. Just assuming that it's like the same kind of population. But it's more likely also that the persons that scored the data set have a <…> different style of scoring.

She also noted that one of the five classes in the categorisation is labelled least consistently. Another interviewee, an industry specialist who had previously worked with sleep staging data during his PhD, also noted that the doctors who labelled the data he worked with did not fully agree in their labelling. As he explained, this made it more difficult to estimate how well the model did compared to a 'human level', but for the task itself, they could bracket this problem by following only one annotator 'because again, there is no reason to believe one of them is better than the other one, and, solving the task from a machine learning perspective, [it] doesn't matter which one it is, because it's the same thing we're doing'.

Both interviewees present ground truth not as an unquestioned fact but as a result of expert and yet imperfect human practice (Henriksen & Bechmann, 2020; Muller et al., 2019). However, for his project, this uncertainty in labels could be bracketed and thus arguably did not generate friction for that task. Indeed, he brought up this subjective labelling matter when discussing the claim that the resulting algorithm works at the level of the human, rather than presenting it as a challenge to his work. For the PhD student, however, this uncertainty seemed more problematic:

> It is because it's the only ground truth we have, and we know that it's not golden truth, and you know that it differs per clinic, and we know that it differs per person, and it depends on all kinds of things, how people score. So, yeah, it's kind of the best you can do. And then even if you try to go beyond those labels and try to observe other patterns, the only thing you can do is always refer back to the labels, of which we know that they are fuzzy. So, yeah, it's a bit odd because it makes it much harder to move out of that and to show that there's other ways of representing that data.

For her project, which used unsupervised modelling to 'go beyond those labels and try to observe other patterns', these labels were a bind as they were the

main tool of evaluation, demonstrating the formative role of the ground truth in unsupervised modelling as well (Jaton, 2025). She thus explained that the type of problem and method (such as unsupervised vs. supervised modelling) shaped how much these domain issues affected ML work, and that if one approaches sleep-staging data as a 'just classification problem', this variability in labels would be less of an issue. This task definition would mean less immersion in the domain but would still require knowledge about the classes. Illustrating this point, she told the story of a presentation she watched, where someone claimed a model's high level of accuracy on that class of sleep stages where expert labellers' agreement was low:

> which is not possible if the ground truth labels don't even have this accuracy, right? And then someone in the audience asked: why, how do you achieve that? And then this person answered like, oh, how do you mean? Like, it's around in the same range as the other classes. So, I don't see the problem. So, I mean, then you're really considering the problem as a classification problem and don't have any knowledge of the meaning of the different classes.

In telling this story, the interviewee draws boundaries against this agnostic technical approach. The three different approaches to labels and their fuzziness illustrate how different framings of the problem contribute to different approaches to knowing data – and the differing degrees of friction their features can generate with evaluation practices in ML work. They also suggest that frictions during the work can sometimes be reframed as safeguards against data cascades, as tensions can be absent in the ML work relationship but arise later (Sambasivan et al., 2021). Overall, the frictions in knowing biomedical data demonstrate that friction is the product of not just data features or the combination of different epistemic cultures but of situated work relationships, generated between different elements.

Discussion

I set out to explain the frictions ML practitioners experience in projects using biomedical data. Based on the interview narratives, I organised these frictions along two dimensions of their relationships with biomedical data: accessing and knowing. However, the frictions in accessing also turn out to be intertwined with the ML practitioners' epistemic practices, highlighting the naturalised elements of ML work and adding to the differences between 'methods' and applied ML in healthcare. These insights emphasise that issues of data governance and sharing are not absent or arising *barriers* but elements shaping epistemic practices in ML through friction in relations to other contexts (also see Zając et al., 2023).

This finding is part of the paper's main argument that data friction emerges in the overall working relationship, or 'data settings' (Loukissas, 2019). Wirth et al. (2025) argued that 'data are never "clean" – to use the terminology of statistical practitioners – but adapted or not to a particular use, situated both in time and in the work-relationship' (p. 65). This paper suggests extending this argument to notions of noisy or messy data resulting from biological and healthcare data (re)use in machine learning – they are noisy or messy compared to the goals of the work and its underlying ideals and assumptions. In the context of biomedical data in ML, this includes the relationship with data governance procedures, infrastructures supporting clinical data sharing, data features, and established practices and tacit assumptions of machine learning work.

This paper does not claim to comprehensively outline all the potential discontinuities in biological and health ML projects. Part of the argument is that these frictions arise in situated relationships, and what generates friction in one context would not in another, similar to infrastructures (Star & Ruhleder, 1996). Highlighting this relational and contextual nature of friction is a key contribution of the paper.

I draw on an interview study and use friction to explore the challenges described by the interlocutors. While the writing of the paper implies that I see this analytical strategy as productive, I must acknowledge the particularity and limitations of this 'window' into the world of ML work. This interpretive approach has allowed me to bring in notions and assumptions of ML work and develop this relative conceptualisation. In doing so, I approach an issue raised by Edwards et al. (2011): all planned action needs to include some flexibility for improvisation and thus a bit of friction. However, seeing friction everywhere would diminish the analytical potential of the concept. Not all improvisation and contingencies are experienced as problematic and disruptive to the process, and in the words of Fujimura (1987, p. 260), there are reasons why 'one person's emergency is another person's routine'. These considerations emphasise the role of naturalised second-order notions like skill difficulty (Pinch et al., 1996) or methods vs. domain distinction (Ribes et al., 2019) in what is friction and what is business as usual.

However, this interpretive approach also has limitations. Such an inquiry cannot account for all the issues that can arise in ML knowledge infrastructures, as the interviewees' perspectives are partial and specific to their locations. Crucially, issues in ML/AI can manifest somewhere else in space and time compared to where they originate, making it difficult to find and deal with the underlying causes (Sambasivan et al., 2021; Widder & Nafus, 2023). Here, the important consideration I make based on the sleep-staging cases is that in certain work relationships, different types of potentially generative

friction can also *not* arise. Second, this approach focuses on the perspective of one community of practice. Following the theoretical lens used here, it is only logical to assume that other groups doing different work, such as data stewards or clinicians, would have different experiences, and what is friction for ML practitioners might not be friction for them (and vice versa), or their experiences of generative versus cumbersome would differ. Hoeyer (2023, p. 245) notes that his analysis focuses on the extra invisible work data reuse creates for healthcare workers rather than productive insights such as those described by Bonde et al. (2019). Considering that domain specialists tend to be deskilled and devalued (Diaz & Smith, 2024; Sambasivan & Veeraraghavan, 2022), their perspectives might be drastically different.

These considerations draw attention to the normativities and politics of friction as part of knowledge infrastructures (Star & Ruhleder, 1996) and suggest several important areas for future empirical work utilising the friction lens. How do actors experiencing friction understand it: as cumbersome and/or productive (Bonde et al., 2019), constraint and/or resource, friction or work? Who gets to decide what friction is productive and what is a loss or waste? What implications can these understandings have for the data infrastructures and their politics? Attention to such relative matters would make the friction lens an even more useful tool for denaturalising computational practices.

Conclusion

In this paper, I employed friction as a conceptual lens to nuance the understanding of machine learning practitioners' experiences with biomedical data, which many described as challenging to work with. Exploring the frictions practitioners experience in accessing and knowing biomedical data, I demonstrate that these challenges come not necessarily from the features of the biomedical data or the collision between machine learning and domain knowledge. Rather, friction generates between the different elements of the situated working relationships constructed in these projects, including data governance procedures, infrastructures supporting clinical data sharing, data features, and established practices and tacit assumptions of machine learning work. These relationships also matter for whether frictions are considered cumbersome or productive, and whether they are experienced as frictions at all or seen as the name of the game.

References

Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others*. Duke University Press. https://doi.org/10.1515/9781478009276

Aula, V. (2019). Institutions, infrastructures, and data friction – Reforming secondary use of health data in Finland. *Big Data & Society*, *6*(2). https://doi.org/10.1177/2053951719875980

Bates, J. (2017). The politics of data friction. *Journal of Documentation*, *74*(2), 412-429. https://doi.org/10.1108/jd-05-2017-0080

Bates, J., Lin, Y.-W., & Goodale, P. (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society*, *3*(2), 2053951716654502. https://doi.org/10.1177/2053951716654502

Bonde, M., Bossen, C., & Danholt, P. (2019). Data-work and friction: Investigating the practices of repurposing healthcare data. *Health Informatics Journal*, *25*(3), 558-566. https://doi.org/10.1177/1460458219856462

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world* (1 ed.). MIT Press. https://doi.org/10.7551/mitpress/9963.001.0001

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE.

Davies, S. R., & Díaz-Valderrama, M. (2026). Attending to 'computational universalism': Practices, frictions, events. *Social Studies of Science*, *56*(3), 340-352. https://doi.org/10.1177/03063127261445803

Deterding, N. M., & Waters, M. C. (2021). Flexible coding of in-depth interviews: A twenty-first-century approach. *Sociological Methods & Research*, *50*(2), 708-739. https://doi.org/10.1177/0049124118799377

Diaz, M., & Smith, A. D. R. (2024). What makes an expert? Reviewing how ML researchers define "expert". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*(1), 358-370. https://doi.org/10.1609/aies.v7i1.31642

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. The MIT Press.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667-690. https://doi.org/10.1177/0306312711413314

Engdahl, I. (2024). Wuthering weights—Localisation trajectories of machine learning models for local ends. In V. Fors, M. Berg, & M. Brodersen (Eds.), *The De Gruyter handbook of automated futures: Imaginaries, interactions and impact* (pp. 321-336). De Gruyter. https://doi.org/https://doi.org/10.1515/9783110792256-020

Felt, U. (2025). *Academic times: Contesting the chronopolitics of research*. Springer. https://doi.org/https://doi.org/10.1007/978-981-96-4609-8

Fujimura, J. H. (1987). Constructing 'do-able' problems in cancer research: Articulating alignment. *Social Studies of Science*, *17*(2), 257-293. https://doi.org/10.1177/030631287017002003

Hartley, J. M., & Thylstrup, N. B. (2025). The algorithmic gut feeling – Articulating journalistic doxa and emerging epistemic frictions in AI-driven data work. *Digital Journalism*, *13*(3), 438-456. https://doi.org/10.1080/21670811.2024.2319641

Henriksen, A., & Bechmann, A. (2020). Building truths in AI: Making predictive algorithms doable in healthcare. *Information, Communication & Society*, *23*(6), 802-816. https://doi.org/10.1080/1369118X.2020.1751866

Hoeyer, K. (2023). *Data paradoxes: The politics of intensified data sourcing in contemporary healthcare* (1st ed.). The MIT Press. https://doi.org/https://doi.org/10.7551/mitpress/14926.001.0001

Jaton, F. (2025). Examining algorithms in the light of their ground truth datasets: Results, objections, and avenues of reflection. *Digital Society*, *4*(2), 61. https://doi.org/10.1007/s44206-025-00197-4

Jiménez-Sánchez, A., Avlona, N.-R., Juodelyte, D., Sourget, T., Vang-Larsen, C., Rogers, A., Zając, H. D., & Cheplygina, V. (2024). Copycats: the many lives of a publicly available medical imaging dataset. *Advances in Neural Information Processing Systems*, *37*, 113383-113404. https://doi.org/https://doi.org/10.52202/079017-3602

Knorr, K. D. (1979). Tinkering toward success: Prelude to a theory of scientific practice. *Theory and Society*, *8*(3), 347-376. http://www.jstor.org/stable/656764

Lee, F., & Ribes, D. (2025). Computational universalism, or, Attending to relationalities at scale. *Social Studies of Science*, *56*(3). https://doi.org/10.1177/03063127251345089

Loukissas, Y. A. (2019). *All data are local: Thinking critically in a data-driven society*. The MIT Press. https://doi.org/10.7551/mitpress/11543.001.0001

Luitse, D., Blanke, T., & Poell, T. (2025). AI competitions as infrastructures of power in medical imaging. *Information, Communication & Society*, *28*(10), 1735-1756. https://doi.org/10.1080/1369118X.2024.2334393

Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., Dugan, C., & Erickson, T. (2019). How data science workers work with data: Discovery, capture, curation, design, creation. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper 126. https://doi.org/10.1145/3290605.3300356

Pelizza, A. (2016). Disciplining change, displacing frictions. Two structural dimensions of digital circulation across land registry database integration. *Tecnoscienza – Italian Journal of Science & Technology Studies*, *7*(2), 35-60. https://doi.org/10.6092/issn.2038-3460/17326

Pinch, T., Collins, H. M., & Carbone, L. (1996). Inside Knowledge: Second order measures of skill. *The Sociological Review*, *44*(2), 163-186. https://doi.org/10.1111/j.1467-954X.1996.tb00420.x

Ribes, D., Hoffman, A. S., Slota, S. C., & Bowker, G. C. (2019). The logic of domains. *Social Studies of Science*, *49*(3), 281-309. https://doi.org/10.1177/0306312719849709

Ruckenstein, M. (2023). *The feel of algorithms*. University of California Press. https://doi.org/https://doi.org/10.1525/9780520394568

Ruckenstein, M., & Lehtiniemi, T. (2026). Friction and promise in data labor. *Science, Technology, & Human Values*, *51*(2), 277-303. https://doi.org/10.1177/01622439251358900

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/https://doi.org/10.1145/3411764.3445518

Slota, S. C., Hoffman, A. S., Ribes, D., & Bowker, G. C. (2020). Prospecting (in) the data sciences. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720906849

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1), 111-134. http://www.jstor.org/stable/23010792

Thakkar, D., Ismail, A., Kumar, P., Hanna, A., Sambasivan, N., & Kumar, N. (2022). When is machine learning data good?: Valuing in public health datafication. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Article 322. https://doi.org/10.1145/3491102.3501868

Thylstrup, N. B., Hansen, K. B., Flyverbom, M., & Amoore, L. (2022). Politics of data reuse in machine learning systems: Theorizing reuse entanglements. *Big Data & Society*, *9*(2). https://doi.org/10.1177/20539517221139785

Thylstrup, N. B., & Hartley, J. M. (2025). "Argh! The world doesn't fit the model!": Small acts of worldmaking in data annotation for news media. *Media Theory*, *9*(2), 105-132. https://doi.org/10.70064/mt.v9i2.1273

Tsing, A. L. (2005). *Friction: An ethnography of global connection*. Princeton University Press. https://doi.org/ https://doi.org/10.1515/9781400830596

van der Vlist, F., Helmond, A., & Ferrari, F. (2024). Big AI: Cloud infrastructure dependence and the industrialisation of artificial intelligence. *Big Data & Society*, *11*(1). https://doi.org/10.1177/20539517241232630

Widder, D. G., & Nafus, D. (2023). Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231177620

Wirth, A., Ludec, C. L., Girard-Chanudet, C., & Cornet, M. (2025). Who decides when artificial intelligence works? Algorithmic production as a continuous evaluation process. *Digital Society*, *4*, Article 65. https://doi.org/10.1007/s44206-025-00224-4

Zając, H. D., Avlona, N. R., Kensing, F., Andersen, T. O., & Shklovski, I. (2023). Ground truth or dare: Factors affecting the creation of medical datasets for training AI. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 351–362. https://doi.org/10.1145/3600211.3604766

Zilberstein, S. (2024). Ethical Dilemmas and Collaborative Resolutions in Machine Learning Research for Health Care. *Socius*, *10*. https://doi.org/10.1177/23780231241259671

Chapter 8
Discussion and conclusion

> Concepts like 'nation,' 'society,' and 'culture' name bits and threaten to turn names into things. Only by understanding these names as bundles of relationships, and by placing them back into the field from which they were abstracted, can we hope to avoid misleading inferences and increase our share of understanding (Wolf, 1982, p. 3).

This quote illustrates anthropologist Eric Wolf's argument that world history is made of interconnections rather than 'things'. On a smaller scale, my research journey, the twists and turns of which are described in the previous chapters, led me to examine ML work 'bit' by 'bit': from labels, to ethics, to (biomedical) data.

Drawing on interviews with ML practitioners and an eclectic theoretical toolkit, each inquiry has revealed bundles of relationships. In paper 1 (see chapter 5), the analysis of practitioners' meaning-making about 'machine learning' and 'AI' labels demonstrated how labelling is intertwined with technologists' relations with non-specialists, popular representations of AI as large language models, and distinctions among technologists. In paper 2 (see chapter 6), I explored how ethics – the way practitioners themselves understood this concept – fits into their representations of ML work and showed how these representations are rooted in technologists' social location and experience, but also draw on shared concepts and narratives. In paper 3 (chapter 7), I departed from questions of representation and, at the same time, narrowed my analytic attention to consider friction in ML work with biomedical data. This analysis situates notions and practices of ML work within integrative working relationships, along with other elements such as data-sharing infrastructures, time plans, and knowledge encoded in biomedical data labels; it argues that friction is generated from the uneasy fit among elements in the overall working relationship.

Elucidating these relations is the central contribution of this dissertation, which addresses how ML work is *co-constituted by its relations with labels, ethics, and (biomedical) data as its 'others'*. The next sections further discuss the findings and contributions of my research along the three following lines: investigating vernacular concepts, examining ML work as both situated and embedded in broader contexts, and illuminating practitioners' relationships with society. Before concluding this chapter, I reflect on the limitations of the study and avenues for future research it opens up.

8.1 Analysing vernacular concepts in computational cultures

Recent debates around artificial intelligence and other concepts of computation have (re)drawn attention to language in STS and related fields. On the one hand, scholars argue that anthropomorphising and using language related to learning and intelligence reproduce hype and mystification of technologies (Inie et al., 2026; Kajava & Sawhney, 2023). On the other hand, critical scholars have noted that social and humanistic research on digital and computational phenomena has become too comfortable with their vernaculars, problematising these terms as 'hardening' categories that reify 'elements of which computational actors say the world is made' (Lee & Ribes, 2025, p. 326). This particularly applies to the 'AI' signifier, which, as noted by Lucy Suchman (2023), is ascribed 'uncontroversial thingness'. My analysis of concepts employed by practitioners in their thinking and talking about their work contributes both empirically and theoretically to this discussion.

Paper 1 approaches 'AI' and 'machine learning' as labels and showcases the meanings practitioners' ascribe to these labels as descriptors of their work. In doing so, it provides insight into one of the contexts in which the figure of 'AI' appears (Suchman, 2023): in communication with non-specialists, as an imperfect but useful outward umbrella for different kinds of ML work. Drawing on Stuart Hall's encoding-decoding model (Hall, 1973), my analysis highlights the connection between this representational practice and the tendency to equate AI with generative tools such as ChatGPT. Practitioners note and anticipate the 'generative' association from their 'non-technical' interlocutors and can use it in communication even if it is, in one interviewee's words, a 'wrong shortcut'. These findings emphasise the role of representation and communication in the continuous (re) construction of AI (Galanos, 2023).

They also highlight that the vernacular categories of computation (that social scholarship worries about naturalising) are not equally settled and 'thingified' among technologists themselves. For instance, 'machine learning' seems to be more of a stable 'thing' than 'AI', which the interviewees readily discussed as a term and a communication device. Rather, the 'thingness' is produced via the dynamic that can be seen as a representational analogue of MacKenzie's 'certainty trough' (1998). MacKenzie notes that technology producers express more uncertainty about that technology in private but limit it in public, which might contribute to certainty in actors invested in technology but not involved in knowledge production (such as managers, decision-makers, and some users). He posits that some uncertainty in knowledge production is normal, but these dynamics can be conducive to misplaced certainty outside of it. Based on my analysis, I suggest that the same might happen with the

connotations and 'thingness' of technology categories, which are enacted with greater certainty on the 'outside', which can have performative effects for medialised representations (Felt & Fochler, 2012) and technology hype (Galanos et al., 2026).

The tension between the desire to use concepts familiar to one's audience to establish understanding, and the reproduction of dominant significations and categories, is not unique to ML practitioners or other actors in proximity to buzzwords. Problematising 'artificial intelligence' and other concepts, some scholars suggest using alternative terminologies; for instance, Inie et al. (2026) argue for language that emphasises functionality, such as 'probabilistic automation', over language that supposes capabilities for learning or intelligence. While I do not offer a definitive argument on this matter, since my analysis shows these decisions are situated in relations to interlocutors and communicative goals, these insights draw attention to the connections between micro-level interactions and representation practices and broader cultural dynamics.

With the 'ethical turn' in ML/AI (Andrejevic, 2020), the same concerns about categories, and the work done by them and under their banners, can be applied to 'ethics' (see section 1.2 on the problematisation of 'ethics'). My approach to ethics through the conceptual register of representation thus complements the existing literature in empirical studies of ethics (Frahm & Schiølin, 2024; Helm & Gerlek, 2026), by showing how practitioners understand this notion and its relation to their work.

The key theoretical contribution of this dissertation is the conceptual metaphor of constellations, developed to capture representations of ethics as products of meaning-making. Constellations are sets of stars that *appear* together; as such, constellations exemplify how interpretation makes meaningful shapes out of disparate objects based on familiar cultural concepts. With this metaphor, I identified three representations of ethics I find in practitioners' narratives (ethics as applications, ethics as data, and ethics as a research topic). The analogy of objects in the night sky and the patterns in which they constellate, facilitated analysing individual elements of these representations as well as the figures their combinations produce. Drawing on this metaphor, the analysis demonstrates that ethics as applications and ethics as data have different loci not only in terms of topic but also scope of consideration: the 'whole topic' level of ethics as applications versus the attention to granular practices in ethics as data. In relation to these constellations, the particularities of applications and the broader discussions on data fade into the background. Moreover, the constellations metaphor captures these representations as both situated (as different constellations are visible depending on the date and viewer's location) and drawing on shared cultural concepts (more on

this in the next section). This conceptual metaphor highlights how a broad notion of ethics can be ascribed very particular understandings and roles in the practice, and what these representations invisibilise (Bowker & Star, 1999; Suchman, 1995).

Another computational category this dissertation provides insight into is that of 'domains', a vernacular term noted by Ribes (2019) and considered as a part of the 'logic of domains' (Ribes et al., 2019). This logic separates abstract computational methods from the 'domains' as areas of knowledge and activity to which these methods can be applied. Scholars have connected this logic to the universalising tendencies of the digital and computational ways of knowing, as it constructs a picture of the world as being made of concrete domains that are unified in that they are amenable to, and reliant on, computational analysis that can either replace or incorporate domain knowledge (Ribes et al., 2019; Slota, Hoffman, et al., 2020).

This study complements the existing literature on this concept and logic by showing how the notion of domains (and their distinction from methods) is entangled in the ways practitioners think and talk about their work. The methods versus applications distinction appears in paper 2 as part of the 'applications' constellation, which ascribes ethics to (potential) applications rather than methods. Choosing or avoiding work in certain domains is then part of doing ethics. This representation echoes the notion of computation as neutral or ambivalent (double-use) tools which gain sociopolitical valence in applications (Malazita & and Resetar, 2019). In paper 1, I show that the methods vs. applications distinction is used in differentiations *among* ML practitioners: between machine learning researchers who work on methods versus, for example, natural language processing researchers who use and develop these methods but also are interested in language. Similarly, paper 3 depicts the work in the biomedical 'domain'; several of the interviewees contrasted this work with more 'methods' or 'fundamental' machine learning using similar-modality data (signals, vision). Moreover, the 'domains' or application fields might not necessarily be on equal footing: computer vision has been historically connected to machine learning research (Dobson, 2023), while language is becoming more integrated into machine learning as a result of both exchange of methods, e.g., the transformer architecture developed initially for NLP (Vaswani et al., 2017), and the popularity of large language models as a topic. These distinctions among types of ML work and practitioners add nuance to the binary between data scientists or ML/AI engineers and 'domain experts', and remind us that machine learning is not a unified field or monolithic object, but bundles of interacting communities and practices. Here, the notion of 'domain' and the fractal distinctions it seems to articulate can add to the repertoire of concepts capturing these relations, such as supply chains

(Widder & Nafus, 2023) and epistemic hierarchies (Sambasivan et al., 2021; Widder, 2024).

8.2 ML work as both situated and embedded in broader narratives and structures

One of the challenges in studying machine learning is capturing both situated practices, often highlighted by laboratory studies (Engdahl, 2024; Jaton, 2021), and 'computational relationalities at scale' (Lee & Ribes, 2025) and global knowledge infrastructures (Edwards, 2010). Attending to both situatedness and embeddedness of ML work within broader discourses and institutional structures is the overall contribution of this dissertation.

The analysis of friction occurring in particular data settings (paper 3) articulates the situatedness most explicitly. Drawing on the relational thinking foregrounded in infrastructure scholarship (Star & Ruhleder, 1996), I approached practitioners' experiences of working with biomedical data as holistic working relationships, or what Loukissas (2019) termed 'data settings'. These include local encounters of regulations, research timescapes, 'general-purpose' methods, and (re)used data. The analysis then demonstrates that the frictions in ML practitioners' accessing and knowing relationships with data stem not from features of the data, or the collision between the computational and 'domain' epistemic cultures, but from the uneasy fit between different elements of situated working relationships. For instance, data features such as 'fuzzy' ground truth labels can present more or less of an obstacle for machine learning depending on the goals and framing of the task. This also shows how friction is relative to what is expected and supposed to work smoothly (Aula, 2019), highlighting the role of meanings and tacit normativities of work in generating friction. This relational understanding of friction accounts for why one project's friction is another one's business as usual.

At the same time, this dissertation's findings indicate the presence of shared culture and meaning-making among practitioners working with different methods and located in different countries, fields or sectors. This shared culture is illustrated by the joke about ML/AI/regression labelling at the beginning of paper 1 and includes codes such as 'AI' as an outward-facing signifier of their work for communication with the non-technical others and understanding of machine learning as a key part of AI.

The duality of situatedness and embeddedness is also reflected in the ethics constellations (paper 2). On the one hand, my analysis connects each constellation to certain locations and types of ML work: ethics as applications in applications-oriented methods research located up- or mid-stream of the development supply chain (Widder & Nafus, 2023); ethics as data – in work that

involved intensive collection and/or wrangling of personal data; and ethics as a topic – in specialised projects. Representations of ethics as data seem to be particularly connected to interviewees' immediate experiences and granular practices of data work. At the same time, these constellations are underscored by stable and shared narratives (for example, the methods vs. applications separation mentioned above), i.e., their narrative infrastructures (Felt, 2017).

Tracing these infrastructures adds to the understanding of how notions of ethics are intertwined with other narratives and practices in ML/AI development, such as tech industry values (Metcalf et al., 2019), epistemic hierarchies of types of work(Widder, 2024), open-source values and principles (Widder et al., 2022), etc. This literature emphasises that 'ethics' is not isolated from the work of science and technology but co-constitutes it (Pickersgill, 2012). As they underscore ethics notions, these infrastructural, 'background' (Abend, 2014), or second-order (Pinch et al., 1996; Star & Ruhleder, 1996) narratives and notions that are not necessarily *about* ethics are important to take into account in developing approaches to ethical ML/AI. Otherwise, the first-order ethics solutions might not work if they are incongruent with the meta-context of the practice, potentially creating tensions (Star & Ruhleder, 1996). These are the kind of tensions Ulrike Felt and colleagues note regarding promotion of societal responsibilities in life sciences (Felt, 2017; Sigl et al., 2020).

The analysis also shows that the narrative infrastructures shaping practitioners' representations of 'ethics' stem partially from institutionalised procedures, such as conference requirements (which include ethics statements and checklists) and ethical approval and data protection procedures. Tracing these narrative infrastructures contributes to the understanding of the relationship between institutionalised ethics and the compliance work it creates, and ethics as an actor's category (Helm & Gerlek, 2026). Previous studies indicate that while procedural ethics are supposed to support and ensure good practice, their relationship is more complex. For instance, Heimer argues that institutionalised ethics 'defines problems narrowly and then diverts scarce attention and resources to compliance work, making dissent difficult' (2013, p. 371). Samuel et al. (2021) find that health AI researchers' construction of research ethics is collated with (and solved by) data protection and ethical approval procedures, and it is in the issues out of these institutions' purview where ethics becomes more problematic. My results echo some of these findings. Indeed, my analysis identified these constellations based on the interviewees' delineations of non-ethical issues and areas of work as much as the ethical ones, and noted the concerns and practices backgrounded by the specific loci of the constellations (the 'whole topic' scope in applications vs. granular practices in data, and the focus on specialised projects in ethics as a topic). These foci and blind spots are important, but as they are studied by following the

'ethics' category, they might be more relevant to situations where this category is invoked.

Previous studies have assumed a connection between the word 'ethics' and the academic discipline of philosophy, governance approaches, and other 'official systems' rather than everyday practices of doing 'good' work (Heimer, 2013; Pickersgill, 2012). The finding that institutional procedures underscore constellations of ethics as their narrative infrastructures can to some extent be seen as an empirical confirmation of these connotations. However, this study also emphasises that this category should not be ignored in the quest to find out how actors *really* think and practice ethics on the ground. Rather, it can be seen as one of the 'regimes of normativity' or categories and narratives practitioners draw on in making sense of the world around them or going about their work. What shapes these narratives and categories and the work they do when intertwined with other meanings and practices are important considerations for both research and intervention. As such, the term itself might be what Shklovski and Némethy (2023) conceptualised as one of the 'nodes of certainty', that is, familiar and more routinised approaches. However, these constellations would not cover all the 'regimes of normativity' or practices of striving for 'good' machine learning. These could be more tacit and arise from certain 'values levers' (Shilton, 2013), or in contexts that work in 'spaces for doubt' (Shklovski & Némethy, 2023), or even break away from routinised concepts and distinctions (Smith et al., 15.05. 2024).

Governance procedures, including data protection regulations and ethical approvals in the medical sector, can also contribute to the friction practitioners experience in working with biomedical data, if or when these procedures fit uneasily with other elements of working relationships, such as project timelines, the use of cloud computing, etc. These findings add to the growing body of studies that show how data protection and other regulatory issues contribute to ML epistemic practices (Engdahl, 2024; Zając et al., 2023) by empirically illustrating these frictions and emphasising that the access arrangements data governance and sharing procedures create are not just barriers to data (re)use in machine learning, but an important element of its epistemics and politics (Thylstrup et al., 2022). These shared requirements, and the work and (potential) challenges they create for ML work, connect practitioners distributed across geographic and social locations. However, this study's findings on both friction and representations of ethics emphasise that these requirements do not reach or affect all practitioners in the same way. Depending on the organisation, the types of projects they work with, local distribution of labour, and other contextual factors, institutionalised ethics and data governance procedures would contribute differently to meanings and practices of ML work.

8.3 ML work and/in the societal context

Demonstrating that relations and shared concepts and narratives go beyond the professional sphere of ML practitioners is an important contribution of this study. The inquiries into both labels and ethics highlight that practitioners take note of the broader societal discussions on technology, which they encounter in traditional and social media as well as conversations with colleagues and people outside their field. Paper 1 frames the findings on how practitioners envision communication with non-specialists and anticipate popular representations of 'AI' when they describe their work in an inquiry centred on the signifiers. However, its findings also speak to the literature on science-society relations and communication, even though my analysis also includes industry specialists. ML practitioners' self-descriptions privilege familiarity and relatability while aiming to be concise; this echoes the findings of previous studies on communication in the interactions of science and publics (Davies, 2008; Felt & Fochler, 2012). Moreover, my findings illustrate those studies' argument that communication constructs identities for science and publics and the relationship between them, as I show that terminology marks boundaries in the construction of technology between 'insiders' and 'outsiders' to knowledge production (MacKenzie, 1998).

These representation dynamics draw on the positive associations with AI, but practitioners also encounter concerns and anxieties about computational technologies. Moreover, practitioners showed that they took notice of the discussions around data sourcing and artificial intelligence at the time and mentioned cases publicised in the media. These findings provide new insights into the discussion about abstraction and the tendency in computational cultures to 'delete the social' (Leigh Star, in Forsythe, 2001) as part of computational cultures (Malazita & and Resetar, 2019) and about societal considerations in science and technology more broadly (Felt, 2017; Sigl et al., 2020).

Encounters with societal discussions and concerns described in the interview narratives indicate that they are not invisible or 'deleted' completely. At the same time, my analysis indicates that these discussions, and society more generally, are kept at arm's length. For instance, paper 2 shows that encounters with societal debates via the media are recognised as encounters with ethics, even if they are mostly framed as occasional rather than part of their work (except for practitioners whose work focused on 'ethics'-related topics and projects). Furthermore, the limited engagement with the 'societal context' is reflected in the applications constellation of ethics. On the one hand, this representation of ethics draws connections between models and their future or potential applications in the world. However, it also assigns ethical significance to the applications (rather than methods) and expresses a limited

control over the ethicality of potential applications. Moreover, this representation of ethics emphasises the avoidance of particularly problematic problems or application fields, while the potential for negative social implications is to be reflected on (e.g., in ethics statements) but is not necessarily addressed. As such, this constellation reflects what Moore (2019) poignantly characterises as 'AI for not bad' rather than 'AI for good'.

Yet, even if the personal and professional roles are separated (Malazita & and Resetar, 2019), I suggest that the non-expert roles practitioners take on in algorithmic systems (as data subjects and users) and their more personal relations to society through their social milieu, who are affected by algorithmic systems, can become points for discussion and intervention for future ethics or responsible research and innovation efforts. These points add to the call to incorporate lived experience in AI ethics (Widder, 2024), and their relational character could complement approaches embracing the feminist STS notion of 'located accountability' (Suchman, 2002; Widder & Nafus, 2023) or the lens of care (Zakharova & Jarke, 2024).

8.4 Limitations of the study

The findings and contributions discussed above are based on the conceptual and methodological approaches that I have adopted as techniques of seeing (Haraway, 1988). Here, I reflect on the limitations of this vantage point.

The first limitation is the object constructed in and by this study: machine learning work. ML work is not an emic notion or collective identity. Rather, it is an analytical construct I employed to capture a community of practice bound by shared knowledge objects (machine learning models) and practices across lines drawn by fields, occupation titles, and organisations. This construction raises questions of whether such a community of practice indeed exists and what this grouping of diverse technologists together does in the analysis. The findings of the dissertation demonstrate the presence of shared narratives and practices, which suggests that, to some degree, connections can be made. However, the distinctions such as those between the 'general' ML and 'applications' fields, reflected in paper 1, make me avoid claims of a unified epistemic culture or singular ML/AI field (Berman et al., 2025). Thus, the main reason for this construction pertains to its analytical affordances. Avoiding titles and identities as organising elements of the research design allowed me to denaturalise labels and examine differently positioned practitioners' relationships to them, thus noting the distinctions mentioned above. Moreover, it afforded inquiry into practitioners' relations with ethics and biomedical data across the research and development supply chain (Widder & Nafus, 2023). However, the use of this analytical construct can be said to stabilise the machine learning category and object, which, as noted by Lee (2025), is just as

constructed. My analysis does not take this construction for granted. For instance, paper 1 shows that, compared to 'AI', the machine learning category is more stable and agreed-upon among the interviewees. The analysis treats this as a code (Hall, 1997), fixed in social convention and conducive to change. The dissertation did not take this investigation further as I directed my analytical and deconstructive attentions elsewhere.

The research design employed in the dissertation also poses limitations. First, as previously noted in the discussion of methods (see Chapter 4), the interviewing method was best suited to my interest in meaning-making and potentially shared culture. Yet it has limitations due to the distance from practice (Soss, 2006) and tendency to 'lead to an image of individual selves as more coherent, with less contradiction and unpredictability than real lives normally encompass' (Lamont & Swidler, 2014, p. 163). This limitation is arguably most significant for paper 3, as engagement with the data settings (Loukissas, 2019) in which ML practitioners work with biomedical data could produce additional insights. However, the interviewing method allowed me to consider more contexts and instances of ML work, and thus note, for example, the different degrees of friction that sleep-data labels posed for practitioners in different projects. Secondly, this dissertation presents a 'nested' study; its three themes are explored based on the same dataset, generated with the same sampling approach and the same guide. This design shaped the holistic and relational framing of the dissertation and the insights my analysis produced. On the other hand, individual studies into the three 'others' could yield more information on each individual subject. As all data were collected in spring-summer 2024, this study also produces a 'framing effect' of time (Hyysalo et al., 2019). In paper 1, I address the changeability of meanings with a representation theory that considers codes as conducive to change (Hall, 1997) and by situating the exploration of the contemporary dynamics in the longer interrelated history of ML and AI as discursive constructions (Galanos, 2023). Third, this inquiry has often followed language: both in analysing practitioners' relationships with labels, and their understandings of 'ethics'. While this way of directing analytic attention led to interesting findings, the focus on language and representation might overemphasise the importance of specific categories (e.g., 'ethics'), as well as the role of discourse more generally, in sociotechnical systems. Material and infrastructural aspects of machine learning are no less important, and my inquiry into friction in paper 3 reflects my desire to better account for the joining of imaginaries and materialities (Suchman, 2012, p. 48).

Chapter 4 addresses the limitations of recruitment choices and the contingencies of access for the study. Hence, the last limitation to note here is the lack of comparative analysis, which could have potentially disappointed a

reader excited by, for instance, the data collected across three countries. While the study noted the implications of particular social locations (such as constellations of ethics appearing in certain types of ML work), I did not aim to trace the meanings and experiences of ML practitioners to individual explanans such as national context, field or sector of work, organisation, gender, etc. Not pursuing comparative analysis is a choice rooted in the research design, which treated sampling across three countries as accessing an international and highly mobile community of practice rather than as cases for comparison, and is intertwined with my relational analytical approach. However, the appearances of these explanans in the negotiated codes depicted in paper 1 (for instance, a PhD student mentions Danish norms against promoting oneself and the discussion of 'cocky' labelling on invokes gender dynamics) could serve as an inspiration for future research.

8.5 Directions for future research

The findings of this dissertation also point to at least two potential avenues of further inquiry. As my analysis explored the connection between ML practitioners' situated meaning-making about their work and shared cultural concepts and narratives, it encourages further investigation of the sources of these concepts and narratives. Amoore et al. propose engaging with ML texts such as textbooks and influential papers as 'sites where concepts are forged and contested' (Amoore et al., 2023, p. 1); my analyses point to other potentially important sites. In paper 2, I link one of the ethics constellations to ML conferences' ethical statements and checklists as its narrative infrastructure. Further research can focus on the narratives constructed by conferences and other 'field-configuring events' (Lampel & Meyer, 2008). For instance, previous studies have pointed to the influence of benchmarking competitions in directing development (Luitse et al., 2025). Moreover, directing attention to narratives produced by conferences like the ACM Conference on Fairness, Accountability, and Transparency or the AAAI Conference on AI, Ethics, and Society could provide new insights into how the AI ethics field co-produces its object (Helm & Gerlek, 2026).

A related avenue for inquiry refers to '[w]hat is read in computer science' (Amoore et al., 2023, p. 2) in the everyday. Paper 2 notes that reading to keep up with new developments is part of ML work in both academia and industry. In addition to academic papers, this involves reading newsletters, blogs, and social media communications by researchers who intensively communicate their work across audiences and platforms (Gallagher et al., 2024) and influential users such as tech managers like Sam Altman, 'godfather' research figures like Geoffrey Hinton, etc. Engaging with these texts (in their multimodality), which connect distributed communities, could provide further insights

into computational cultures. Taking inspiration from the joke featured in paper 1 and Ziewitz's (2019) findings on the role of humour in the negotiations of ethicality in the search engine optimisation industry, future studies can turn their attention particularly to humour as a window into meaning-making and normativities among technologists.

A different direction for future research stems from the dissertation's findings on how data governance and data sharing between organisations contribute to (parts of) ML practice. In addition to ethnographic or interview studies into how navigating data governance fits into ML knowledge practices of different kinds, further research can also focus on data governance as a point of interaction between ML work and other forms of expertise. Previous critical research has produced fruitful insights on data annotation and the relationship between annotators and ML practitioners (e.g., Bennett et al., 2025; Miceli et al., 2022; Wirth et al., 2025). Another stream of studies has focused on the role and experiences of 'domain experts' in computational projects (Diaz & Smith, 2024; Sambasivan et al., 2021; Sambasivan & Veeraraghavan, 2022). A similar focus on data stewards, compliance specialists, and/or data sharing officers could also yield fruitful insights. A promising example of such inquiry is the study by Thakkar et al. (2022), who traced how different actors in the ML production chain value, including data stewards, strive to produce high-quality data in population health projects. Further studies on the working relationships between ML practitioners and data stewards and similar specialists would shed light on this aspect of ML work and its relations with other forms of knowledge and ways of dealing with data, and thus potentially friction (Edwards et al., 2011).

8.6 Conclusion

The interweaving of machine learning systems into the fabric of our daily lives, amid the contradictory valence of excitement and concern about artificial intelligence, has made it a fascinating case for this doctoral research. On the one hand, these technologies are 'hyped' and mystified; thus, social studies need to elucidate their mundane work and practice (Lee, 2025). On the other hand, researchers are warned against the naturalisation of technological categories and urged to 'make computation strange again' (Davies & Díaz-Valderrama, 2026; Lee & Ribes, 2025). In this dissertation, I explored how machine learning practitioners think and talk about their work, focusing on labels, ethics, and experiences with biomedical data. In this study, I aimed to capture both mundane practices of machine learning (and frictions that can arise there) and practitioners' experiences of being the 'popular topic' in societal discussions. One of the findings of this dissertation is indeed that these relations and interactions with broader society seep into how practitioners make sense of their

work. Overall, the study emphasises that machine learning work is comprised of bundles of relations, some of which are highly local while others spread far and wide across the world. This relational thinking is a way of de-essentialising categories, objects, and distinctions one risks taking for granted, and of gaining a better understanding of sociotechnical systems.

Summary

In recent years, machine learning and artificial intelligence (AI) have attracted unprecedented attention in academic and media discussions, as well as in everyday conversations. Some discussions look to the future risks and potential of these technologies; others focus on existing applications in research and business analytics, risk prediction, content generation, and more. Academics and activists have also drawn attention to the harmful aspects of these technologies, such as environmental costs of computation, labour exploitation, and biased predictions. In response, both research organisations and businesses have launched initiatives to make AI more ethical and socially responsible. Legislators have also turned their attention to data and AI with the General Data Protection Regulation (GDPR) and the AI Act. This dissertation explores the culture and experiences of working in machine learning amid this societal and regulatory attention through interviews with researchers, data scientists, and engineers working in Denmark, the Netherlands, and the United Kingdom. It addresses *how machine learning work is co-constituted by its relations with the labels 'AI' and 'machine learning', ethics, and (biomedical) data*. The study's framing emphasises that none of these phenomena are concrete, independent 'things', but acquire their meanings in relation to one another.

First, this research provides insight into practitioners' relationship with the 'AI' label. They see 'AI' as an imperfect but useful descriptor for communicating with people outside their field. Practitioners experience and anticipate that non-specialists associate 'AI' with large language models (such as ChatGPT). While not everyone finds this representation accurate, practitioners accept and use it, as it makes their work more relatable than 'machine learning' and other more 'technical' descriptions. In this way, the 'AI' term can mark the distinction between specialists and non-specialists. However, the analysis also identifies more ambivalent connotations of 'AI' and some connections between popular and professional representations.

The second important finding concerns how practitioners understand the notion of 'ethics' and how it fits into their work. From the interviews, the analysis identifies three representations of ethics: 1) ethics of applications, 2) ethics of data, and 3) ethics as a research topic. The first portrays machine learning models' (potential) applications as the main ethical consideration, often focusing on the potential misuse of these models for undesirable purposes. Practitioners also articulate the potential (negative) impacts of their research when submitting to conferences. Ethics of applications apply to entire topics

or research areas, while ethics of data operate at the level of specific practices. Finally, research on topics such as fairness and explainability constitutes a third understanding of ethics in machine learning. These representations highlight that some issues and practices are more central to how practitioners understand ethics than others. Moreover, professionals also encounter ethics outside their professional roles when engaging with traditional or social media content about the societal implications of technology and in conversations with people around them.

Furthermore, the study examines practitioners' experiences of working with biomedical data. Accessing biomedical data and dealing with its features can present challenges for machine learning practitioners. The analysis then argues that these challenges, or *frictions*, are not caused by the inherent properties of the biomedical data or the clash between machine learning and biomedical cultures of knowledge production. Rather, they come from tensions within overall working relationships which include many different elements: data governance procedures and data infrastructures, project timelines, data features, established practices and tacit assumptions of machine learning work (for example, how difficult the task is supposed to be, or how much difference certain data properties make for the training process), and more. This also suggests that data access should be considered an ongoing dimension of working relationships that shapes the results of machine learning research and development, rather than just an initial barrier to data use.

In summary, this dissertation demonstrates the interplay of narratives, practices, and materialities that co-constitute machine learning work as a whole, as well as its particular aspects such as labelling, ethics, and biomedical data. It highlights that social research about machine learning needs to consider how it is shaped by both situated practice and broader cultural and institutional contexts. Regarding the latter, it demonstrates how governance mechanisms and relations to broader society contribute to the ways practitioners understand and experience their work. These relations matter for the produced technologies and need to be considered in future dialogue and promotion of responsible research and technology development.

Dansk resumé

I de senere år har maskinlæring og kunstig intelligens (AI) tiltrukket sig en hidtil uset opmærksomhed i akademiske debatter, i medierne og i hverdagens samtaler. Nogle diskussioner retter sig mod disse teknologiers fremtidige risici og potentialer, mens andre fokuserer på deres eksisterende anvendelser inden for forskning og forretningsanalyse, risikoforudsigelse, indholdsgenerering med mere. Forskere og aktivister har desuden henledt opmærksomheden på teknologiernes skadelige aspekter, herunder de miljømæssige omkostninger ved databehandling, udnyttelse af arbejdskraft og skævvredne forudsigelser. Som reaktion herpå har både forskningsinstitutioner og virksomheder iværksat initiativer, der skal gøre AI mere etisk og socialt ansvarlig. Lovgivere har ligeledes rettet opmærksomheden mod data og AI gennem databeskyttelsesforordningen (GDPR) og AI-forordningen (AI Act). Gennem interviews med forskere, data scientists og ingeniører, der arbejder i Danmark, Holland og Storbritannien, undersøger denne afhandling kulturen omkring og erfaringerne med at arbejde med maskinlæring i en tid præget af omfattende samfundsmæssig og regulatorisk opmærksomhed. Afhandlingen undersøger herunder specifikt, *hvordan arbejdet med maskinlæring medkonstitueres gennem dets relationer til betegnelserne 'AI' og 'maskinlæring', til etik og til (biomedicinske) data*. Undersøgelsens analytiske ramme understreger, at ingen af disse fænomener er konkrete og uafhængige 'ting', men at de får deres betydning i relation til hinanden.

Undersøgelsen giver for det første indsigt i maskinlærings-praktikernes forhold til betegnelsen 'AI'. De betragter 'AI' som en upræcis, men nyttig betegnelse, når de kommunikerer med personer uden for deres eget fagområde. Praktikerne oplever og forventer, at ikke-specialister forbinder 'AI' med store sprogmodeller (som ChatGPT). Selvom ikke alle opfatter denne fremstilling som retvisende, accepterer og anvender praktikerne den, fordi den gør deres arbejde lettere at relatere til end betegnelsen 'maskinlæring' og andre mere tekniske beskrivelser. På denne måde kan betegnelsen 'AI' markere en skelnen mellem specialister og ikke-specialister. Analysen identificerer imidlertid også mere ambivalente konnotationer knyttet til 'AI' samt visse forbindelser mellem populære og professionelle fremstillinger af teknologien.

Det andet væsentlige resultat vedrører maskinlærings-praktikernes forståelse af begrebet 'etik', og hvordan etik indgår i deres arbejde. På baggrund af interviewene identificerer analysen tre repræsentationer af etik: 1) anvendelsesetik, 2) dataetik og 3) etik som forskningsemne. Den første repræsentation udpeger maskinlæringsmodellers mulige anvendelser som det primære

etiske hensyn og fokuserer ofte på risikoen for, at modellerne misbruges til uønskede formål. Praktikerne forholder sig eksempelvis til de potentielle (negative) konsekvenser af deres forskning, når de indsender bidrag til konferencer. Anvendelsesetik knytter sig til hele emner eller forskningsområder, mens dataetik gør sig gældende på niveauet for konkrete praksisser. Endelig udgør forskning i emner som fairness og forklarlighed en tredje forståelse af etik inden for maskinlæring. Disse repræsentationer viser, at visse problemstillinger og praksisser står mere centralt end andre i praktikernes forståelse af etik. Fagpersonerne møder desuden etiske spørgsmål uden for deres professionelle roller, eksempelvis når de beskæftiger sig med indhold i traditionelle og sociale medier om teknologiens samfundsmæssige konsekvenser, og når de taler med mennesker i deres omgangskreds.

Undersøgelsen analyserer endvidere praktikernes erfaringer med at arbejde med biomedicinske data. Det kan være udfordrende for maskinlæringspraktikere at få adgang til biomedicinske data samt håndtere deres særlige karakteristika. Analysen argumenterer imidlertid for, at disse udfordringer eller *friktioner* hverken skyldes biomedicinske datas iboende egenskaber eller et sammenstød mellem maskinlæringens og biomedicinens kulturer for vidensproduktion. De udspringer snarere af spændinger i de samlede arbejdsrelationer, som omfatter en lang række forskellige elementer: procedurer for dataforvaltning og datainfrastruktur, projekttidsplaner, dataegenskaber, etablerede praksisser og uudtalte antagelser i arbejdet med maskinlæring, eksempelvis antagelser om, hvor vanskelig en opgave forventes at være, eller hvor stor betydning bestemte dataegenskaber har for træningsprocessen. Dette peger samtidig på, at dataadgang bør betragtes som en løbende dimension af arbejdsrelationer, der er med til at forme resultaterne af forskning og udvikling inden for maskinlæring, snarere end blot som en indledende hindring for anvendelsen af data.

Sammenfattende viser afhandlingen, hvordan fortællinger, praksisser og materialiteter spiller sammen og medkonstituerer arbejdet med maskinlæring som helhed såvel som dets specifikke aspekter, herunder betegnelser, etik og biomedicinske data. Afhandlingen fremhæver, at samfundsvidenskabelig forskning i maskinlæring må tage højde for, hvordan området formes af både situerede praksisser og bredere kulturelle og institutionelle kontekster. Hvad angår sidstnævnte, viser afhandlingen, hvordan styringsmekanismer og relationer til det omgivende samfund bidrager til den måde, hvorpå praktikerne forstår og oplever deres arbejde. Disse relationer har betydning for de teknologier, der udvikles, og bør derfor indgå i den fremtidige dialog om og fremme af ansvarlig forskning og teknologiudvikling.

Bibliography

Abend, G. (2014). *The Moral Background*. Princeton University Press. https://doi.org/doi:10.1515/9781400850341

The AI Resist List. https://airesistlist.org

Al Kuwaiti, A., Nazer, K., Al-Reedy, A., Al-Shehri, S., Al-Muhanna, A., Subbarayalu, A. V., Al Muhanna, D., & Al-Muhanna, F. A. (2023). A review of the role of artificial intelligence in healthcare. *Journal of Personalized medicine*, *13*(6), 951. https://doi.org/https://doi.org/10.3390/jpm13060951

American Psychological Association. (2020). *Concise guide to APA Style (7th ed.)*. https://doi.org/10.1037/0000173-000

Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others*. Duke University Press. https://doi.org/10.1515/9781478009276

Amoore, L. (2024). The deep border. *Political Geography*, *109*, 102547. https://doi.org/https://doi.org/10.1016/j.polgeo.2021.102547

Amoore, L., Campolo, A., Jacobsen, B., & Rella, L. (2023). Machine learning, meaning making: On reading computer science texts. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231166887

Andrejevic, M. (2020). Data civics: A response to the "ethical turn". *Television & New Media*, *21*(6), 562-567. https://doi.org/10.1177/1527476420919693

Antoft, R., & Salomonsen, H. H. (2007). *Studying organizations by a pragmatic research design: The case of qualitative case study designs* European Group of Organization Studies "Beyond Waltz - Dances of Individuals and Organizations", Austria. https://vbn.aau.dk/ws/portalfiles/portal/12015316/Pragmatisme.pdf

Ashurst, C., Hine, E., Sedille, P., & Carlier, A. (2022). AI ethics statements: Analysis and lessons learnt from NeurIPS broader impact statements. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2047–2056. https://doi.org/10.1145/3531146.3533780

Aula, V. (2019). Institutions, infrastructures, and data friction – Reforming secondary use of health data in Finland. *Big Data & Society*, *6*(2). https://doi.org/10.1177/2053951719875980

Avnoon, N. (2021). Data scientists' identity work: Omnivorous symbolic boundaries in skills acquisition. *Work, Employment and Society*, *35*(2), 332-349. https://doi.org/10.1177/0950017020977306

Avnoon, N. (2024). The gates to the profession are open: The alternative institutionalization of data science. *Theory and Society*, *53*(2), 239-271. https://doi.org/10.1007/s11186-023-09529-0

Avnoon, N., Kotliar, D. M., & Rivnai-Bahir, S. (2024). Contextualizing the ethics of algorithms: A socio-professional approach. *New Media & Society*, *26*(10), 5962-5982. https://doi.org/10.1177/14614448221145728

Bareis, J., & Katzenbach, C. (2022). Talking AI into being: The narratives and imaginaries of national AI strategies and their performative politics. *Science,*

Technology, & Human Values, *47*(5), 855-881. https://doi.org/https://doi.org/10.1177/01622439211030007
Barocas, S., & boyd, d. (2017). Engaging the ethics of data science in practice. *Communications of the ACM*, *60*(11), 23-25. https://doi.org/https://doi.org/10.1145/3144172
Bassett, C., & Roberts, B. (2023). Automation anxiety: A critical history-the apparently odd recurrence of debates about computation, AI and labour. In S. Lindgren (Ed.), *Handbook of Critical Studies of Artificial Intelligence* (pp. 79-93). Edward Elgar Publishing. https://doi.org/https://doi.org/10.4337/9781803928562.00012
Bassett, D. R. (2011). Notions of identity, society, and rhetoric in a speech code of science among scientists and engineers working in nanotechnology. *Science Communication*, *34*(1), 115-159. https://doi.org/10.1177/1075547011417891
Bates, J. (2017). The politics of data friction. *Journal of Documentation*, *74*(2), 412-429. https://doi.org/10.1108/jd-05-2017-0080
Bates, J., Lin, Y.-W., & Goodale, P. (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society*, *3*(2), 2053951716654502. https://doi.org/10.1177/2053951716654502
Bechmann, A., & Bowker, G. C. (2019). Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society*, *6*(1). https://doi.org/10.1177/2053951718819569
Bengesi, S., El-Sayed, H., Sarker, M. K., Houkpati, Y., Irungu, J., & Oladunni, T. (2024). Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. *IEEE Access*, *12*, 69812-69837. https://doi.org/10.1109/ACCESS.2024.339777
Bennett, S. (2025). Artificial Intelligence and the ethics of navigating ambiguity. *Big Data & Society*, *12*(2). https://doi.org/https://doi.org/10.1177/20539517251347594
Bennett, S. J., Catanzariti, B., & Tollon, F. (2025). "Everybody knows what a pothole is": Representations of work and intelligence in AI practice and governance. *AI & Society*, *40*, 3283–3294. https://doi.org/https://doi.org/10.1007/s00146-024-02162-0
Bensaude Vincent, B. (2014). The politics of buzzwords at the interface of technoscience, market and society: The case of 'public engagement in science'. *Public Understanding of Science*, *23*(3), 238-253. https://doi.org/10.1177/0963662513515371
Berman, G., Williams, K., & Cohen, E. (2025). The benefits of being between (many) fields: Mapping the high-dimensional space of AI research. *Big Data & Society*, *12*(1). https://doi.org/10.1177/20539517241306355
Bernard, H. R. (2017). *Research methods in anthropology: Qualitative and quantitative approaches*. Bloomsbury Publishing.
Bevir, M. (2006). How narratives explain. In D. Yanow & P. Schwartz-Shea (Eds.), *Interpretation and method: Empirical research methods and the interpretive turn* (pp. 281-290). M.E. Sharpe.

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

Bijker, W. E., & Law, J. (1992). *Shaping technology/building society: Studies in sociotechnical change*. The MIT Press. https://mitpress.mit.edu/9780262521949/shaping-technology-building-society/

Bonde, M., Bossen, C., & Danholt, P. (2019). Data-work and friction: Investigating the practices of repurposing healthcare data. *Health Informatics Journal*, *25*(3), 558-566. https://doi.org/10.1177/1460458219856462

Bonde Thylstrup, N., Flyverbom, M., & Helles, R. (2019). Datafied knowledge production: Introduction to the special theme. *Big Data & Society*, *6*(2). https://doi.org/10.1177/2053951719875985

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world* (1 ed.). MIT Press. https://doi.org/10.7551/mitpress/9963.001.0001

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2010). Toward information infrastructure studies: Ways of knowing in a networked environment. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International Handbook of Internet Research* (pp. 97-117). Springer Netherlands. https://doi.org/10.1007/978-1-4020-9789-8_5

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press. https://doi.org/https://doi.org/10.7551/mitpress/6352.001.0001

Brause, S. R., Zeng, J., Schäfer, M. S., & Katzenbach, C. (2023). Media representations of artificial intelligence: Surveying the field. In S. Lindgren (Ed.), *Handbook of Critical Studies of Artificial Intelligence* (pp. 277-288). https://doi.org/10.4337/9781803928562.00030

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199-231. https://doi.org/10.1214/ss/1009213726

Calvert, J. (2006). What's special about basic research? *Science, Technology, & Human Values*, *31*(2), 199-220. https://doi.org/10.1177/0162243905283642

Cardon, D., Cointet, J.-P., Mazières, A., & Carey-Libbrecht, L. (2018). Neurons spike back. *Réseaux*, *211*(5), 173-220. https://doi.org/https://doi.org/10.3917/res.211.0173

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE.

Clarke, A. E., & Star, S. L. (2008). The social worlds framework: A theory/methods package. In E. J. Hackett, O. Amsterdamska, M. Lynch, & J. Wajcman (Eds.), *The handbook of science and technology studies* (Vol. 3, pp. 113-137). The MIT Press. https://www.dhi.ac.uk/san/waysofbeing/data/data-crone-wyatt-2007b.pdf

Collins, H. (1990). *Artificial experts: Social knowledge and intelligent machines*. The MIT Press. https://doi.org/https://doi.org/10.7551/mitpress/1416.001.0001
Collins, H., & Evans, R. (2007). Rethinking expertise. In. University of Chicago Press. https://doi.org/https://doi.org/10.7208/9780226113623
Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence* (1 ed.). Yale University Press. https://doi.org/10.2307/j.ctv1ghv45t
Crawford, K., & Joler, V. (2023). *Calculating empires: A genealogy of technology and power since 1500*. https://calculatingempires.net
D'ignazio, C., & Klein, L. F. (2020). *Data feminism*. The MIT press. https://doi.org/https://doi.org/10.7551/mitpress/11805.001.0001
Davenport, T. H., & Patil, D. (2012, October). Data scientist. *Harvard business review*, *90*(5), 70-76. https://blogs.sun.ac.za/open-day/files/2022/03/Data-Scientist-Harvard-review.pdf
Davies, S. R. (2008). Constructing communication:Talking to scientists about talking to the public. *Science Communication*, *29*(4), 413-434. https://doi.org/10.1177/1075547008316222
Davies, S. R., & Díaz-Valderrama, M. (2026). Attending to 'computational universalism': Practices, frictions, events. *Social Studies of Science*, *56*(3), 340-352. https://doi.org/10.1177/03063127261445803
Davison, J. (2018). No, machine learning is not just glorified statistics. *Medium-Toward-Data-Science. com*. https://medium.com/data-science/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3
Delacroix, S., & Wagner, B. (2021). Constructing a mutually supportive interface between ethics and regulation. *Computer Law & Security Review*, *40*, 105520. https://doi.org/https://doi.org/10.1016/j.clsr.2020.105520
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248-255. https://doi.org/https://doi.org/10.1109/CVPR.2009.5206848
Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, *8*(2). https://doi.org/10.1177/20539517211035955
Deterding, N. M., & Waters, M. C. (2021). Flexible coding of in-depth interviews: A twenty-first-century approach. *Sociological Methods & Research*, *50*(2), 708-739. https://doi.org/10.1177/0049124118799377
Diaz, M., & Smith, A. D. R. (2024). What makes an expert? Reviewing how ML researchers define "expert". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*(1), 358-370. https://doi.org/10.1609/aies.v7i1.31642
Dobson, J. E. (2023). *The birth of computer vision*. University of Minnesota Press. https://doi.org/https://doi.org/10.5749/9781452968889
Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, *3*(2). https://doi.org/10.1177/2053951716665128

Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly*, *5*(1). https://doi.org/https://doi.org/10.63744/r4ysrh7ae534

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., & Morgan, G. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM computing surveys*, *55*(9), 1-33. https://doi.org/https://doi.org/10.1145/3561048

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. The MIT Press.

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., Burton, M., & Calvert, S. (2013). *Knowledge infrastructures: Intellectual frameworks and research challenges*.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667-690. https://doi.org/10.1177/0306312711413314

Elish, M. C., & boyd, d. (2018). Situating methods in the magic of Big Data and AI. *Communication Monographs*, *85*(1), 57-80. https://doi.org/10.1080/03637751.2017.1375130

Engdahl, I. (2024). Wuthering weights—Localisation trajectories of machine learning models for local ends. In V. Fors, M. Berg, & M. Brodersen (Eds.), *The De Gruyter handbook of automated futures: Imaginaries, interactions and impact* (pp. 321-336). De Gruyter. https://doi.org/https://doi.org/10.1515/9783110792256-020

Felt, U. (2017). "Response-able practices" or "new bureaucracies of virtue": The challenges of making RRI work in academic environments. In L. Asveld, R. van Dam-Mieras, T. Swierstra, S. Lavrijssen, K. Linse, & J. van den Hoven (Eds.), *Responsible Innovation 3: A European Agenda?* (pp. 49-68). Springer International Publishing. https://doi.org/10.1007/978-3-319-64834-7_4

Felt, U. (2025). *Academic times: Contesting the chronopolitics of research*. Springer. https://doi.org/https://doi.org/10.1007/978-981-96-4609-8

Felt, U., & Fochler, M. (2012). Re-ordering epistemic living spaces: On the tacit governance effects of the public communication of science. In S. Rödder, M. Franzen, & P. Weingart (Eds.), *The Sciences' Media Connection –Public Communication and its Repercussions* (pp. 133-154). Springer Netherlands. https://doi.org/10.1007/978-94-007-2085-5_7

Flick, U. (2009). *An introduction to qualitative research* (4th ed.). Sage.

Forsythe, D. E. (1993). Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence. *Social Studies of Science*, *23*(3), 445-477. https://doi.org/10.1177/0306312793023003002

Forsythe, D. E. (2001). *Studying those who study us: An anthropologist in the world of artificial intelligence*. Stanford University Press. https://doi.org/10.1515/9781503619371

Frahm, N., & Schiølin, K. (2024). The rise of tech ethics: Approaches, critique, and future pathways. *Science and Engineering Ethics*, *30*(5), 45. https://doi.org/10.1007/s11948-024-00510-3

Franzke, A. S. (2022). An exploratory qualitative analysis of AI ethics guidelines. *Journal of Information, Communication and Ethics in Society*, *20*(4), 401-423. https://doi.org/https://doi.org/10.1108/JICES-12-2020-0125

Frith, L., Jacoby, A., & Gabbay, M. (2011). Ethical boundary-work in the infertility clinic. *Sociology of Health & Illness*, *33*(4), 570-585. https://doi.org/https://doi.org/10.1111/j.1467-9566.2010.01308.x

Fujimura, J. H. (1987). Constructing 'do-able' problems in cancer research: Articulating alignment. *Social Studies of Science*, *17*(2), 257-293. https://doi.org/10.1177/030631287017002003

Galanos, V. (2023, 21.11.2023). *To have done with the metaphor of summers and winters: Can AI and Internet history cure hype?* http://dx.doi.org/10.2139/ssrn.4640305

Galanos, V., Gonçalves, A. B., Alvial-Palavicino, C., Bareis, J., Blackwell, B., Charette, M., Cheng, M. Y., Dorofeeva, O., Gloerich, I., & Klingebiel, J. (2026). Towards Mapping and Defining critical hype studies: Multidisciplinary insights and future directions. https://doi.org/10.5281/zenodo.18764034

Galanos, V., & Stewart, J. K. (2024). Navigating AI beyond hypes, horrors and hopes: Historical and contemporary perspectives. In A. Ponce del Castillo (Ed.), *Artificial intelligence, labour and society* (pp. 27-46). ETUI. https://www.etui.org/sites/default/files/2024-03/Chapter2_Navigating%20AI%20beyond%20hypes%2C%20horrors%20and%20hopes%20historical%20and%20contemporary%20perspectives.pdf

Gallagher, J. R., Avgoustopoulos, R. E., Hamilton, A., & Seilkhanova, T. (2024). The challenges and writing practices of communicating artificial intelligence and machine learning in an era of hype. *AI & Society*, *40*, 2895-2908. https://doi.org/10.1007/s00146-024-02049-0

Gieryn, T. F. (1983). Boundary-work and the demarcation of science from non-science: strains and interests in professional ideologies of scientists. *American sociological review*, *48*(6), 781-795. https://doi.org/https://doi.org/10.2307/2095325

Gieryn, T. F. (1999). *Cultural boundaries of science: Credibility on the line*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226824420

Goñi, J. I., Eugenia, R., Jesus, P. M., Martín, I., & and Millán, M. J. (2024). Tooling with ethics in technology: A scoping review of responsible research and innovation tools. *Journal of Responsible Innovation*, *11*(1). https://doi.org/10.1080/23299460.2024.2360228

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Graham, R. (2022). Discourse analysis of academic debate of ethics for AGI. *AI & Society*, *37*(4), 1519-1532. https://doi.org/10.1007/s00146-021-01228-7

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt.
Greene, D., Hoffmann, A. L., & 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*. https://doi.org/https://doi.org/10.24251/HICSS.2019.258
Grodal, S. (2017). Field expansion and contraction: How communities shape social and symbolic boundaries. *Administrative Science Quarterly*, *63*(4), 783-818. https://doi.org/10.1177/0001839217744555
Guest, G., MacQueen, K. M., & Namey, E. E. (2012). *Applied thematic analysis*. Sage.
Hacking, I. (1990). *The taming of chance*. Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9780511819766
Hall, S. (1973). Encoding and decoding in the television discourse. In D. Morley (Ed.), *Essential Essays, Volume 1: Foundations of Cultural Studies* (pp. 257–276). Duke University Press. https://doi-org.ez.statsbiblioteket.dk/10.2307/j.ctv11cw7c7.16
Hall, S. (1997). The work of representation. In *Representation: cultural representations and signifying practices*. Sage in association with the Open University. https://us2.sagepub.com/sites/default/files/upm-binaries/55352_Hall_ch_1.pdf
Hall, S. (2019). Encoding—decoding (1980). In *Crime and media* (pp. 44-55). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9780367809195-6/encoding—decoding-1980-stuart-hall
Hansen, K. B. (2020). The virtue of simplicity: On machine learning models in algorithmic trading. *Big Data & Society*, *7*(1).
Hansen, K. B., & Borch, C. (2021). The absorption and multiplication of uncertainty in machine-learning-driven finance. *The British Journal of Sociology*, *72*(4), 1015-1029. https://doi.org/https://doi.org/10.1111/1468-4446.12880
Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, *14*(3), 575-599. https://doi.org/https://doi.org/10.2307/3178066
Haraway, D. (1992). The promises of monsters: a regenerative politics for inappropriate/d others. *Cultural Studies*, 295-337. https://pages.mtu.edu/~jdslack/readings/CSReadings/Haraway_Promises_of_Monsters.pdf
Hartley, J. M., & Thylstrup, N. B. (2025). The algorithmic gut feeling – Articulating journalistic doxa and emerging epistemic frictions in AI-driven data work. *Digital Journalism*, *13*(3), 438-456. https://doi.org/10.1080/21670811.2024.2319641
Heimer, C. A. (2013). 'Wicked'ethics: Compliance work and the practice of ethics in HIV research. *Social Science & Medicine*, *98*, 371-378. https://doi.org/https://doi.org/10.1016/j.socscimed.2012.10.030

Helm, P., & Gerlek, S. (2026). Empirical AI ethics: Grounded, relational, transformative. *Cambridge Forum on AI: Culture and Society*, *2*, Article e13. https://doi.org/10.1017/cfc.2026.10023
Henriksen, A., & Bechmann, A. (2020). Building truths in AI: Making predictive algorithms doable in healthcare. *Information, Communication & Society*, *23*(6), 802-816. https://doi.org/10.1080/1369118X.2020.1751866
Hoeyer, K. (2006). "Ethics wars": Reflections on the antagonism between bioethicists and social science observers of biomedicine. *Human Studies*, *29*, 203-227. https://doi.org/https://doi.org/10.1007/s10746-006-9022-9
Hoeyer, K. (2023). *Data paradoxes: The politics of intensified data sourcing in contemporary healthcare* (1st ed.). The MIT Press. https://doi.org/https://doi.org/10.7551/mitpress/14926.001.0001
Hoeyer, K. L., & Tutton, R. (2005). 'Ethics was here': Studying the language-games of ethics in the case of UK Biobank. *Critical Public Health*, *15*(4), 385-397. https://doi.org/10.1080/09581590500523533
Horst, M. (2013). A field of expertise, the organization, or science itself? Scientists' perception of representing research in public communication. *Science Communication*, *35*(6), 758-779. https://doi.org/10.1177/1075547013487513
Hyysalo, S., Pollock, N., & Williams, R. (2019). Method matters in the social study of technology: Investigating the biographies of artifacts and practices. *Science & Technology Studies*, *32*(3), 2-25. https://doi.org/10.23987/sts.65532
Ienca, M., Ferretti, A., Hurst, S., Puhan, M., Lovis, C., & Vayena, E. (2018). Considerations for ethics review of big data health research: A scoping review. *PloS one*, *13*(10), e0204937-e0204937. https://doi.org/10.1371/journal.pone.0204937
Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, *3*(2). https://doi.org/https://doi.org/10.1177/2053951716674238
Inie, N., Zukerman, P., & Bender, E. M. (2026). De-anthropomorphizing "AI": From wishful mnemonics to accurate nomenclature. *First Monday*, *31*(2). https://doi.org/10.5210/fm.v31i2.14366
Jaton, F. (2017). We get the algorithms of our ground truths: Designing referential databases in digital image processing. *Social Studies of Science*, *47*(6), 811-840. https://doi.org/10.1177/0306312717730428
Jaton, F. (2021). *The constitution of algorithms: Ground-truthing, programming, formulating*. MIT Press. https://doi.org/https://doi.org/10.7551/mitpress/12517.001.0001
Jaton, F. (2025). Examining algorithms in the light of their ground truth datasets: Results, objections, and avenues of reflection. *Digital Society*, *4*(2), 61. https://doi.org/10.1007/s44206-025-00197-4
Jerolmack, C., & Khan, S. (2014). Talk is cheap: Ethnography and the attitudinal fallacy. *Sociological Methods & Research*, *43*(2), 178-209. https://doi.org/https://doi.org/10.1177/0049124114523396
Jiménez-Sánchez, A., Avlona, N.-R., Juodelyte, D., Sourget, T., Vang-Larsen, C., Rogers, A., Zając, H. D., & Cheplygina, V. (2024). Copycats: the many lives of a

publicly available medical imaging dataset. *Advances in Neural Information Processing Systems*, *37*, 113383-113404. https://doi.org/https://doi.org/10.52202/079017-3602
Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399. https://doi.org/https://doi.org/10.1038/s42256-019-0088-2
Jones-Imhotep, E. (2020). The ghost factories: histories of automata and artificial life. *History and Technology*, *36*(1), 3-29. https://doi.org/10.1080/07341512.2020.1757972
Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S. G., Noble, S. U., & Shestakofsky, B. (2021). Toward a sociology of artificial intelligence: A call for research on inequalities and structural change. *Socius: Sociological Research for a Dynamic World*, *7*. https://doi.org/10.1177/2378023121999581
Kajava, K., & Sawhney, N. (2023). Language of algorithms: Agency, metaphors, and deliberations in AI discourses. In S. Lindgren (Ed.), *Handbook of Critical Studies of Artificial Intelligence* (pp. 224-236). Edward Elgar Publishing. https://doi.org/10.4337/9781803928562.00025
Kerr, A., Barry, M., & Kelleher, J. D. (2020). Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720915939
Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences* (1st ed.). SAGE Publications. https://doi.org/10.4135/9781473909472
Kitchin, R. (2022). *The data revolution : a critical analysis of big data, open data & data infrastructures* (2nd ed.). Sage Publications.
Kitzinger, C., & Mandelbaum, J. (2013). Word selection and social identities in talk-in-interaction. *Communication Monographs*, *80*(2), 176-198. https://doi.org/10.1080/03637751.2013.776171
Knorr Cetina, K. (2007). Culture in global knowledge societies: knowledge cultures and epistemic cultures. *Interdisciplinary science reviews*, *32*(4), 361-375. https://doi.org/10.1179/030801807X163571
Knorr, K. D. (1979). Tinkering toward success: Prelude to a theory of scientific practice. *Theory and Society*, *8*(3), 347-376. http://www.jstor.org/stable/656764
Krarup, T., & Horst, M. (2023). European artificial intelligence policy as digital single market making. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231153811
Lammar, D., Horst, M., & Müller, R. (2025). AI in the german media: Narratives of AI-in-particular and AI-in-general in german media reporting about artificial intelligence. *Digital Journalism*, 1-19. https://doi.org/https://doi.org/10.1080/21670811.2025.2493759
Lamont, M., & Molnár, V. (2002). The study of boundaries in the social sciences. *Annual Review of Sociology*, *28*(1), 167-195. https://doi.org/https://doi.org/10.1146/annurev.soc.28.110601.141107

Lamont, M., & Swidler, A. (2014). Methodological pluralism and the possibilities and limits of interviewing. *Qualitative Sociology*, *37*(2), 153-171. https://doi.org/10.1007/s11133-014-9274-z

Lampel, J., & Meyer, A. D. (2008). Field-configuring events as structuring mechanisms: How conferences, ceremonies, and trade shows constitute new technologies, industries, and markets. *Journal of Management Studies*, *45*(6), 1025-1035. https://doi.org/https://ssrn.com/abstract=1862360

Lareau, A. (2021). *Listening to people: A practical guide to interviewing, participant observation, data analysis, and writing it all up*. University of Chicago Press.

Lee, F. (2025). The practices and politics of machine learning: A field guide for analyzing artificial intelligence. *AI & Society*, *40*(8), 6135-6148. https://doi.org/10.1007/s00146-025-02430-7

Lee, F., & Ribes, D. (2025). Computational universalism, or, Attending to relationalities at scale. *Social Studies of Science*, *56*(3). https://doi.org/10.1177/03063127251345089

Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. *Big Data & Society*. https://doi.org/https://doi.org/10.1177/2053951714534395

Liu, D., Nanayakkara, P., Sakha, S. A., Abuhamad, G., Blodgett, S. L., Diakopoulos, N., Hullman, J. R., & Eliassi-Rad, T. (2022). Examining responsibility and deliberation in AI impact statements and ethics reviews. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 424–435. https://doi.org/https://doi.org/10.1145/3514094.3534155

Lottick, K., Susai, S., Friedler, S. A., & Wilson, J. P. (2019). Energy usage reports: Environmental awareness as part of algorithmic accountability. *Workshop on Tackling Climate Change with Machine Learning at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. https://doi.org/https://doi.org/10.48550/arXiv.1911.08354

Loukissas, Y. A. (2019). *All data are local: Thinking critically in a data-driven society*. The MIT Press. https://doi.org/10.7551/mitpress/11543.001.0001

Luitse, D., Blanke, T., & Poell, T. (2025). AI competitions as infrastructures of power in medical imaging. *Information, Communication & Society*, *28*(10), 1735-1756. https://doi.org/10.1080/1369118X.2024.2334393

MacKenzie, D. (1998). The certainty trough. In R. Williams, W. Faulkner, & J. Fleck (Eds.), *Exploring Expertise: Issues and Perspectives* (pp. 325-329). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-13693-3_15

MacKenzie, D. (2021). *Trading at the speed of light: How ultrafast algorithms are transforming financial markets*. Princeton University Press.

Magalhães, J. C., & Smit, R. (2026). Less hype, more drama: Open-ended technological inevitability in journalistic discourses about AI in the US, The Netherlands, and Brazil. *Digital Journalism*, *14*(2), 323-340. https://doi.org/10.1080/21670811.2025.2522281

Malazita, J. W., & and Resetar, K. (2019). Infrastructures of abstraction: How computer science education produces anti-political subjects. *Digital Creativity*, *30*(4), 300-312. https://doi.org/10.1080/14626268.2019.1682616
Markelius, A., Wright, C., Kuiper, J., Delille, N., & Kuo, Y.-T. (2024). The mechanisms of AI hype and its planetary and social costs. *AI and Ethics*, *4*(3), 727-742. https://doi.org/https://doi.org/10.1007/s43681-024-00461-2
McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12-12. https://doi.org/https://doi.org/10.1609/aimag.v27i4.1904
Metcalf, J., & Crawford, K. (2016). Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society*, *3*(1). https://doi.org/https://doi.org/10.1177/2053951716650211
Metcalf, J., Moss, E., & boyd, d. (2019). Owning ethics: Corporate logics, Silicon Valley, and the institutionalization of ethics. *Social Research: An International Quarterly*, *86*(2), 449-476. https://doi.org/https://doi.org/10.1353/sor.2019.0022
Miceli, M., & Posada, J. (2022). The data-production dispositif. *Proceedings of the ACM on human-computer interaction*, *6*(CSCW2), Article 460. https://doi.org/10.1145/3555561
Miceli, M., Posada, J., & Yang, T. (2022). Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on human-computer interaction*, *6*(GROUP), Article 34. https://doi.org/10.1145/3492853
Moats, D., & Seaver, N. (2019). "You Social Scientists Love Mind Games": Experimenting in the "divide" between data science and critical algorithm studies. *Big Data & Society*, *6*(1). https://doi.org/10.1177/2053951719833404
Moore, J. (2019). AI for Not Bad. *Frontiers in Big Data*, *2*. https://doi.org/10.3389/fdata.2019.00032
Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., Dugan, C., & Erickson, T. (2019). How data science workers work with data: Discovery, capture, curation, design, creation. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper 126. https://doi.org/10.1145/3290605.3300356
Munk, A. K., Jacomy, M., Ficozzi, M., & Jensen, T. E. (2024). Beyond artificial intelligence controversies: What are algorithms doing in the scientific literature? *Big Data & Society*, *11*(3). https://doi.org/10.1177/20539517241255107
Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*, *3*(3), 869-877. https://doi.org/https://doi.org/10.1007/s43681-022-00209-w
Narayan, A., & Kapoor, S. (2025). AI as Normal Technology. (14.04.2025). Retrieved 21.06.2025, from https://perma.cc/HVN8-QGQY
Natale, S. (2021). *Deceitful media: Artificial intelligence and social life after the Turing test*. Oxford University Press. https://doi.org/https://doi.org/10.1093/oso/9780190080365.001.0001

Niemiec, E. (2022). Will the EU Medical Device Regulation help to improve the safety and performance of medical AI devices? *DIGITAL HEALTH*, *8*. https://doi.org/10.1177/20552076221089079
Norton, P. (2021). *Autonorama: The illusory promise of high-tech driving*. Island Press. https://doi.org/https://doi.org/10.5822/9781642832419
Offenhuber, D. (2024). Shapes and frictions of synthetic data. *Big Data & Society*, *11*(2). https://doi.org/10.1177/20539517241249390
Olazaran, M. (1996). A sociological study of the official history of the Perceptrons controversy. *Social Studies of Science*, *26*(3), 611-659. https://doi.org/10.1177/030631296026003005
Orr, W., & Crawford, K. (2024). The social construction of datasets: On the practices, processes, and challenges of dataset creation for machine learning. *New Media & Society*, *26*(9), 4955-4972. https://doi.org/10.1177/14614448241251797
Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society*, *23*(5), 719-735. https://doi.org/10.1080/1369118X.2020.1713842
Pasquinelli, M., & Joler, V. (2021). The Nooscope manifested: AI as instrument of knowledge extractivism. *AI & Society*, *36*(4), 1263-1280. https://doi.org/10.1007/s00146-020-01097-6
Pelizza, A. (2016). Disciplining change, displacing frictions. Two structural dimensions of digital circulation across land registry database integration. *Tecnoscienza – Italian Journal of Science & Technology Studies*, *7*(2), 35-60. https://doi.org/10.6092/issn.2038-3460/17326
Pentzold, C., & Knorr, C. (2024). When data became big: revisiting the rise of an obsolete keyword. *Information, Communication & Society*, *27*(3), 600-617. https://doi.org/10.1080/1369118X.2023.2227673
Phan, T., Goldenfein, J., Mann, M., & Kuch, D. (2022). Economies of virtue: The circulation of 'ethics' in Big Tech. *Science as Culture*, *31*(1), 121-135. https://doi.org/https://doi.org/10.1080/09505431.2021.1990875
Pickering, A. (1992). *Science as practice and culture*. University of Chicago Press. https://doi.org/https://doi.org/10.7208/9780226668208
Pickersgill, M. (2012). The co-production of science, ethics, and emotion. *Science, Technology, & Human Values*, *37*(6), 579-603. https://doi.org/10.1177/0162243911433057
Pinch, T., Collins, H. M., & Carbone, L. (1996). Inside Knowledge: Second order measures of skill. *The Sociological Review*, *44*(2), 163-186. https://doi.org/10.1111/j.1467-954X.1996.tb00420.x
Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, *14*(3), 399-441. https://doi.org/https://doi.org/10.1177/030631284014003004
Pink, S. (2022). Trust, ethics and automation: Anticipatory imaginaries in everyday life. In *Everyday automation* (pp. 44-58). Routledge.

https://www.taylorfrancis.com/chapters/oa-edit/10.4324/9781003170884-4/trust-ethics-automation-sarah-pink
Pollock, N., & Williams, R. (2011). Who decides the shape of product markets? The knowledge institutions that name and categorise new technologies. *Information and Organization*, *21*(4), 194-217. https://doi.org/https://doi.org/10.1016/j.infoandorg.2011.08.001
Pols, J. (2015). Towards an empirical ethics in care: Relations with technologies in health care. *Medicine, Health Care and Philosophy*, *18*(1), 81-90. https://doi.org/https://doi.org/10.1007/s11019-014-9582-9
Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.
Raley, R., & Rhee, J. (2023). Critical AI: A field in formation. *American Literature*, *95*(2), 185-204. https://doi.org/10.1215/00029831-10575021
Raz, A., Pontarotti, G., & Weitzman, J. B. (2019). Epigenetic metaphors: An interdisciplinary translation of encoding and decoding. *New Genetics and Society*, *38*(3), 264-288. https://doi.org/10.1080/14636778.2019.1601009
Ribes, D. (2019). How I learned what a domain was. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), Article 38. https://doi.org/10.1145/3359140
Ribes, D., Hoffman, A. S., Slota, S. C., & Bowker, G. C. (2019). The logic of domains. *Social Studies of Science*, *49*(3), 281-309. https://doi.org/10.1177/0306312719849709
Rotem, N., & Shani, L. (2025). Cutting across disciplines: Co-production and the rise of a computational culture. *Minerva*. https://doi.org/10.1007/s11024-025-09620-z
Ruckenstein, M. (2023). *The feel of algorithms*. University of California Press. https://doi.org/https://doi.org/10.1525/9780520394568
Ruckenstein, M., & Lehtiniemi, T. (2026). Friction and promise in data labor. *Science, Technology, & Human Values*, *51*(2), 277-303. https://doi.org/10.1177/01622439251358900
Ryan, G. W., & Bernard, H. R. (2003). Techniques to Identify Themes. *Field Methods*, *15*(1), 85-109. https://doi.org/10.1177/1525822x02239569
Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/https://doi.org/10.1145/3411764.3445518
Sambasivan, N., & Veeraraghavan, R. (2022). The deskilling of domain expertise in AI development. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Article 587. https://doi.org/10.1145/3491102.3517578
Samuel, G., Chubb, J., & Derrick, G. (2021). Boundaries between research ethics and ethical research use in artificial intelligence health research. *Journal of Empirical Research on Human Research Ethics*, *16*(3), 325-337. https://doi.org/10.1177/15562646211002744

Schaffer, F. C. (2006). Ordinary language interviewing. In D. Yanow & P. Schwartz-Shea (Eds.), *Interpretation and method: Empirical research methods and the interpretive turn* (pp. 150-160). M.E. Sharpe.
Schjøtt, A. (2026). Slicing the past to predict the future: Recasting data slicing as curatorial work in ML development and evaluation. *Cambridge Forum on AI: Culture and Society*, *2*, Article e7. https://doi.org/10.1017/cfc.2026.10014
Schjøtt, A., & Blanke, T. (2026). Making machine learning good enough – studying the political endeavour of finding 'right' metrics and thresholds. *Big Data & Society*, *13*(2). https://doi.org/10.1177/20539517261429198
Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2). https://doi.org/10.1177/2053951717738104
Seaver, N. (2022). *Computing taste*. University of Chicago Press. https://doi.org/doi:10.7208/chicago/9780226822969
Sheikh, H., Prins, C., & Schrijvers, E. (2023). AI is leaving the lab and entering society. In H. Sheikh, C. Prins, & E. Schrijvers (Eds.), *Mission AI: The New System Technology* (pp. 43-84). Springer International Publishing. https://doi.org/10.1007/978-3-031-21448-6_3
Shestakofsky, B. (2024). Cleaning up data work: Negotiating meaning, morality, and inequality in a tech startup. *Big Data & Society*, *11*(3). https://doi.org/10.1177/20539517241285372
Shilton, K. (2013). Values levers: building ethics into design. *Science, Technology, & Human Values*, *38*(3), 374-397. https://doi.org/10.1177/0162243912436985
Shklovski, I., & Némethy, C. (2023). Nodes of certainty and spaces for doubt in AI ethics for engineers. *Information, Communication & Society*, *26*(1), 37-53. https://doi.org/10.1080/1369118X.2021.2014547
Sigl, L., Felt, U., & Fochler, M. (2020). "I am primarily paid for publishing…": The narrative framing of societal responsibilities in academic life science research. *Science and Engineering Ethics*, *26*(3), 1569-1593. https://doi.org/10.1007/s11948-020-00191-8
Slota, S. C., Fleischmann, K. R., Greenberg, S., Verma, N., Cummings, B., Li, L., & Shenefiel, C. (2020). Good systems, bad data?: Interpretations of AI hype and failures. *Proceedings of the Association for Information Science and Technology*, *57*(1), e275. https://doi.org/https://doi.org/10.1002/pra2.275
Slota, S. C., Fleischmann, K. R., Greenberg, S., Verma, N., Cummings, B., Li, L., & Shenefiel, C. (2023). Locating the work of artificial intelligence ethics. *Journal of the Association for Information Science and Technology*, *74*(3), 311-322. https://doi.org/https://doi.org/10.1002/asi.24638
Slota, S. C., Hoffman, A. S., Ribes, D., & Bowker, G. C. (2020). Prospecting (in) the data sciences. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720906849
Small, M. L., & Calarco, J. M. (2022). *Qualitative literacy: A guide to evaluating ethnographic and interview research* (1 ed.). University of California Press. https://doi.org/10.2307/j.ctv2vr9c4x

Smith, B. C. (2019). *The promise of artificial intelligence: Reckoning and judgment*. The MIT Press.
Smith, R. D., Kawamura, K., Szymanski, E., Shineha, R., & Calvert, J. (15.05. 2024). The soba restaurant and the oyster bar: peripheral spaces for responsible research and innovation. https://doi.org/10.31235/osf.io/sprhu
Song, M. (2020). Rethinking minority status and 'visibility'. *Comparative Migration Studies*, *8*, Article 5. https://doi.org/https://doi.org/10.1186/s40878-019-0162-2
Soss, J. (2006). Talking our way to meaningful explanations: A practice-centered view of interviewing for interpretive research. In D. Yanow & P. Schwartz-Shea (Eds.), *Interpretation and method: Empirical research methods and the interpretive turn* (pp. 127-149). M.E. Sharpe.
Soss, J. (2021). On casing a study versus studying a case. In E. S. Simmons & N. R. Smith (Eds.), *Rethinking Comparison: Innovative Methods for Qualitative Political Inquiry* (pp. 84-106). Cambridge University Press. https://doi.org/https://doi.org/10.1017/9781108966009.005
Star, S. L. (1989). The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. In L. Gasser & M. N. Huhns (Eds.), *Distributed artificial intelligence* (pp. 37-54). Elsevier. https://doi.org/https://doi.org/10.1016/B978-1-55860-092-8.50006-X
Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, *19*(3), 387-420. https://doi.org/10.1177/030631289019003001
Star, S. L., & Ruhleder, K. (1994). Steps towards an ecology of infrastructure: Complex problems in design and access for large-scale collaborative systems. *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, 253–264. https://doi.org/10.1145/192844.193021
Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1), 111-134. http://www.jstor.org/stable/23010792
Stark, L. (2023). Breaking up (with) AI ethics. *American Literature*, *95*(2), 365-379. https://doi.org/https://doi.org/10.1215/00029831-10575148
Steinhoff, J. (2024). AI ethics as subordinated innovation network. *AI & Society*, *39*(4), 1995-2007. https://doi.org/10.1007/s00146-023-01658-5
Stevens, M., & Beaulieu, A. (2024). A careful approach to artificial intelligence: The struggles with epistemic responsibility of healthcare professionals. *Information, Communication & Society*, *27*(4), 719-734.
Stevens, M., Wehrens, R., Kostenzer, J., Weggelaar-Jansen, A. M., & de Bont, A. (2022). Why personal dreams matter: How professionals affectively engage with the promises surrounding data-driven healthcare in Europe. *Big Data & Society*, *9*(1), 20539517211070698. https://doi.org/10.1177/20539517211070698

Su, N. M., Lazar, A., & Irani, L. (2021). Critical affects: Tech work emotions amidst the techlash. *Proceedings of the ACM on human-computer interaction*, *5*(CSCW1), 1-27. https://doi.org/10.1145/3449253
Suchman, L. (1995). Making work visible. *Communications of the ACM*, *38*(9), 56-64. https://doi.org/https://dl.acm.org/doi/pdf/10.1145/223248.223263
Suchman, L. (2002). Located accountabilities in technology production. *Scandinavian Journal of Information Systems*, *14*(2), 7. https://doi.org/http://aisel.aisnet.org/sjis/vol14/iss2/7
Suchman, L. (2012). Configuration. In *Inventive methods* (pp. 48-60). Routledge.
Suchman, L. (2023). The uncontroversial 'thingness' of AI. *Big Data & Society*, *10*(2). https://doi.org/10.1177/20539517231206794
Suchman, L. A., & Trigg, R. H. (1993). Artificial intelligence as craftwork. In S. Chaiklin & J. Lave (Eds.), *Understanding Practice: Perspectives on Activity and Context*. Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9780511625510.007
Tavory, I. (2020). Interviews and inference: making sense of interview data in qualitative research. *Qualitative Sociology*, *43*(4), 449-465. https://doi.org/10.1007/s11133-020-09464-x
Tavory, I., & Timmermans, S. (2014). *Abductive analysis: Theorizing qualitative research*. University of Chicago Press.
Thakkar, D., Ismail, A., Kumar, P., Hanna, A., Sambasivan, N., & Kumar, N. (2022). When is machine learning data good?: Valuing in public health datafication. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Article 322. https://doi.org/10.1145/3491102.3501868
Thomas, S. L. (2019). Migration versus management: the global distribution of computer vision engineering work. *2019 ACM/IEEE 14th International Conference on Global Software Engineering (ICGSE)*, 12-17. https://doi.org/10.1109/ICGSE.2019.00017
Thompson, J. (2022). A guide to abductive thematic analysis. *The qualitative report*, *27*(5), 1410-1421. https://doi.org/https://doi.org/10.46743/2160-3715/2022.5340
Thylstrup, N. B. (2022). The ethics and politics of data sets in the age of machine learning: deleting traces and encountering remains. *Media, Culture & Society*, *44*(4), 655-671. https://doi.org/10.1177/01634437211060226
Thylstrup, N. B., Hansen, K. B., Flyverbom, M., & Amoore, L. (2022). Politics of data reuse in machine learning systems: Theorizing reuse entanglements. *Big Data & Society*, *9*(2). https://doi.org/10.1177/20539517221139785
Thylstrup, N. B., & Hartley, J. M. (2025). "Argh! The world doesn't fit the model!": Small acts of worldmaking in data annotation for news media. *Media Theory*, *9*(2), 105-132. https://doi.org/10.70064/mt.v9i2.1273
Tsing, A. L. (2005). *Friction: An ethnography of global connection*. Princeton University Press. https://doi.org/ https://doi.org/10.1515/9781400830596
Valdivia, A. (2025). The supply chain capitalism of AI: A call to (re) think algorithmic harms and resistance through environmental lens. *Information,*

Communication & Society, *28*(12), 2118-2134. https://doi.org/https://doi.org/10.1080/1369118X.2024.2420021
van der Vlist, F., Helmond, A., & Ferrari, F. (2024). Big AI: Cloud infrastructure dependence and the industrialisation of artificial intelligence. *Big Data & Society*, *11*(1). https://doi.org/10.1177/20539517241232630
van Dijk, N., Casiraghi, S., & Gutwirth, S. (2021). The 'ethification' of ICT governance. Artificial intelligence and data protection in the European Union. *Computer Law & Security Review*, *43*, 105597. https://doi.org/https://doi.org/10.1016/j.clsr.2021.105597
van Maanen, G. (2022). AI ethics, ethics washing, and the need to politicize data ethics. *Digital Society*, *1*(2), 9. https://doi.org/10.1007/s44206-022-00013-3
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, Article 7181. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
Verbeek, P.-P. (2006). Materializing morality: Design ethics and technological mediation. *Science, Technology, & Human Values*, *31*(3), 361-380. https://doi.org/https://doi.org/10.1177/0162243905285847
Vipra, J., & West, S. M. (2023). *Computational power and AI*. A. N. Institute. https://ainowinstitute.org/wp-content/uploads/2023/09/AI-Now_Computational-Power-an-AI.pdf
Wagner, B. E. N. (2018). Ethics as an escape from regulation: From ethics-washing to ethics-shopping. In E. BayamlioĞLu, I. Baraliuc, L. Janssens, & M. Hildebrandt (Eds.), *BEING PROFILED* (pp. 84-89). Amsterdam University Press. https://doi.org/10.2307/j.ctvhrd092.18
Wainwright, S. P., Williams, C., Michael, M., Farsides, B., & Cribb, A. (2006). Ethical boundary-work in the embryonic stem cell laboratory. *Sociology of Health & Illness*, *28*(6), 732-748. https://doi.org/https://doi.org/10.1111/j.1467-9566.2006.00539.x
Wehrens, R., Stevens, M., Kostenzer, J., Weggelaar, A. M., & de Bont, A. (2023). Ethics as discursive work: The role of ethical framing in the promissory future of data-driven healthcare technologies. *Science, Technology, & Human Values*, *48*(3), 606-634. https://doi.org/10.1177/01622439211053661
Weiss, R. S. (1994). Respondents: Choosing them and recruiting them. In *Learning from strangers: The art and method of qualititave interview studies* (pp. 15-37). Free Press.
Widder, D. G. (2024). Epistemic power in AI ethics labor: Legitimizing located complaints. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1295-1304. https://doi.org/https://doi.org/10.1145/3630106.3658973
Widder, D. G., & Nafus, D. (2023). Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231177620

Widder, D. G., Nafus, D., Dabbish, L., & Herbsleb, J. (2022). Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2035–2046. https://doi.org/10.1145/3531146.3533779

Widder, D. G., Whittaker, M., & West, S. M. (2024). Why 'open' AI systems are actually closed, and why this matters. *Nature*, *635*(8040), 827-833. https://doi.org/https://doi.org/10.1038/s41586-024-08141-1

Widder, D. G., Zhen, D., Dabbish, L., & Herbsleb, J. (2023). *It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them?* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago, IL, USA. https://doi.org/10.1145/3593013.3594012

Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*(1), 121-136. http://www.jstor.org/stable/20024652

Wirth, A., Ludec, C. L., Girard-Chanudet, C., & Cornet, M. (2025). Who decides when artificial intelligence works? Algorithmic production as a continuous evaluation process. *Digital Society*, *4*, Article 65. https://doi.org/10.1007/s44206-025-00224-4

Wolf, E. R. (1982). *Europe and the People without History*. University of California Press.

Yanow, D. (2006). Thinking interpretively: Philosophical suppositions in the human sciences. In D. Yanow & P. Schwartz-Shea (Eds.), *Interpretation and method: Empirical research methods and the interpretive turn* (pp. 5-26). M.E. Sharpe.

Yanow, D., & Schwartz-Shea, P. (Eds.). (2006). *Interpretation and method: Empirical research methods and the interpretive turn*. M.E. Sharpe.

Young, E., Wajcman, J., & Sprejer, L. (2023). Mind the gender gap: Inequalities in the emergent professions of artificial intelligence (AI) and data science. *New Technology, Work and Employment*, *38*(3), 391-414. https://doi.org/https://doi.org/10.1111/ntwe.12278

Zając, H. D., Avlona, N. R., Kensing, F., Andersen, T. O., & Shklovski, I. (2023). Ground truth or dare: Factors affecting the creation of medical datasets for training AI. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 351–362. https://doi.org/10.1145/3600211.3604766

Zakharova, I., & Jarke, J. (2024). Care-ful data studies: Or, what do we see, when we look at datafied societies through the lens of care? *Information, Communication & Society*, *27*(4), 651-664. https://doi.org/10.1080/1369118X.2024.2316758

Zhang, A. X., Muller, M., & Wang, D. (2020). How do data science workers collaborate? Roles, workflows, and tools. *Proceedings of the ACM on human-computer interaction*, *4*(CSCW1), Article 22. https://doi.org/10.1145/3392826

Zhang, D., Maslej, N., Brynjolfsson, E., Etchemendy, J., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Sellitto, M., & Sakhaee, E. (2022). *The AI Index 2022 annual report*. S. I. f. H.-C. A. AI Index Steering Committee, Stanford

University. https://hai.stanford.edu/assets/files/2022-ai-index-report_master.pdf
Ziewitz, M. (2016). Governing Algorithms:Myth, Mess, and Methods. *Science, Technology, & Human Values*, *41*(1), 3-16. https://doi.org/10.1177/0162243915608948
Ziewitz, M. (2019). Rethinking gaming: The ethical work of optimization in web search engines. *Social Studies of Science*, *49*(5), 707-731. https://doi.org/10.1177/0306312719865607
Zilberstein, S. (2024). Ethical Dilemmas and Collaborative Resolutions in Machine Learning Research for Health Care. *Socius*, *10*. https://doi.org/10.1177/23780231241259671

Appendix

Appendix I: Interview guide

Intro

- Thank you for taking the time to talk to me.
- Short description of the project.
- Information on recording & data handling, consent form
- Then I will turn the recorder on?

Background and self-identification/presentation

Main questions	**Probes & probing instructions**
1. Could we start with you telling me a bit about your research area and background?	- How would you position your work in terms of disciplines, fields? - And how did you end up working on/at / studying…? - What did you find interesting about…? - *If transitioned between sectors or in-between:* how do you find industry/NGO compared to academia?
2. How do you usually describe your research specialization? / area of research?	- [*audiences*] Within the field / broader academic audiences / dissemination to non-academic audiences / outside of work - [*reactions*] What do people say to that?
3. There are a lot of popular terms related to your broad area of research. Machine learning, big data, AI, data science. What do you think about them?	- Do you use them to describe your research and yourself as a researcher? - Why/why not?

II Research projects

4. Could you tell me a bit about your most recent projects, where you work with machine learning and large-scale human data?	- Main goal/purpose, basic/applied (where) - Role of the interviewee in the project - Data, methods - What do you think about the project? - Could you give me another example of your projects with machine learning and human data?

III Valuation of research and impact

5. In thinking about what kind of research you would want to do, what are the aspect that are important for you to consider?	- [remember not give answer options] - Could you give me an example of a research you look up to? A kind of research you aspire to do in the future? What is it about it that is aspirational/ admirable? - Is there a kind of research you would never do?
6. Ideally, what kind of impact would you want your research to have?	- How conducive do you think *field/sector* is to that? - And do you have any concerns about the impact of your research? - Do you think this vision is common in your field?
8. How do you think your field / research community is perceived from the outside?	- [audiences] by researchers in other fields or sectors/ greater society? - How do you feel about that?

IV Ethics

We talked a bit about different considerations about research and what is important. One aspect of data-intensive research and machine learning that has been discussed more recently is ethics.

a) warm-up – ethics domain

9. Could you tell me what comes to mind if you think about "ethics" in your area of research?	[*here follow those topics in Q10 or note them down to come back*]

b) within own research

10. Could you tell me about some of the ethical considerations you have had in relation to your research? [*probe in relation to the issues mentioned in Q9*]	[*try to get a few cases: details & attitude to them*] - Could you tell me when this became an issue to consider? - Could you tell me about the last time – if any – you felt you had an ethical question related to your research? - Could you tell me about some moments within your research projects, if any, when you made decisions related to ethics? - Could you tell me a little about how you solve these issues?
11. The way you see and approach these ethical issues…do you think it is similar to other researchers in your field?	- Why/ why no? What do you think about that? - And do you think it is different from researchers in other fields?
12. Could you tell me, how is it for you -to deal with these ethical issues?	- Main challenges / tools - What aspects of ethics within your research are easier / more difficult?

c) ethics talk in different moments of their (professional) life

Ethics is also something that seems to be talked about a lot nowadays in relation to AI, machine learning, big data etc… I would like now to talk about your experiences with conversations about ethics.

13. In your experience, what are the main things that are discussed in relation to ethics in your research area?	- Could you tell me something about the last time you saw, read, or heard something about ethics in your research area? - Could you give me a recent example in another context? - How do these discussions relate to the ethical considerations in your own research?
14. Do you yourself have conversations about ethics?	- Could you tell me about the last time you talked about something related to ethics? [probe for details & attitude]

d) knowledge about ethics

15. From where you do think you learned what you know about ethics?	- Have you received any ethics training in your education or during your work? What do you think about what you learned there? - Have you followed any discussions about ethics in field / ML / AI etc.? - Have you consulted any ethics policies or guidelines?
16. Between now and then, what has changed – if anything – in terms of what you think about ethics?	- From where you are now, what do you think about your past knowledge and opinions about ethics? - Has anything changed in terms of how you approach ethics practically, in your research?

V Governance

Nowadays, doing research, researchers also encounter different regulations. But from what I see, it is very different for different projects. /
You mentioned … before. Now I would like to talk a bit more about your experience with regulation.

17. In your research projects, do you deal with any regulations, including ethics regulations, in relation to your research?	[*probe for examples, details & attitude*] - How did you learn about this regulation? At what stage of the project were you dealing with it? - Did you have any experience with any other regulations within this project? [*probe a little about particular regulations*] - Were there any guidelines/policies you had to apply in your research? - Did you have to apply for any kind of (ethics) approvals? - How did you get access to the data? - Requirements of conferences / journals / funding agencies - What happens to the data and code after the project is finished? - Was it similar to other projects of yours? - Do you think this is typical for your field – from talking to colleagues, at conferences…? Why/why not?
18. From what I can see, the organizational approaches to dealing with regulation vary a lot from one place to another. How does it work for you?	- Whose job is it to know about these regulations and what you are supposed to do? Within/outside project team - Is there anyone supporting or assisting you with dealing with regulations? Also outside of your institution? - Do you think this is typical for your field – from talking to colleagues, at conferences…? Why/why not?
19. Could you tell me, how do you find dealing with regulations, including ethics regulations?	- What makes it easier / more difficult? - What are more easy / difficult aspects to it?
20. If you could change the ways this system of regulation of research, including ethics governance, works, what would you change?	- Ideally, how would regulation related to ethics work? - [probe in relation to mentioned issues]. You said … is difficult. What would make it better?

Outro

- Double-check if I need to clarify something / get details / factual information
- Check in about how I could refer to the person in a pseudonymous fashion
- I think I have asked all the questions I had. Before we finish, is there something you would like to add or elaborate on?
- Repeat agreements & ask for snowballing if needed

Appendix II: Ethical approval for the interview study

SCHOOL OF BUSINESS AND SOCIAL SCIENCES
AARHUS UNIVERSITY

To
PhD Student Oksana Dorofeeva
Department of Political Science
Aarhus BSS

The Research Ethics Committee, Aarhus BSS

Lone Sønnemann Jacobsen
Chief consultant

Date: 27 November 2023

Mobile Tel.: +45 3056 2189
E-mail: los@au.dk
Web: au.dk/en/los@au.dk

Journal no: 2023-0604103
Approval number:
BSS-2023-119-K

Sender's CVR no.: 31119103

Re the 'Interview study on researchers' experiences with ethics and governance in machine learning research with human data' research study
Person in charge of the study: PhD Student Oksana Dorofeeva
Contact/study manager: PhD Student Oksana Dorofeeva
Study period: December 2023 – March 2024

Aarhus University's Research Ethics Committee/Research Ethics Committee - BSS (Institutional Review Board) discussed the study at its meeting on 23 November 2023 and came to the following decision:

Decision

The study *is approved* in accordance with Aarhus University's guidelines for the university's Research Ethics Committee (IRB) and the considerations listed in the guidelines.

Approval is provided that applications are submitted for approvals in other involved countries, when relevant or required, in order to ensure that the study and the personal data collection also comply with the relevant/local ethical requirements and standards of these countries.

The approval is granted with reference to the following documents:

- Information sheet with appendices:
  - Application
  - Protocol

Dean's Office, Aarhus BSS
Aarhus University
Bartholins Allé 14
DK-8000 Aarhus C
Denmark

Tel.: +45 8715 2454
E-mail: bss@au.dk
Web: bss.au.dk/en

- Recruitment material
- Participant information
- Declaration of consent
- Study manager's CV

The person or persons identified as being in charge of the study guarantee that the study is carried out as described in the referenced documents and is otherwise in compliance with applicable research ethics and data protection regulations.

As a result of the approval, subsequent academic publications which are based on the findings of the study may be provided with the endorsement that 'the study was approved by the Institutional Review Board at Aarhus University' (indicating the approval number).

The following Committee members participated in the discussion of the study:

- Professor Jesper Wiborg Schneider (BSS) (Chairman)
- Professor Heidi Houlberg Salomonsen (BSS)
- Lektor Kristian Hvidtfelt Nielsen (NAT)

Kind regards,

Lone Sannemann Jacobsen
Chief consultant

Appendix III: Consent form for the interview study

DANISH CENTRE FOR STUDIES IN RESEARCH AND RESEARCH POLICY
DEPARTMENT OF POLITICAL SCIENCE
AARHUS UNIVERSITY

Participant information

Interview study on researchers' experiences with ethics and governance in machine learning research with human data

This document provides you with the information you need to give informed consent to participation in this study and to the processing of your personal data pursuant to the rules of the General Data Protection Regulation.

The study

The project is carried out by Oksana Dorofeeva, PhD Student at the Danish Centre for Studies in Research and Research Policy, and is funded by the Social Science and Business PhD program at Aarhus University. This study is part of her PhD dissertation that explores how researchers navigate issues related to ethics and governance within data-intensive research. The study aims to get a better understanding of the phenomena of ethics and governance related to data-intensive research by engaging with the perspectives of researchers who work with machine learning and human data. The study will generate knowledge that can contribute to more informed discussions on ethics and governance in data-intensive research, which should be beneficial to both the research community and society at large. The findings of the study will be communicated in a PhD dissertation, in other forms of scientific dissemination (e.g., articles), and in other forms of more popular dissemination (e.g., talks, media, and social media publications, etc.). In case the participants are interested, the research publications based of the results of this study will be forwarded by the author.

The project is a qualitative semi-structured interview study. Researchers with relevant experience in terms of machine learning and human data are recruited through online pages research institutions, relevant companies, and research events, as well as through existing contacts, and will be contacted with a written email invitation to participate in the study. Participants will be asked to kindly suggest other potential participants.

Upon request by the participant, the researcher would provide the transcribed interview and quotes. The researcher would also be available for a follow-up meeting if requested by the participant.

Processing and handling of data

If you agree to participate in the project, information, incl. personal information, provided during the interview will be processed based on 6(1)(a) and Article 9(2)(a) of the General Data Protection Regulation. The interview will be recorded and transcribed verbatim with assistance from students employed by the Danish Centre for Studies in Research and Research Policy at Aarhus University. The transcripts will

Danish Centre for Studies in Research and Research Policy

Oksana Dorofeeva
PhD fellow

Date: 23 October 2023

E-mail: oksana.dorofeeva@ps.au.dk
Web: au.dk/en/oksana.dorofeeva@ps

Danish Centre for Studies in Research and Research Policy
Aarhus University
Bartholins Allé 7
DK-8000 Aarhus C
Denmark

oksana.dorofeeva@ps.au.dk
au.dk/en/oksana.dorofeeva@ps

AACSB ACCREDITED · ASSOCIATION AMBA ACCREDITED · EQUIS ACCREDITED

be analysed by the researcher (Oksana Dorofeeva). The interview data will be treated with confidentiality and will only be shared or published in pseudonymized form. Your data will be stored and processed by Aarhus University for 5 years since the study is completed in accordance with the rules on responsible conduct of research, after which it will be destroyed.

Risk and inconveniences

Potentially critical ethical implications of the interview study with regard to human dignity and integrity, or the privacy of persons are not expected.

Consent

By signing this, I acknowledge that I have read and understood the information provided to me about the study and the processing of personal data. I have had the opportunity to ask questions and my questions have been sufficiently answered. I have had enough time to decide whether I would like to participate.

I hereby consent to participate in the study and consent to Aarhus University registering and processing my personal data in the research project 'Interview study on researchers experiences with ethics and governance in machine learning research with human data' in accordance with the information that was given to me.

Date and participant's signature:	Date and researcher's signature:
_____	_27.11.2023_____
Name in capital letters:	Name in capital letters:
	OKSANA DOROFEEVA

Giving your consent is voluntary and you may at any time stop your participation and withdraw your consent to the processing of your personal data, without having to provide any justification. You do this by contacting Oksana Dorofeeva at oksana.dorofeeva@ps.au.dk. If you withdraw your consent, your personal data given before the withdrawal will continue to be included in the work carried out in the project before you withdrew your consent.

Danish Centre for Studies in Research and Research Policy
Aarhus University
Bartholins Allé 7
DK-8000 Aarhus C
Denmark

oksana.dorofeeva@ps.au.dk
au.dk/en/oksana.dorofeeva@ps

Appendix IV: Secondary data sources list

Code	Conference	Source Title	URL	Access date
ACL	*	Responsible NLP checklist	https://aclrollingreview.org/responsibleNLPresearch/	29.04.2025
CVRP-24a	CVPR	CVPR 2024 Ethics guidelines for authors	https://cvpr.thecvf.com/Conferences/2024/EthicsGuidelines	29.04.2025
CVPR-24b	CVPR	CVPR 2024 Submission Policies	https://cvpr.thecvf.com/Conferences/2024/AuthorGuidelines	29.04.2025
CVPR-24c	CVPR	CVPR 2024 Suggested Practices for Authors	https://cvpr.thecvf.com/Conferences/2024/AuthorSuggestedPractices	29.04.2025
CVPR-25a	CVPR	CVPR 2025 Ethics guidelines for authors	https://cvpr.thecvf.com/Conferences/2025/EthicsGuidelines	29.04.2025
CVPR-25b	CVPR	CVPR 2025 Submission Policies	https://cvpr.thecvf.com/Conferences/2025/AuthorGuidelines	29.04.2025
CVPR-25c	CVPR	CVPR 2025 Suggested Practices for Authors	https://cvpr.thecvf.com/Conferences/2025/AuthorSuggestedPractices	29.04.2025
ICLR-24	ICLR	ICLR 2024 Author Guide	https://iclr.cc/Conferences/2024/AuthorGuide	29.04.2025
ICLR-25a	ICLR	ICLR 2025 Author Guide	https://iclr.cc/Conferences/2025/AuthorGuide	29.04.2025
ICLR-25b	ICLR	ICLR Code of Ethics	https://iclr.cc/public/CodeOfEthics	29.04.2025
ICML-24a	ICML	From ICML 2024 Program chairs: Impact Statements in ICML Submissions	https://medium.com/@icml2024pc/impact-statements-in-icml-submissions-0ba1ea22119b	29.04.2025
ICML-24b	ICML	ICML 2024 Call for papers	https://icml.cc/Conferences/2024/CallForPapers	29.04.2025
ICML-24c	ICML	ICML 2024 Publication ethics	https://icml.cc/Conferences/2024/PublicationEthics	29.04.2025
ICML-25a	ICML	ICML 2025 Call For Papers	https://icml.cc/Conferences/2025/CallForPapers	29.04.2025
ICML-25b	ICML	ICML 2025 Publication ethics	https://icml.cc/Conferences/2025/PublicationEthics	29.04.2025
NeurIPS-25a	NeurIPS	NeurIPS Code of Ethics	https://neurips.cc/public/EthicsGuidelines	29.04.2025
NeurIPS-25b	NeurIPS	NeurIPS Paper Checklist Guidelines	https://neurips.cc/public/guides/PaperChecklist	29.04.2025

Appendix V: Coding list for Paper 1

Name	Description	Cases	References
AI as a term	Codes related to 'artificial intelligence'/'AI' as a term.		
affordances of 'AI'	Mentioning AI in terms of what it does or affords in communication, invoking (supposed) effects. + and – indicate portraying these effects as positive or negative.		
'AI' brings in concerns (–)	Mentioning that the term invokes concerns and worries. Includes mentions that doing so might be undesirable.	4	6
'AI' creates (too) high expectations (–)	Describing that the term 'AI' can invoke unrealistic expectations, the field being ahead of where it actually is; portraying those expectations as a potential problem.	3	3
'AI' is what people understand and engage with (+)	Invoking the recognition of the word 'AI' and having connotations with it (an idea of what it is or could be) by non-specialists and general publics. Also making a narrative link to establishing communication, engagement – showing this understanding/recognition is good for communication.	11	23
AI is known more than ML	indicating that non-specialists are more familiar with 'AI' than 'machine learning', which would make the former more preferable term for communication.	8	16
'AI' looks good promotes (+)	Indicating that invoking AI could be helpful and promote their work. Compared to the 'recognition' one, it more reflects prestige or something like that.	3	4
AI is for communicating with outsiders, non-specialists	Indicating that AI is a term for talking to people outside the field / of different backgrounds.	14	27
AI vs. ML	Discussing the relationship between AI and machine learning categories, both in a 'realist' way (methods, approaches) and about terms and the joke. Some of this could also be in the definitions (which is why there is no ML vs AI code there), but that one is more realist, and I did not want to make a distinction between the 'communicative' and the 'realist' but look at all meanings about terms altogether.		
ML ⊂ AI	Saying ML is part of AI (but not all of it)	6	7
ML, AI (=)	Portraying ML and AI as interchangeable	3	3
THE JOKE	Telling me the AI, ML, regression joke	3	3

Name	Description	Cases	References
R Python PowerPoint joke	A joke with same regression ML AI elements but about coding languages vs. PowerPoint	1	1
'AI's changing meaning over time	Invoking changes in AI's meaning over time, what was AI at some earlier point is not now and vice versa	3	6
old school symbolic AI		1	1
associating 'AI' with promotional behaviour	Making a connection between someone (including themselves) using the 'AI' term and the goal to promote one's work business etc. Includes both portrayals where this is problematised and where it is not.	8	9
critiques of AI as a term	More critical opinions on 'AI' as a term. Not necessarily its 'effects' in communication but also the concept more generally.	0	0
'AI' is a buzzword	Calling 'AI' a buzzword	5	8
'AI' is vague too broad	Characterising the 'AI' term as too broad or vague to be properly meaningful	4	6
it's new naming, relabelling stuff that was already there		1	1
how non-specialists interpret 'AI'	Discussing how non-specialists (including e.g. clients) understand 'AI' (beyond know/recognise what it is). Includes reactions to self-descriptions (in that code) but also more general things.	1	5
ChatGPT as the first or strong association interpretation	Noting that members of the general public understand AI as ChatGPT as main example (or at least indicating a strong association)	10	11
indicating change in the last couple years	noting that how non-specialist understand that concept has recently shifted (also usually connected to ChatGPT/LLMs)	4	5
on (not) using 'AI' when talking about their work or themselves	Discussing whether they would use 'AI' and/or 'ML' when describing themselves and their work.	0	0
avoiding 'AI'	Indicating that they (try to) avoid using the 'AI' term and discussing why.	6	10
machine learning over 'AI'	Narratives about preferring to present themselves with ML over AI	3	5
on using or not 'AI' for funding	Discussing whether they use 'AI' in funding-related situations and why/why not	5	6

Name	Description	Cases	References
using 'AI'	Narratives about using 'AI' in describing their work.	13	29
identifying with ML but using AI	Explaining that genuinely they identify with ML but use AI for communicative purposes. These are excerpts where the interviewees more explicitly make the distance between themselves and AI compared to other notions where AI is invoked for communication.	3	10
ChatGPT as AI	References to ChatGPT as the example/application of ML/AI. Using it in comparison to other ML topics/methods and making other connections. Also includes references to the ChatGPT release as a moment that separates 'before' and 'after'. Does not include 'how people interpret AI', this is a code of its own in the 'term' theme.	17	31
ChatGPT brought the field attention	Attributing increase in the interest in/attention to ML research to ChatGPT.	4	6
connecting people's understanding of and interest in their work to ChatGPT	Indicating people around them and those they describe their work to are more familiar and/or are interested since ChatGPT came out or because of ChatGPT.	9	18
data science	Codes on the notion of data science and data scientist as a job title.		
data science as term	Discussing the term 'data science'.	3	6
data science vs. ML	Characterising the relationship between data science and machine learning. Includes both distinctions and non-distinctions.	3	5
data science vs. statistics	Characterising the relationship between data science and statistics. Includes both distinctions and non-distinctions.	2	2
identifying as a data scientist	Describing themselves as data scientists.	3	6
the job of the data science is and includes	Describing and characterising data science as a title/job, incl. in contrast to other positions. Includes discussions on how 'data scientist' can mean very different things.	8	22
definitions ML AI	More 'realist' (compared to the buzzword/communication device) discussions of what is artificial intelligence, machine learning, etc.		
machine learning vs. deep learning	Portraying the relationship between machine learning and deep learning. Also includes not making a strong distinction or saying it is difficult to do so.	11	18

Name	Description	Cases	References
what is AI and what is not	More 'realist' inclusion and exclusion regarding AI. Also includes notions that it is difficult to draw these boundaries.	8	13
criterion of definition of AI	Invoking a criterion or definition of AI.	5	6
inclusion and exclusion from AI work and expertise	Including and excluding certain types of work or projects from the notion of work on AI. Includes a distinction of the work "with AI" (e.g., using ChatGPT but not doing training or something).	8	15
how they present or introduce themselves and their research or work	Describing how they would describe and/or label their work when commuting with non-specialists. Includes both professional, e.g. science communication contexts and social/outside of work contexts (this is also how I sometimes asked them the questions, asking them about being at a party).		
depends on the audience or background	Indicating that they take their interlocutors' background or other characteristics into account in how they present their work. Sometimes this is a response to my question about how they would portray or describe themselves, and they clarify that it depends.	6	7
describing how they would introduce themselves among specialists	Explaining how they describe or introduce their work among other researchers or technologists.	11	13
depends on which conference	Indicating that venue/audience of different conferences would matter for how they described their work.	2	4
it is good to highlight applications	Noting that in presenting one's work (in presentations or papers etc.) it is helpful to note the applications of models.	4	7
how they describe themselves and their work to general public	Explaining how they would characterise and label their work when talking to members of the general public/ people outside of their professional sphere. The codes describing using 'AI' have moved to the 'AI as a term' theme.	28	45
how lay people react to what they do	Descriptions of reactions to the interviewees' presentations descriptions of their work.	18	37
people do not know what my work is	Indicating that people around the interviewees know or understand little about their work.	2	3

Name	Description	Cases	References
portraying technical work not being interesting for others	Describing that they find or expect interlocutors without a technical or similar background to be interested in what they do.	7	10
simplified relatable	Describing that they try to go for descriptions that would include something familiar and understandable (terms and/or examples), and relatable to people's experiences.	11	15
joking	Describing that they use a joke or a funny description to interest people.	2	2
lay people's attitudes to AI	This code is for narratives where interviewees portray how, in their opinion, non-specialists perceive ML/AI and also describe their encounters with these opinions.		
advising informing non-specialists on AI	Interviewees describing how they try to inform on ML/AI or answer critiques.	4	5
considering expectations from AI ML	Describing that non-specialists have unrealistic expectations of ML/AI. Includes both in specific interactions and talking 'in general'.	4	8
encountering concerns from people	Descriptions of encounters (both in professional and non-professional, e.g., being at a party) with non-specialists where the latter ask worried questions or voice critiques and concerns to the interviewees.	8	12
impressed interested positive	Describing that they think people are interested positively in ML AI.	5	8
noting the gulf between insider and outsider perceptions	Describing that describing a significant difference between how people working on ML/AI perceive things and how people on the outside perceive and discuss ML/AI. Interestingly, some of these narratives also portray that the specialists' perspective is not too of interest to these non-specialist interlocutors.	11	17
opinions or caring about terminology	Codes indicating interviewees' opinions on matters of terminology around their work.		
hard to distinguish between the terms	Noting that it is difficult for them to (definitively) differentiate the terms.	3	3
not caring about terminology	Expressing that they do not ascribe a lot of significance to terminology.	5	5
not judging other people for ai vs. ml choice	Saying they would not judge others for which labels they would choose, 'ML' or 'AI'.	2	2

Name	Description	Cases	References
relating their work to the concept of AI	Positioning themselves and their work in relation to AI. Includes both more declarative and naturalised, in passing connections. Does not include the communication scenarios. This code includes the excerpts that did not fit into any of the categories.	**3**	**3**
ascribing AI to themselves	in the interaction with me, in passing / naturalised rather than in response to questions about this or discussions about terminology, not about how they present themselves to other people	6	13
explicitly 'not AI'	Explicitly distancing themselves and saying what they do is not AI.	1	5
I'm on the border re AI	Describing their position as on the border, not fully in or out.	1	1
technically is AI	Indicating that 'AI' 'technically fits' but this is not how they think about it themselves. Accurate but not definitive.	5	9
relations within the broad ML/AI landscape	Codes for narratives on the different 'parts' of the ML/AI landscape (fields, approaches) and different (and changing?) relationship between them: boundaries, hierarchies, exchanges, etc.		
computer vision as central in ML research	Indicating a strong connection between ML research and computer vision field and data, and noting the latter's dominance in ML compared to other types or modalities of data.	4	4
fast pace in general ML	Mentioning the rapid development in general ML/AI as a matter of significance for something.	6	12
general, methods ML research and field	Interviewees invoking the descriptions of the mainstream/methods/general ML/AI research that is focused on improving methods, architectures on benchmark datasets. Includes both those belonging to it and describing it from the 'inside' and those describing it from other positions (see contrast).	9	19
contrasting to general methods ML field	Interviewees making a distinction contrast between their work/field and the 'general', 'mainstream', 'methods' ML research focused on improving methods on benchmarks.	8	21
NLP vs. general ML research	Discussing the relationship between the relationship between the field of natural language processing and machine learning.	6	20
bridging NLP and ML	Indicating that NLP and ML research got closer together in the last 10 years. Describing history as well as personal experience.	4	8

Name	Description	Cases	References
noting hierarchies of worth popularity	Interviewees describing how some topics or approaches (not necessarily theirs) are seen as more prestigious, mainstream etc., while others are looked down on or not taken as seriously.	3	5
generative methods as THE popular MLAI subfield at the moment	Indicating that generative AI and LLMs are where attention and interest concentrate (among other ML/AI approaches and topics). Some of them call it 'hype' or 'hyped' but not all so I am choosing this slightly more neutral naming. Includes both 'internal' (among researchers) and 'around' them: clients, funders, etc.	9	13
noting trends and changes in ML research	Indicating that trends in ML research come and go, and the popular topics change	4	5
statistics	Codes related to statistics.		
identifying with statistics	Say they identify as statisticians.	2	6
statistics vs. ML/AI	Describing the relations between ML/AI and statistics.	3	12

Appendix VI: Coding list for Paper 2

A: Coding list for interview data

Name	Description	Cases	References
energy and efficiency	Discussing energy and compute use as ethical issues in ML (different from discussions about how not everyone has the same access to GPUs within ML).	**5**	**8**
considering their own energy use	Describing using energy and compute resources as the ethics aspects of their own work.	2	2
ethics a research or project topic	Codes capturing the understanding of ethics as work on ethics-related topics: e.g., FAccT topics.		
'ethical' model qualities have to be a specialization	Describing that making the model more 'ethical' via dealing with bias, enforcing fairness measures, etc. is something to be done in specialised projects and is not something to do every time. Includes expressing that this would be worthy to be a specialised paper or contribution.	**4**	**6**
if there's a chance of writing some flavour of fairness or differential privacy into the paper, I will try to	In vivo code. The interviewee also works on privacy.	1	1
encounters with ethics as a topic to read about or discuss	Codes capturing ethics as discussions or topics of discourse.		
ethics popping up in their feed	Captures encountering discussions of topics interviewees associate with ethics and ML as occasional, rather than more intentional consumption, 'following'. This is an in vivo code initially but also includes traditional media (news, radio, etc.)	9	15
reading & discussing papers	Mentioning reading papers related to ethics. Some are broad sentiments 'I see papers', couple more concrete examples. Interestingly, two of them mention that this is because they read about LLMs (because this is their topic) and papers on LLMS mention ethical issues.	5	5
specialised talks meetings events	Mentioning attending talks or events that included ethics discussions (NB 2 of the interviewees went together to the same one).	6	8
within conferences	Describing discussions related to ethics they encountered at conferences with broader focus (e.g., ML conferences). These seem to be more unexpected for the interviewees.	3	4

Name	Description	Cases	References
ethics-coded topics as ML research topics	Discussing that 'ethics-coded' topics like fairness or privacy are also 'normal' ML topics: also optimising, but for other metrics. Includes a mention that these topics can be interesting as research problems rather than normative commitments.	2	8
I know because I care	An interviewee indicating that they intentionally follow discussions related to ethics and they find it personally important to be informed on this.	2	2
referencing	Codes capturing instances when practitioners reference 'ethics' content and discussions.	0	0
I want AI to do laundry	Interviewees referencing the expression "I want AI to do my laundry and dishes so that I can do art and writing, not for AI to do my art and writing so that I can do laundry and dishes." I remember seeing it a lot online around the time of the interviews.	2	2
mentioning controversy cases	References involving cases of problematic models or datasets.	6	7
notable ethics people	Mentioning (who they consider) notable ML ethics experts.	2	2
towards ethics knowledge	Codes capturing interviewees' descriptions of their relationship to ethics as an area or a way of knowledge.		
ethics (more) natural due to field discipline background	Noting that their field or discipline makes thinking about ethics-coded considerations more natural. 2 of these are statistical background and bias, another is interpretability.	3	3
following only matching topics at conferences	Interviewees expressing that at conferences, they attend talks and posters related to their immediate specialisation or research interests, which excludes ethics and 'ethics'-coded topics like fairness.	4	5
not an ethics expert natural due to field background	Expressing that they are not specialists in what they consider ethical issues and approaches, connecting it to background training or their different research/technical specialisation: find it hard to understand, don't have the terminology, etc. This is not necessarily distancing since some of them work on these topics! But explain that there is a learning curve and barriers. Also comparison with philosophers and social scientists.	6	9

Name	Description	Cases	References
not my field	Expressing that ethics related topics are not their topics. Not sure if this is a distancing thing, but this contributes to the representation of ethics as a (set of) topic(s) one can specialise in (or not).	4	5
work on ethics-related topics as ethics in their work	Discussing work focused on ethics-related topics (e.g., FACCT topics) as the representation of ethics in their work.	13	37
previous work	Noting that ethics was more salient in their work when they worked on specialised projects, but now there has been a long time since they talked or did something related to ethics.	2	2
ethics about data	Captures codes related to the representation of ethics as data and data practices.		
data & ethics	Connecting ethics to data: saying ethics is about data, noting data-related issues as examples of concerns. This is more general and about various actors compared to the concrete examples in own practice in another code.	16	20
misuse of data	Mentioning misuse of data as a potential ethical issue (noting since misuse is usually about applications).	2	2
data as criterion for relevance of ethics or ethical concern	Narratives connecting the type of data used or not used to whether a project or work in general has ethical concerns, needs to express or deal with them, etc., and the degree of ethical valence.	20	40
ethically relevant data	Connecting the relevance of ethics to the type of data they use. Includes mentions of 'real', 'human', and 'personal' and 'sensitive' (another link to data protection).	5	5
relating my research and their research re human subjects	Making a connection between my research and theirs based on the fact that both involve human subjects and related considerations.	3	4
no personal data – less or no ethical issues	Narratively connecting the non-use (absence) of personal data (a data protection category) with not having ethics issues.	4	6
public data = free of issues	Expressing that using public and open-source data means the project does not have ethical issues. Includes mentions of then (not) needing approvals, expressing that data is public as a reason not to have ethical issues in statements or needing further discussion.	7	10

Name	Description	Cases	References
data work as ethics in their own practice	Discussing their data-work practices (collection, wrangling) as examples/representations of ethics in their work, with concrete examples These excerpts then also show ethics as something that is part of their regular practice.		
annotation (but for data not annotators)	Describing choices that get encoded into annotation schemas and data as an example of ethics.	2	2
communicating data limitations biases	Mentioning communicating potential biases or limitations related to data as examples of ethics in their practice.	2	3
designing and doing data collection	Describing designing and executing data collection as examples of doing ethics in their own work. The 'design' part also includes protocols and preparing for ethical approval, but did not make sense to separate that as they discuss e.g. deciding on phrasing things for participants, so things connected to how data generation would go.	9	29
designing data infrastructure	Data infrastructure (hardware, moving data around) etc. as examples of ethics.	2	2
doing data protection and compliance	Invoking communication with data protection and compliance specialists as examples of doing ethics.	5	6
treating data workers, annotators right	Mentioning their (also company's?) practices of caring towards annotators as examples of ethics in their work.	3	4
wrangling preprocessing	Includes the preparation of data: cleaning, wrangling, filtering, etc.	6	7
excluding masking cleaning data	Describe practices of selecting data NOT to be used, cleaning, masking, removing, filtering etc., as ways of making data that will be used in the analysis ethical. Removing personal data + removing or masking features that could lead to bias or deanonymisation (and associated risks for people featured).	4	7
ethics and data protection intertwined, bundled	Captures narratives where ethics and data protection are connected, doing one is doing the other, or they are hard to distinguish. Hence also includes the excerpts from the code that is about doing data protection as examples of ethics in their own work, but also other narratives (e.g., about others or ML more broadly).	10	11
problematising public data(sets)	Expressing that public datasets actually have ethical problems. A bit of a countertrend compared to the codes that portray public data as free of issues.	3	8

Name	Description	Cases	References
WCTM about data	Invoking themes and concerns related to data in response to the "What comes to mind when you think about ethics?" question.	20	22
ethics in practice	Pre-themes code that captures ethics in their practice broadly. Based on stories and their answers to questions like: What do they see in their work that is doing ethics? In what contexts do they encounter and 'do' ethics?		
company ethics, what they do towards ethics as a company	Describing their company's practices related to ethics.	4	4
ethics match in hiring	Describing that when they got hired, the company checked on similar ethics (the opposite/counter of descriptions that they choose work based on ethics, here it is companies choosing them).	2	3
conversations about ethics	Captures them describing talking about ethics.		
ethical conversations they had	Captures stories/examples of conversations regarding ethics.	12	21
how often or much they talk about ethics	Captures their descriptions/qualifications of the prominence of ethical conversations in their lives.	13	17
what is talked about as ethics with colleagues	Captures topics or issues covered in conversations about ethics with colleagues.	12	16
division of work re ethics	Descriptions of who does "ethics" part of the work (here of course there is a confusion between ethics and approvals and other procedures): them, domain experts, etc.	9	11
mental division of labour	In vivo code.	1	1
from where they (are) learn(ing) about ethics issues	Captures them describing the sources from which they hear and learn about ethics (often I asked about this specifically).	25	49
when ethics is addressed in the project	When in the project - stage or scenario - ethics is discussed or is something to obviously think about.	11	21
ethics is about application	Codes capturing the representation of ethics as considerations for applications of ML.		
considering potential issues of intended application (-)	Discussing things that could go wrong and/or be harmful or problematic within the intended (good) application rather than misuse for different malicious ones.	3	11

Name	Description	Cases	References
considering potential misuse for bad applications	Codes related to (potential/downstream) negative applications, also called misuse.		
(lack of) control over ethicality of application	Expressing that they personally and/or researchers and developers in general have little control over how their outputs (models) will be used in the world, and whether they would only be used in ethical, good ways.	6	8
can only mention potential misuse	Noting that they do not have control about how other people would use their models, they can only articulate the potential issues.	3	5
release and licensing as practices	Mentioning licensing and release (to open-source or not) as practices or solutions for preventing misuse. However, the mention of licensing includes that malicious actors would not be stopped by licensing.	2	2
ethical issues = misuse	Discussing ethical issues as issues and misuse/use of the technology outside of its original context, for malicious purposes. Includes narrative associations (e.g., I ask about ethical issues, and they mention misuse), making kind of chains of equivalence between ethics and ethical and misuse, and other narratives where ethical issues are equated to or represented as misuse issues.	9	14
need governance to control release of models with potentially harmful applications	Expressing that there needs to be some governance (rules, guidelines) for ML development and companies to consider and limit potential negative applications.	1	3
researchers and developers should consider potential applications more	Expressing that ML researchers and developers should consider what their work is/could be used for.	3	6
ethics of the whole topic	In vivo code, expression noting that in ethical statements people consider the overall ethicality of the research topic based on the range of potential applications.	1	1
linking type of research & negative implications	Captures links made between research or project topics and avenues towards negative applications. Indicates known, visible link and the 'location' of these potential negative impacts in type of topics. NB several of these are in discussions of ethical statements.	6	13

Name	Description	Cases	References
potential bad uses of their (type of) models	Describing how their models, or models of similar type could potentially be misused, for what kinds of applications. Notably, these applications mostly ascribed to types of models or subfields. Even when talking about 'my model', mentioning the general topic and characteristics (that would be shared by other models).	7	11
topics that should not be studied	Describing topics the interviewees think should not be studied because they are problematic, unethical. This is for the field/machine learning in general, not only themselves: for instance, includes mentions of papers with problematic topics that they think should not be published.	5	8
ethics of their own application	Includes codes related to narrating the potential applications toward which the interviewees' work is oriented or for which it can be used.		
applications argument for own work being not problematic	Invoking explanations related to applications as reasons for own work not posing (much) ethical issue or concern. Also includes the "counterevidence" code for when practitioners invoke a type of model, argument as such an explanation, but 2 out of 3 there also connect to applications.		
difficult to misuse	Describing that they find it difficult to come up with a misuse scenario for their specific project.	2	2
far from deployment	Noting that their work is on methods and thus is far from deployment, which is where/when concerns and potential issues would come up and need to be addressed.	4	5
fundamental research	A subset of the 'far from deployment' argument which specifically invokes the notion of basic or fundamental research.	2	4
good or nonproblematic application	Describing the application are they work in or are oriented towards as a positive, non-problematic application. Includes both explicitly good/positive and more 'benign' descriptions.	5	7
method can be used for different things	Noting that the methods they develop can still be "picked up" in both good and bad ways.	2	3
safe(r) type of modelling	Invokes types of models or tasks: time series modelling, pattern discovery, classification and summarisation. Notably, the time series then is still connected to messing with the market (see "hard to misuse"), and pattern discovery contrasted with "decision-making and replacing people", which still involve uses/applications.	3	4

Name	Description	Cases	References
avoiding problematic topics applications as ethics of their work	Describing that they avoid work on certain applications, would rather not work on them as their ethics, examples of what ethics in their practice is. This code also includes which topics or applications they avoid. Includes both more general "would/would not" thinking and a few examples of concrete choices (e.g., projects, job offers).	9	22
fault in the method – ML method motivates data extraction (-)	Noting that the methods developed in ML now kind of motivate companies and other actors to hoard data because they can be modelled. Also kind of a counterevidence since it ascribes fault to the method.	1	1
good and bad application areas	Invoking and describing good and bad (unethical) applications, application fields, or areas. Includes making distinctions and contrasts between application areas.	12	23
locating & fixing ethics in the model (-)	Invoking ethics as something that is in the model and can be fixed by tinkering with the model.	1	1
ML can be used for good and for bad	Expressing that machine learning (and even the same models) can be used for good as well as malicious purposes/applications.	10	15
can be used for bad but can also help understand and deal	Expressing that ML research and development can be used for potential negative applications, and this is a risk/danger that exists with knowledge or technology. But doing this, producing this knowledge/technology can also help understand and help deal with the same negative applications.	3	3
if it's not me, it's going to be someone else	Indicating that if they don't develop this technology (which could potentially be used for unethical applications). Also connects to the notion that this way at tleast there would be knowledge.	3	3
WCTM = application, ethics is about how a model is or can be used	Invoking applications or use of the models/ML when answering the 'What comes to your mind when I ask you to think about ethics' question.	11	13
how data is used	Invoking the purpose, for what data (and not models) are or can be used. Partly similar (use) but also counterevidence as it is about data rather than models, but approaches in a similar way (rather than granular practices).	2	2
ethics through connection to personal relations	Connecting their thinking about ethics to people they care about.	**2**	**2**

Name	Description	Cases	References
formal ethics	Everything related to formal or procedural ethics: assessments, statements, approvals.		
considering changes in formal ethics re ML research over time	Describing their experience with how approaches to ethics have changed in the field over time. Often is about applications and statements.	5	9
contributing to formal ethics	Describing their work on/in formal ethics as part of their work.	2	3
EDI	Mentioning the EDI (Equality, Diversity, Inclusion) framework when discussing ethics.	1	1
ethical approval	Codes related to the ethical approval (includes both university committees and national medical ones).		
discussing experiences with approvals	Discussing their experiences with the ethical approvals procedure. Includes discussion of how/why they did not get one.	14	45
time it takes to get ethical approval	Describing the time it takes to obtain approvals. This is both the preparation of the application & getting a reply.	6	11
who works on getting ethical approval	Describing the distribution of labour for obtaining approvals (often on domain specialists if in collaboration).	9	15
ethical statements and checklists (conferences)	Codes capturing the interviewees' experiences with and attitudes to 'ethical statements' written as part of conference submissions.		
discussing ethical statements	Describing the ethical/impact statements practice, their experience with it, and what they think about it.	5	10
for some topics ethical statement matters more	Differentiating projects or types of papers in terms of how much detail or discussion in ethical statements they need – based on topic.	2	3
good, they make people to think about ethical implications	Expressing that the good thing about the statements (or what makes them a good thing) is that they make (some even say force) researchers think about potential negative implications of their work (impact/applications).	5	5
ethics questions in the submission checklist	Describing submission checklists (ACL NLP and NeurIPS) as places where they encounter and do (discuss) ethics.	4	7
ethics statements as representation of ethics in ML	Narrative association of ethics and ethical statements. For instance, bringing up ethical statement in response to me asking a question about ethics and ethics in the field.	5	6

Name	Description	Cases	References
what they wrote in ethical statements	Describing what they wrote in ethical statements.	9	13
ethics course	Describing courses on issues related to ethics at their organisations (includes both academic for PhDs and non-academic, at companies).	9	13
ethics evaluation in peer review	Describing their experience with considering the ethicality of submitted papers and potentially flagging papers for ethical concerns. Mentions problematic applications (e.g., surveillance) mostly, and twice data issues.	6	14
ethics policies	Mentioning the existence of ethics policies from conferences and organisations.	4	8
FAIR principles	Mentioning the FAIR principles framework.	2	3
struggling with term or definition of ethics	**Expressing that they find it difficult to define 'ethics' or say what exactly it entails. Some excerpts also connect this to the lack of expertise.**	**4**	**6**
is this ethics or not	Mentioning a topic or concern and saying they are unsure whether that is included in ethics.	3	4

B: Coding list for conference materials

Name	Description	Sources	References
data as particular project practices	Asking particular questions about the data used in the submittted research and the data practices (e.g., compensation of participants). Also include FAQs.	**8**	**16**
ethical statements	Excerpts about the 'ethics" or 'impact' statements (or similar discussion suggested to be in the paper or in the checklist).	**9**	**9**
applications	Connecting the potential (negative) impacts to potential applications of the research.	9	10
examples of links between topics or methods, applications, and negative impacts	Providing examples of how certain topics imply a path to applications and potential negative impacts. For example, generative models -> potential use for deepfakes -> misinformation.	3	5
misuse	Mentioning potential misuse or the produced technology being used for malicious applications. Notably, most of these mention that researchers should write if THEIR research products or approach could be used for any malicious applications. So the connection to level/subfield is in the examples and in the differentiations between types of papers.	6	7
attributing to methodology (-)	Distinguishing between the implications of methods and applications, and attributing certain impacts to the methods.	2	2
broader set of potential concerns (-)	Capture a portrayal of broad range of topics and concerns that could be discussed in the ethics statement, with misuse as one of many (if even there) rather than focus (also in terms of how much it is discussed compared to other topics).	3	3
impact comes not from one project		1	1
mitigation (-)	Asking researchers to discuss mitigation strategies they propose/use against potential negative impacts. Not required but asked.	4	4
some papers require more reflection in the statement	Discussing that some papers need to pay more attention to the ethics/impact statements.	0	0
basic research	Noting that basic research might not need this discussion because it is not connected to applications.	2	2

Name	Description	Sources	References
implications established within ML research in general, ICML	Captures the argument that many papers would have "societal implications that are well established when advancing the field of Machine Learning". ICML then gives the option of the template statement (also mentioned by interviewees).	2	2
relationship to application	Attributing the necessity or not (and expected level of discussion) depending on the project's link or path to problematic applications.	4	6
two-sides double use	Mentioning that applications can have good and bad sides or be double-use (both good and bad).	4	5
invoking (other) governance	Mentions of other formal ethics or governance approaches. Introduces these as relevant categories and converges them, following one involves following the other.		
data governance	This code includes data protection (GDPR) mentioned as well as copyright and personally identifiable information. Is invoked in connection with personal data, protected information, sensitive data.	6	8
ethical approval	Mentioning ethical approval or committee (IRB) procedure. Is invoked in connection to human participants.	14	21
the AI Act	Mentioning that the submitted work must be compliant with the AI Act.	1	1
topics that would not be accepted	Mentions of topics and types of research that is not accepted by the conference.	**1**	**3**
other themes and values, concerns	Captures mentions of other topics or issues (in statements and not) interviewees have connected to ethics.		
energy and compute use	Captures mentions of energy and compute use (to be discussed).	2	2
fairness and inequality	Captures mentions of concerns for fairness and inequality. Includes mentions of vulnerable populations.	4	6
licensing, copyright for code models	Questions and suggested practices for copyright for models and code (for data, is included in the data governance).	2	2
limitations scope certainty	Captures aspects of research regarding made claims and their limits and certainty.	4	4
openness reproducibility	Captures instructions and questions regarding open-sourcing and reproducibility of research.	7	12
privacy and security	Captures mentions of privacy and security considerations.	6	8

Name	Description	Sources	References
prohibited, deprecated datasets	Mentioning datasets that should not be used in submissions (with some exceptions when they can be used for comparison). This is mostly from CVPR (+1 NeurIPS mention).	6	6

Appendix VII: Coding list for Paper 3

Name	Description	Sources	References
accessing the data	Codes capturing the experiences of capturing biomedical data and the challenges/tensions there.		
easy data access	Describing getting access to data as (relatively) easy. Two of these are 'open' repositories.	3	6
happy to avoid colleagues' data access issues	Describing an easy process with access and being 'grateful' to avoid issues that their colleagues complain about.	1	1
long time to get access to data	Describing that it takes a (too) long time to get access to data and describing this as an issue.	5	9
not getting access to data after all	Describing projects where they did not end up getting access and/or abandoned the project because it took too much time and effort to get the data (without making progress).	3	5
negotiating access	Describing access as something they negotiate with compliance or data stewards (who control what data they get, e.g. anonymization and filtering).: e.g., need more variables or more data.	3	5
positionality of access	Describing that one needs to be located in a certain group or organisation, or establish collaborations (which is time and work) to gain access to biomedical data.	5	11
request application for access	Describing the process of applying for access to data with health organisations,	7	18
storing heath data	Describing how they store their (biomedical) data. Seems like practices vary a lot, but it seems like there is more flexibility when it is new, own data rather than reusing data from healthcare organisations. Hospital data has to stay with the hospital, on their servers.	13	21
The HOW of access	Describing an issue with access that happens when you are supposedly granted and got access. Is about how that access works for the practitioners and what they can and cannot do with the data.	4	9
cannot publish the data	Describing that these data cannot be published/open-sourced as part of the publication (personal, sensitive, etc.). Access in terms of how you can use, what can you do with them.	2	2

Name	Description	Sources	References
only certain features or variables	Describing role-based access where they get access to data but not all data (e.g., without some variables)	2	6
restriction on models or computation due to data	Captures descriptions of restrictions in terms of what they could do with the data regarding modelling (because of how they can access the data or where they can move/store it).	5	6
challenges of working with biomedical data	Captures instances when they describe biomedical data as difficult or describe challenges/issues with data.	14	35
contrasting methods ML with biomedical ML	Making contrasts between working with data in methods or fundamental ML (usually on open data) and in the biomedical context/domain.		
access (safeguards compliance)	Contrasting methods and biomedical ML based on how one accesses the data and what they can or cannot do with it.	2	2
data work	Contrasting in terms of how one has to prepare data for modelling and account for it in modelling and representing the data.	3	12
knowing the data and how to work with it	Codes capturing working with the data and what one needs to take into account and know, and issues that can arise there.		
adaptation to biomedical data domain	Describing them having to significantly adapt how they work with data and models when moving to the biomedical domain, spending time and effort.	2	7
having to learn more than expected	Describing learning things they did not expect to matter, and having to get into the domain knowledge and data details more than expected.	4	10
classification problem vs. understanding sleep	Describing that how much one needs to know about the data and labels depends on how they see the problem. If it is just a 'classification problem', then bare minimum, otherwise need to dig deeper.	1	2
data qualities features			
imbalanced data	Describing biomedical data as often imbalanced and with potential for bias. Often these descriptions express that this is not a collection issue but that the phenomena in the world (and thus data produced about them) are imbalanced, e.g. who gets certain diseases.	8	16

Name	Description	Sources	References
not clinical, straight genomic	Describing not having bias issues because this is genomic data straight from the patient rather than having been generated by people and around people.	1	2
limited data sizes	Describing that there is limited data in their project or on problems they work in: either as a problem, or something to take into account in modelling.	4	5
noisy artifacts		2	3
variation uncertainty in human decisions labels	Describing that the data come from human decisions and labels, and that includes subjectivity, variability etc. Some describe it as more of an issue for modelling.	4	11
labelling reflecting real social processes	This code captures something a little different: also about labels and them being uncertain, but less with the expectation that they should be. More that the labels reflect complex healthcare processes (which they also do not control, so another code capturing the 'data reflects the world rather than what we need').	3	10
needing to know 'surprisingly little'	Describing that they need to know 'surprisingly little' and one picks up more biomedical knowledge than they need. A bit of counterevidence compared to those who describe needing to know more.	1	1
relationship to the biomedical knowledge and domain (experts)	Codes capturing practitioners' relationship to domain knowledge (less directly connected to data) and collaboration with domain experts.		
asking the specialist rather than learning themselves	Describing relying on the specialists' input for making decisions within the project, saying they know little about the topic.	2	2
being interested in domain topics and knowledge	Describing that they are (and/or have been) interested in the topics they work on and learning about the biomedical subject matter.	4	6
learning about the domain	Described how they were or are learning about the subject matter and topics involved in their research.	6	6
learning by exposure	Describing that they have learned about the domain subject matter via and for the collaborative relationship with the biomedical researchers/clinicians.	7	10
need to know to make sense to domain specialists	Describing that they need to know the biomedical subject matter so that the results of their work make sense to and are useful for the domain specialists.	2	2

Name	Description	Sources	References
regular meetings collaboration	Describing being in close contact with the collaborators from the domain side, with regular check-ins (from monthly to weekly) and doing things together.	4	7

Appendix VIII: Generative AI use declaration

During the preparation of this manuscript, the author used DeepL Write (free version) and Microsoft Copilot (basic version available to AU employees) to improve the spelling and grammar of parts of the text. The Danish version of the summary was first automatically translated with 'GPT 5.6. Think Deeper' and then edited.

All AI-assisted edits were reviewed by the author, who takes full responsibility for the content of the manuscript.