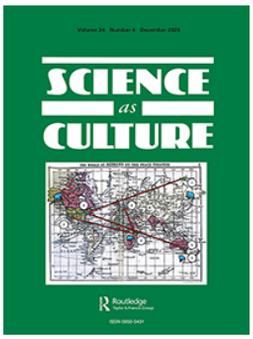

Science as Culture

ISSN: 0950-5431 (Print) 1470-1189 (Online) Journal homepage: www.tandfonline.com/journals/csac20

Making problems: interdisciplinary collaboration and AI ethics

David Moats, Tjitske Holtrop, Nees Jan van Eck, Judit Varga, Ludo Waltman & Francien Dechesne

To cite this article: David Moats, Tjitske Holtrop, Nees Jan van Eck, Judit Varga, Ludo Waltman & Francien Dechesne (25 Nov 2025): Making problems: interdisciplinary collaboration and AI ethics, Science as Culture, DOI: 10.1080/09505431.2025.2584759

To link to this article: https://doi.org/10.1080/09505431.2025.2584759

Published online: 25 Nov 2025.

SCIENCE AS CULTURE
https://doi.org/10.1080/09505431.2025.2584759

Routledge
Taylor & Francis Group

Making problems: interdisciplinary collaboration and AI ethics

David Moats [a], Tjitske Holtrop [b], Nees Jan van Eck [b], Judit Varga [b], Ludo Waltman [b] and Francien Dechesne [c]

[a]Faculty of Social Sciences, University of Helsinki, Helsinki, Finland; [b]Center for Science and Technology Studies (CWTS), Leiden University, Leiden, Netherlands; [c]Center for Law and Digital Technologies (eLaw), Leiden University, Leiden, Netherlands

ABSTRACT

The impacts of AI and algorithmic systems on society are often spoken of as a grand societal challenge and it is increasingly understood that cultivating more fair, accountable, or transparent uses of AI, will require the inclusion of more diverse voices (social scientists, philosophers, users) in the development process. Yet, within the emerging field of AI Ethics, it is frequently observed that these disciplines speak different languages. What computer scientists mean by 'fair' (expressed mathematically) does not necessarily correspond to legal, philosophical or colloquial definitions of the word. Such disagreements are often presented as a barrier to problem solving. Yet others argue that disciplinary differences, including divergences in the meaning of words, are rather an opportunity: the frictions created when ways of knowing rub up against each other are necessary to develop new problem formulations. However, literature from STS around care in interdisciplinarity, suggests that these frictions must be handled carefully – attending to relations between parties. Through collective reflections on an interdisciplinary project combining the social and computer sciences, this paper illustrates how the framing of the problem evolved *as a consequence* of the curation of disciplinary disjunctures. The paper contributes to STS studies of interdisciplinarity in practice by thinking through how to recognize and foster productive disciplinary disjunctures with care. This has implications for how the AI Ethics community and policy makers can think about addressing grand societal challenges through interdisciplinary research.

ARTICLE HISTORY
Received 14 June 2023
Accepted 25 October 2025

KEYWORDS
Interdisciplinarity; AI ethics; data visualization; collaboration; problems, care

CONTACT David Moats david.moats@kcl.ac.uk Department of Digital Humanities, Kings College London, Strand, London, UK, WC2R 2LS

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

Introduction

Alongside climate change, ageing populations, energy security and inequality, harnessing artificial intelligence (AI) and data driven systems is increasingly recognised as one of the grand societal challenges (HM Government (UK) 2017). For some policy makers, AI is merely a tool to address these challenges (NAIRR, 2023) but for many, technically achieving artificial general intelligence and integrating it into business and society is *itself* such a challenge (NITRD, 2019). This is because, as AI systems are increasingly used to assign credit scores, identify people using facial recognition, or deliver medical diagnoses, these systems can lead to discrimination or uneven outcomes, exacerbating existing social inequalities (O'Neil, 2016). A speech by Tim Gardam of the Nufield Foundation (Gardam, 2019) even claims that AI is more complex and unwieldly than previous challenges.

> Today's challenges are of course even more complex [than biomedical ones] as the ethics surrounding the development and use of data driven systems and artificial intelligence involve scenarios well outside the long established institutional and normative frameworks of medical science. […]
>
> These questions break down intellectual and disciplinary barriers – between social scientists and data scientists, both of which also now have to take account of the underpinning disciplines of the humanities – moral philosophy and practical ethics. (Gardam, 2019, n.p.)

It is increasingly acknowledged that these challenges span different disciplinary remits and thus *demand* interdisciplinary or even transdisciplinary collaboration (Maniglier, 2021). Indeed, these perceived problems with AI have led to the formation of a vibrant interdisciplinary space, sometimes referred to as AI Ethics. Often practical questions of which disciplines need to be included or how they are to interact are left implicit in policy discussions. However, Gardam's speech notes one key practical difficulty in including multiple disciplines in the conversation about AI:

> Before one can identify the ethical co-ordinates to guide us through the risks of a data driven world, or indeed find the routes to well-being that the power of these technologies makes possible, we must first delineate and define key concepts and values, and then negotiate the tensions and trade-offs between them.
>
> Only then can we resolve the ambiguities inherent in the way we use such key terms such as consent, bias, ownership, explainability, and privacy. […] Some clarity of definition has to be the precondition for building a more rigorous evidence base on which to construct possible answers to the dilemmas raised, or, at the very least, to bottom out the nature of disagreements. (Gardam, 2019, n.p.)

In short, Gardam posits that the divergent meanings of key terms in this space is a *barrier* which will impede interdisciplinary collaboration or dialogue. For example, what computer scientists mean by fair, as it can be expressed

mathematically, does not necessarily correspond to legal, philosophical or colloquial definitions of the word (Washington, 2018). Gardam suggests that some sort of shared understanding around these terms is a pre-requisite for collaboration. This is echoed in many other reports about AI as a societal challenge (Association of Nordic Engineers 2020; European Commission 2020; IEEE 2020), including one by the Leverhulme Centre for the Future of Intelligence (Whittlestone et al., 2019), who call for identifying tensions and building 'consensus' between perspectives (not just between disciplines but between academia and different societal stakeholders). It seems to be widely assumed that disciplinary tensions need to be resolved before work on these challenges can progress. But what role do disciplinary tensions play in interdisciplinary collaboration? How do these tensions relate to the problems being addressed? How are terminological/methodological/theoretical differences negotiated in practice, within interdisciplinary teams? What role should policy have in managing these disciplinary interactions?

In this paper, we argue that while disciplinary differences, such as the meaning of key terms, may cause disagreements and misunderstandings, they should not be thought of as a barrier to be overcome, as in the formulations above, but rather as an *opportunity* to further develop thinking. The frictions which are created when ways of knowing rub up against each other may, indeed, be necessary to develop new *articulations* of the research problem (including those big research problems branded grand challenges) and thus to new potential solutions.

To illustrate this point, we draw from our experiences working on an interdisciplinary project attempting to map the emerging field of AI Ethics using Scientometric tools, focus group discussions and textual analysis. We will not go into detail about the findings of this project here but, through collective reflection on the process, we show how shifts in the objectives and direction of our project actually emerged from moments of disciplinary tension. While we should caution that our project was not a typical example of work in the AI Ethics field – we were attempting to *analyse* the field as a whole, rather than to develop or evaluate an autonomous system ourselves – our team is representative of the disciplines involved (coming from Sociology, Science and Technology Studies (STS), Law, Philosophy, Computer Science, and Anthropology) and our project involved both computational and interpretive methods. Thus our experiences should be highly relevant to AI Ethics projects.

We start by discussing literature on inter – and transdisciplinary collaborations in relation to grand societal challenges, particularly in STS. We then draw on the work of Maniglier (2021), who argues that disciplinary tensions are crucial for developing new formulations of the research problem. However, we put Maniglier's argument in dialogue with STS work emerging from studies of interdisciplinary collaborations in practice, which emphasizes the importance of relations of care when dealing with such tensions. From

this work, we draw on the notion of 'disconcertment' (Verran, 2001) to help recognize and explore these disjunctures though collective reflection. We also discuss STS studies of interdisciplinarity in practice to think through how these disjunctures could be fostered 'with care.'

We then go on to describe the development of our project, focusing on four key moments in which our disciplinary differences came into sharp relief and led us in interesting new directions. These turning-points included: articulating the problem, delimiting the object of study, analysis, and the presentation of results. Along the way, we reflect on practices that might help care-fully surface productive frictions and explain how these shifted our understanding of the problem our project was meant to address.

Finally, we discuss the implications of our collaboration for STS discussions of interdisciplinarity and the field of AI Ethics. We argue that carefully fostering disjunctures between ways of knowing can be instrumental in transforming our understanding of grand societal challenges and we offer practical lessons for how to do so at the intersection of the computational and social sciences. We suggest that if AI Ethics researchers embrace, rather than sublimate their disciplinary differences, they may gain a more nuanced understanding of what societal challenges they are engaged with. We also suggest that policy makers and funding bodies might foster bolder, more generative research not by resolving differences between disciplines but by provoking them.

Literature

Interdisciplinarity

Scholars have coined diverse terms that describe the meeting and combination of disciplines, including intradisciplinarity, crossdisciplinarity, multidisciplinarity, interdisciplinarity and transdisciplinarity (Stember, 1991). The meanings of these terms, however, have changed over time, and have been subject to ongoing debates (e.g. Klein, 2004; Jensenius, 2012). For the purposes of this paper we will focus on interdisciplinarity, understood as combining knowledge from disciplines that may relate to each other in different, more or less symmetrical ways (Weszkalnys and Barry, 2013). We understand disciplinary boundaries as already porous but, also, not something which can (or should) be dissolved completely (Strathern, 2004; Osborne, 2015). In addition, we take from transdisciplinary discussions that there may be ways of knowing from outside of academic disciplines (users of technologies, activists, NGOs) whose inputs are worth including in collaborations to address societal challenges (Nowotny *et al.*, 2001).

As the editors of this special issue have shown in their introduction, there are many open questions about the role of interdisciplinarity in addressing societal challenges. Empirical work in STS has observed that the way scientists or other

actors like policy makers and research funders define these challenges not only entails certain types of solutions at the expense of others, these definitions also suggest who is best placed to provide these solutions (including which disciplines) (see Callon, 1980 on research problems). For STS there is a politics to how problems are defined. Diedrich et al. (2011), for example, argue that the current framing of the problem of climate change by the EU commission presupposes technical solutions as opposed to, say, behavioural or social interventions. It also presupposes industrial actors or partnerships as the ideal actors poised to solve the problem as opposed to, say, social science disciplines or different types of stakeholders.

However, there is another difficulty in the way policymakers conventionally depict societal challenges. A recent paper by Maniglier (2021) starts from the observation that many academic institutions, governments and funding bodies currently argue that science should be arranged around 'shared problems,' not disciplines. Rather than the world being carved up into the objects of various fields: matter (physics), social phenomena (sociology), the mind (psychology) etc., science should start with problems existing in the world and ask which disciplines or approaches are required to solve them.

Maniglier notes that this is a simplistic way of understanding problems as merely obstacles to be overcome. It also assumes that problems exist entirely out there in the world, independent of scientists' practices. Instead, following Bachelard and Deleuze, he offers a more nuanced way of thinking about problems. He argues that an important part of thinking is about creating new problem articulations (problematisations) which develop our understanding of the problem, and the prompt for constructing new problematisations often comes out of disjunctures between ways of knowing. For example, a new problematization can be generated by understanding the difference between how two disciplines construct 'relatable' problems:

> … problems are not given outside, in the real world, but in between, in the morphisms by which a discipline or, more generally, a particular way of thinking, experiences the possibility of its identity with a completely heterogeneous way of thinking, without this identity being realized elsewhere than in the dynamic of their divergence. (Maniglier, 2021, p. 42)

Productive (interdisciplinary) research may *require* disciplinarity because it entails understanding differences in how other disciplines define terms, pose questions and create explanations, in contrast with one's own discipline. Disciplinary tensions are thus occasions to re-articulate the problem for the project which may generate new research questions and suggest possible solutions. This is a fascinating proposal which is compatible with STS's attention to the framing of problems as outlined by the Special Issue editors, but adds to it a different way of thinking of the relationship of disciplines to problems. The initial framing of problems (including challenges) can shape which disciplines

are able to contribute and on what grounds but the interaction of disciplines may also be essential to shaping or developing problem formulations.

Maniglier's account also seems particularly appropriate to situations, like in AI Ethics, where interdisciplinary encounters have not (yet) developed into established 'trading zones' (Galison, 1997) – in-between spaces where collaboration can occur, facilitated by a hybrid 'creole' or 'pidgin' language. Rather, in AI Ethics, interdisciplinary interactions develop around diverse objects including methods, concepts, tools, sometimes with shared, but oftentimes divergent languages, understandings of the problem and required solutions (e.g. Coeckelbergh, 2020). To apply Collins, Evans and Gorman's (2019) update of the trading zone concept, there is, arguably, not yet a shared language in AI Ethics and disciplinary practices can remain quite separate, but there may be 'ambassadors' or intermediaries trying to facilitate communication between groups. If this semi-shared language is still in development, as the quotes at the beginning of the article suggest, then forced confrontations where differences are reflected on may be the only way to proceed.

However, Maniglier's proposal does not account for the practical labour required for developing disciplinary dialogues through disjunctures and it might understate the potential consequences and politics of this practice. Not all researchers or research participants are equally affected by frictions and conflict, however well-meaning their deployment might be. For this, we might turn to recent work in STS on care in collaborations (e.g. – Latimer and López Gómez, 2019). Drawing on a feminist ethics of care (Tronto, 1987; Gilligan, 1993), concerned with acts of tinkering and repair (Mol et al., 2010), these authors use the care concept to argue for ways of managing both internal and external relationships by seeing different parties as relationally interconnected and attending to the often neglected aspects of affect, the body and the invisible labour involved in managing complex projects (Davies and Horst, 2015). This literature reminds us that if we are going to actively create these disjunctures, this must be done 'care-fully': by which we mean – keeping in mind the different consequences of actions for all the collaborators and participants (as well as for the work itself) (Groot et al., 2019), tending to interpersonal relationships and, in research, cultivating variety of knowledge practices.

Although there is a danger in care literature of simply valorising the interconnectedness of all humans (and non-humans), many authors acknowledge the inevitability of friction and difficulty as an essential aspect of care relationships. Latimer uses the concept of 'being alongside' (Latimer, 2013) to describe cooperative relationships where difference is not erased and Puig de la Bellacasa (2012) acknowledges that some caring relationships might take asymmetrical forms like 'thinking for' someone (or something) less able to articulate their thoughts. Still, we would like to emphasize more, following Maniglier, that some frictions might be helpfully *encouraged*, not just tolerated or managed.

This work on care, by focusing on affective and bodily aspects of interactions often references Helen Verran's (2001) concept of disconcertment, when 'things feel wrong and unsettling' as key prompts for reflection on epistemological and ontological differences, especially in knowledge practices in post-colonial situations (Law and Lin, 2010). This concept helps us nuance what we have been calling disjunctures and helps sensitize us to how to spot them.

Finally, Maniglier emphasizes that problems are not merely out there, waiting to be solved, they are at least partly articulated by us. Puig de la Bellacasa (2011) suggests we need more of a critical edge when attending to problems, which might necessitate 'speculation about neglected things' – for example, in articulating problems, we might need to actively seek out marginalised voices or relationships which are not already called forth by problems.

In any case, much of the above literature refers to collaborations between the natural sciences and humanities. Do these ideas have the same purchase in the social and computer sciences? If so, can we develop insights about how we can foster productive and care-ful disjunctures within interdisciplinary collaboration between the social and computer sciences?

Interdisciplinarity in practice

To answer these questions, we can draw on literature in STS which seeks to study 'practical interdisciplinary work,' as the special issue editors have called it, particularly around digital data and algorithms (Blok *et al.*, 2017; Neff *et al.*, 2017) in practice. This work acknowledges interdisciplinarity is never just the meeting of two pre-existing disciplines, disciplines are made and remade through the interaction (Callard and Fitzgerald, 2015), just as individual researchers are not normally one discipline or another, they have hybrid existences and are situated in existing histories of disciplinary struggle (Zuiderent-Jerak, 2021). More importantly, these situated, 'warts and all' accounts of collaborations demonstrate that disciplinary relations must be addressed in ways particular to the situation or the disciplines in question.

Kaltenbruner (2014), for example, describes how a collaborative project between computer scientists and humanities scholars repeatedly floundered until the different parties 'decomposed' their research processes to each other – reflexively explaining the component parts such as what they mean by research questions or how they understand 'quality.' This shows that sometimes the essential and productive collective examination of knowledge practices only proceeds after a confrontation or a breakdown of sorts. This paper also builds on the work of one of the authors, Moats (2021) who organized a series of interdisciplinary workshops centred around digital tools, such as data visualisations. He observed that disciplinary tensions were often managed through a conventional division of labour between disciplines, but when these routines were

disrupted, more interesting research directions emerged (cf. Fitzgerald and Callard, 2015).

Following this work, and our theoretical framework, we attempted to proactively trouble our normal disciplinary roles (for example, involving qualitative researchers in query design and having computer scientists do qualitative analysis). Some of this space for play and experiment was possible because the project was part-time and informal for most of us, a passion project not subject to deadlines and deliverables. At the end of the project, in an effort to test Maniglier's proposal (2021) each of us wrote down a list of the key turning points – when the project changed direction – and considered what prompted this change. There was some consensus about what the key turning points were (such as a reorientation of our use of data visualizations) as well as which events precipitated them (for example most of us recalled a debate about how best to collect our data and a pivotal zoom meeting where we discussed some non-functioning visualizations) but, interestingly, each of us saw different aspects of these encounters as important. It quickly became clear that, as Maniglier proposed, each of these turning points could be traced back to a moment of tension between knowledge practices, which we might now, following Verran (2001), call a moment of disconcertment, and each turning point, crucially involved a rearticulation of the research problem. In other words, these feelings of unease, experienced differently by each of us, prompted reflection on what we were doing in the project. Through writing this paper we attempted to synthesize these encounters as a collective narrative, but without completely flattening out differences between us (being alongside in Latimer's language). This chimes with other proposals to interrogate disciplinary relations through tasks like diary writing (West and Schill, 2022) but more focused on the collective thinking process. This practice also allowed us to speculate on why our disagreements were productive, rather than ending in stalemates or hurt feelings, which later led us to think about care.

In the next section, we will narrate our project and, in the process, explain more about the emerging field of AI Ethics.

Narrative of the project

The project started in Spring 2020 when Moats, a specialist in STS and Digital Sociology, received a small grant from the Dutch Research Council (NWO) to visit Waltman and van Eck, specialists in Scientometrics based at the Centre for Science and Technology Studies (CWTS), Leiden University. Scientometrics is the quantitative study of science, often using data from scientific publications. The three were all interested in using data visualizations for research as well as for communicating findings to different audiences. Waltman proposed to take the emerging field of AI Ethics as an empirical topic under a larger initiative about interdisciplinarity in AI. They were quickly joined by anthropologist

Holtrop, with science policy expert Ismael Rafols and information technology ethicist Jeroen van den Hoven advising. Philosopher of technology (residing in a department of Law and Technology in a Law School) Francien Dechesne and STS scholar Judit Varga joined later in the project.

The past ten years have seen growing concerns about the role of AI and algorithms in society more generally. Popular books (e.g. O'Neil, 2016; Eubanks, 2018) have raised concerns about the possible effects of autonomous systems. In response, governments, professional bodies and corporations have increasingly proliferated principles to guide a more ethical or trustworthy use of AI. These include 'fairness,' 'accountability,' 'transparency' and 'explainability' (Fjeld et al., 2020; Floridi and Cowls, 2019; Hagendorff, 2020; Jobin et al., 2019). These themes, however, have also long been addressed by fields like Computer Science (Dwork, 2012), Law (Pasquale, 2015), Geography (Kitchin, 2014), the humanities (Gitelman, 2013), Philosophy (Friedman and Nissenbaum, 1996, Van den Hoven, 1997), Media Studies (Boyd and Crawford, 2012), Science and Technology Studies (STS) (Bowker and Star, 1999), Information Science (Introna and Nissenbaum, 2000) and Anthropology (Wilf, 2013).

This loose constellation of research, which we will for convenience sake, refer to as AI Ethics could be seen as a branch of Ethics of Technology, which combines elements of applied ethics and the philosophy of technology (Hansson, 2017). At the same time, because a large part of the advancements in AI were made thanks to the availability of so called big data, AI Ethics might also be seen to emerge from discussions in Data Ethics (Floridi and Taddeo, 2016).[1] After some of the moral discussions around Big Data had become part of legal regulation, especially in Europe with the General Data Protection Regulation (text adopted in 2016), the discussion moved beyond *personal* data and collection towards how data were used, especially in Machine Learning, to build systems with a degree of autonomy (European Commission, 2020). AI Ethics also draws on Applied Ethics but also from moral theory (Bietti, 2020), political philosophy (Binns, 2018) and bioethics (Mittelstadt, 2019).

Because of the tremendous growth of AI research (and hype around AI technologies) all of these strands have been forcibly brought into dialogue, through policy discussions and more or less interdisciplinary conferences like ACM FAccT (Association for Computing Machinery Conference on Fairness Accountability and Transparency). In undertaking this research, we did not assume that these developments were sufficiently novel to warrant a new field, we only observed that there were attempts to consecrate such a field by the community through acts of naming. We should emphasize also that there is nothing particularly unique about AI Ethics in comparison to other emerging areas of research spurred by societal challenges and the disciplinary tensions this entails, except perhaps for the speed at which this has happened, due to the intense public interest in the topic and the rapid development and circulation of AI systems.

Interdisciplinarity as a problem

In one of the first group meetings, Moats tried to articulate the research problem for the project, in the sense of the real-world issue and/or research gap which the project would attempt to address. He suggested, in a draft Google Doc, that the problems related to AI might come down to asymmetries in interdisciplinary collaborations (who is included) and that this might be detected through scientometric analyses. But the rest of the team cautioned that this might not be the best framing. Although the name AI Ethics, in its conjuncture of Artificial Intelligence and Ethics implied the combination of two disciplines (and these two standing as a potential proxy for many more), it was suggested that AI Ethics might actually involve computer scientists considering the ethical implications of their work as opposed to working with a philosopher or social scientist or even consulting past work on ethics or moral philosophy.[2] It might also involve social scientists critiquing autonomous systems without any direct engagement with computer scientists. This made Moats realize that his implicit definition of interdisciplinarity had centred around *direct collaboration* (this being what STS scholars tend to observe ethnographically), but this was not the only possible understanding of how disciplines can mingle. So we agreed to leave what counts as interdisciplinarity as an open question.

Despite the fact that our working group came from different backgrounds and were answerable to different academic communities, we all seemed to agree that the case we were investigating (i.e. the field of AI Ethics) came first. That is, we agreed that our primary allegiance was to the empirical object of the study rather than particular theories or frameworks. This common ground seemed key to making our collaboration work. Yet we only realised this shared commitment *because* our different understandings of the term interdisciplinarity were brought into question. This is an example of how there is no necessary conflict between care in the sense of giving voice to all our relational positions and actively provoking disagreements, the disagreements in this case helped bring us together without also merging our positions into one. As a team, we translated the initial problem of how to *detect* interdisciplinarity into a problem of how to study the manifold forms that interdisciplinarity might take. Our question became: 'how could we characterise the status of interdisciplinarity in the field?'

Queries

Given our shared interest in data visualisations, we soon set out to study AI Ethics with Scientometric visualizations using data from the Microsoft Academic database (Wang and Schneider, 2020), a repository of scientific articles and metadata about them. Although, as we explained earlier, we will not go

into detail about our data analysis in this paper, it is worth noting that this data (which was already publicly available) was collected in accordance with Leiden University data policies and European Regulations. We used this data to make scientometric graphs and followed these up with qualitative analyses of the abstracts and articles themselves. These graphs were to be used to gain different overviews of the field, which could be refined through feedback with practitioners.

In mid-2020, we collected papers containing 'artificial intelligence' and 'ethic*' in their title or abstract in the period 2000–2019.[3] However, we were surprised to find only 1287 papers. Where was this massively hyped field? Was AI Ethics simply the wrong term (what about 'data ethics' 'politics of AI')?

One of us suggested a score-based query (e.g. Mahieu et al., 2018). Instead of picking some obvious name which signifies a domain, one would pick a series of terms which might capture it and assign scores to them depending on how relevant they might be to the field. Others however, felt that this would again, impose our own, 'arbitrary' assumptions on the proceedings.

Moats was sceptical for a slightly different reason, because of his anthropological or STS commitment to define communities how they define themselves. Waltman similarly wanted to engage with papers which 'self-identified' with the term AI Ethics but preferred to first group publications into 'research areas' (Waltman and van Eck, 2012) detected through their citation connections (visualized in the left image in Figure 1) and then sized according to the number of publications containing the terms 'artificial intelligence' and 'ethic*' (the right image in Figure 1). Thus, we would not be limiting the search to papers mentioning the term but also include publications connected through citations.

One of the clusters highlighted in this analysis seemed to contain papers from computer science, social sciences and humanities and included some of

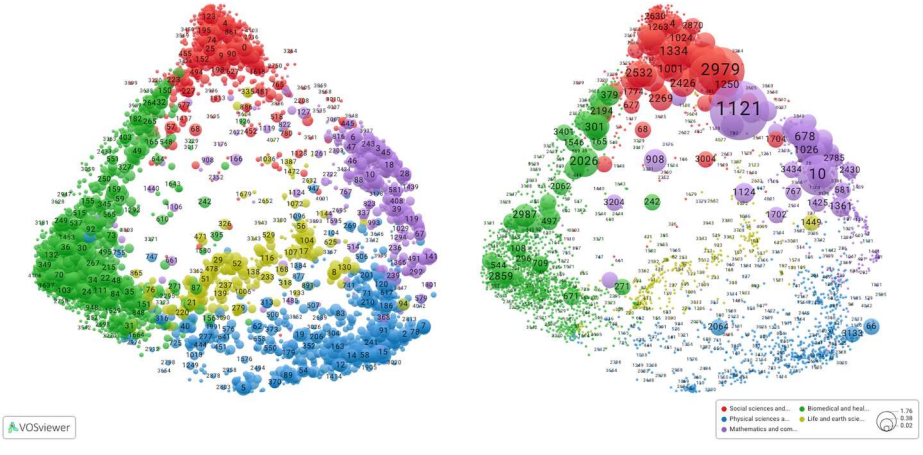

Figure 1. VOSviewer 'map of all science' using the Microsoft Academic database.[8]

the key papers cited earlier in our summary. However, analysing the papers in the cluster, qualitatively, Moats felt that the corpus (n = 1743) was characterized by only certain types of disciplinary interactions (citations across disciplines and broad statements of principles) but not others (e.g. direct collaborations). He wanted a query which would allow for more diverse interactions – if they existed that is. Moats, felt disconcertment here, torn between his methodological commitments, those of other team members and an uncooperative object of study. In the end we decided to use the score-based query after 'exhausting other options.' We consulted ethicist Van den Hoven to help us come up with a list of AI-related terms and a list of ethics-related terms (see Table 1).

We gave a score from 1 to 5 for each of the terms based on their presumed centrality to the field as we saw it. To select papers, we summed the scores of AI and Ethics terms in the title and abstract of each paper. There were 5455 papers which scored at least 5 from each list. We were pleased to find that these papers included more research about specific *applications* of AI, beyond reviews and commentaries.[4]

So even though we agreed philosophically about our commitments to the object of study, it turns out we did not initially agree on how to define its boundaries: how best to be faithful to the empirical, how to 'care' for the case? To move forward, we did not need to resolve our different preferred approaches but only pick a pragmatic solution which accommodated the contours of our empirical object. From our experiments and discussions, however, we realised that our different field delineations favoured certain people and disciplines at the expense of others: popular terms like 'AI Ethics' might privilege more self-promotional work, for example at the expense of more marginal terms. The score-based query, while requiring arbitrary choices, acknowledged the possible diversity and polyvocality of AI Ethics. So the problem of how to

Table 1. List of query terms and their scores.

AI Terms	Scores	Ethics Terms	Scores
artificial intelligence	5	ethics	5
machine learning	4	fair*	4
artificial neural network	4	harm	3
Bayesian OR baye*	4	risk	3
deep learning	4	moral	3
big data	3	discrimination	3
algorithm*	3	trust	3
data science	3	principle*	3
robot*	3	social	2
self-driving	3	accountability	2
gradient descent	3	interpretability	2
Turing Test	3	privacy	2
automation	2	equality	2
computer vision	2	bias	2
intelligent agent	2	oversight	2
supervised	2	responsibility	2
		safety	1
		security	1

map different forms of interdisciplinarity in the field then became a problem of 'how to respect different disciplinary understandings of the field?'

Conflicts and harmony

Although happy with the size and breadth of the corpus of publications, Moats was finding it difficult to make visualizations of the data. In a team meeting on zoom, he admitted that he had not been able to make a map that 'works.' Waltman then asked: 'what do you mean by "works"?' (Figure 2)

Moats was previously part of the field of Mapping Controversies (Venturini and Munk 2021), which uses digital tools to visualize debates on social media platforms like Twitter and Facebook, among other sources. He had been using tools from this community (Gephi, Table2Net), while Waltman and van Eck had been using VOSviewer (van Eck and Waltman, 2010, 2014), a programme they had developed, which also produces networks. These different combinations of software packages accomplish similar things but have very different visualization options and default settings.

In a follow-up meeting, Moats tried to explain that he felt the maps he was making were too homogeneous: he was looking for more distinct clusters which might speak to tensions between approaches in AI Ethics. Waltman remarked, jokingly, 'it seems like you are looking for controversy?' to which Holtrop remarked 'does that mean you [Waltman] are looking for harmony?'

This exchange, another moment of disconcertment, was important because it reminded Moats that he was assuming a confrontational form of sociality, coming from the fraught world of social media, which was not necessarily applicable to this different data source (academic papers) and particular

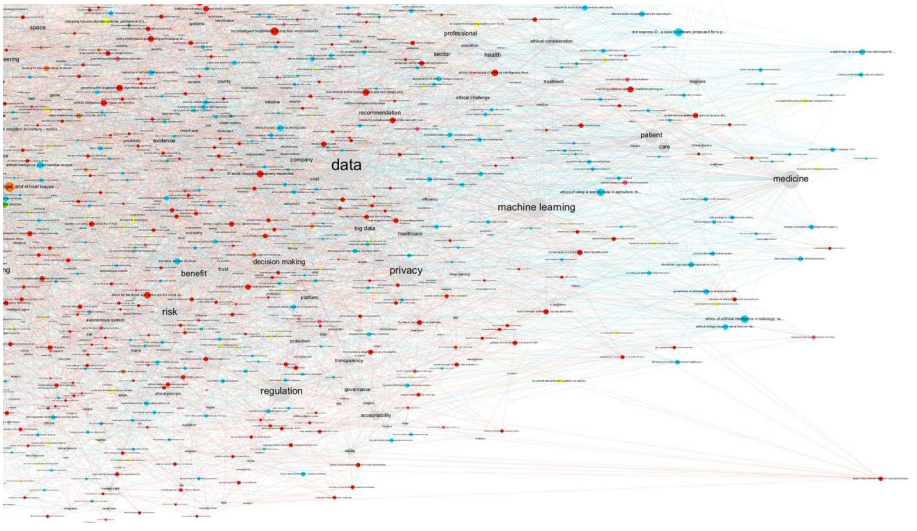

Figure 2. A bi-partite network of articles and keywords contained in abstracts.

scientific culture (AI Ethics). For Waltman, this moment highlighted that he and van Eck were interested in using visualizations for discovering higher-level patterns in one's data, whereas Moats was using visualizations as a way of making the underlying data more accessible for qualitative research (like a map of an unknown territory, Moats, 2021). Holtrop, for her part, could not distinguish between the different types of maps, visually, but the moment of disconcertment only happened because she was invited to a technical meeting which she would not normally have attended and was able to ask some awkward questions. Including non-technical disciplines in so-called technical work could be seen as part of caring for 'neglected things' but doing so also moved the project forward.

We could all take the mutual probing on this occasion because we were, by this point, comfortable with each other and could trust that our interactions would be handled with care: accommodating for different needs, abilities, time commitments and working conditions within the group.

In any case, this led to a discussion about the different perceived uses of visualizations. We agreed that we should use visualizations not to reveal some underlying truth of the matter (as some in Scientometrics or network science might; Jacomy, 2020), but as tools for fostering discussions between the practitioners we were interested in studying – as a kind of photo-elicitation method (Harper, 2002). In this case, the visualizations needed to be more legible than Moats' efforts. van Eck went away with the dataset to tinker with it. A few days later, he showed us the visualization presented in Figure 3.

van Eck explained that this was a network of the 1000 most relevant publications identified using the score-based query.[5] Two papers are connected to each other not only when one cites the other but also when they are both cited by the same papers or when they both cite the same papers: a mix of direct citation connections, co-citation connections, and bibliometric coupling connections (van Eck and Waltman, 2014).[6] Interestingly, while Moats' maps failed to capture possible differences within AI Ethics, van Eck's map elicited a narrative of relatively clear-cut disciplinary divisions. While it is not self-evident which depiction better represents the field, in our interviews with members of the AI Ethics community, many confirmed that discussions were frequently siloed within disciplines, even within collaborative projects and the supposedly interdisciplinary FAccT conferences.

From our struggles with the maps, we learned that visualizations might play an interesting role in interdisciplinary discussions. They are reasonably accessible, but different disciplines orient to them in different ways. They provoke questions like: 'what do we mean by "working?,"' 'What is the goal of a visualization?' and this made us wonder if such visualizations of the AI Ethics field could be used to purposely prompt *dialogue* between the different disciplines in that field? Could visualizations help care for interdisciplinary engagements between the social and computational sciences?

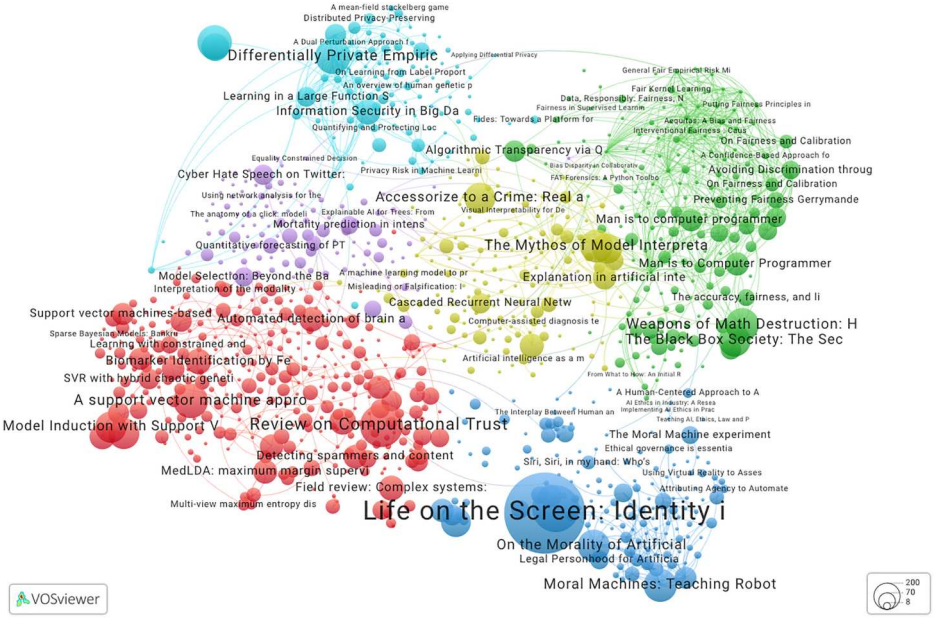

Figure 3. VOSviewer publication citation map based on the score-based query.[9]

Representations

Armed with variations on van Eck's visualization, we decided to organize some focus groups with AI Ethics practitioners (both in the academy and in industry). We would use the maps as a prompt to kick-start discussions about the composition of the AI Ethics field and also ask participants to tell us what might be missing from or skewed in the visualizations, to improve them in future. We gathered a long list of potential participants for interviews and online focus groups who had authored key papers represented in the maps, particularly from the FAccT conferences. Although AI Ethics researchers were in high demand at the time, and many never replied or politely declined, we managed to arrange two focus groups of 5–7 participants each centred around a specific topic: (1) criminal justice and (2) radiology. In each we attempted a geographic and demographic mix of participants, though there was a slight bias towards Europe given our connections. We also, importantly, attempted a mix of disciplines which included the following: computer scientists, philosophers of technology, sociologists, anthropologists and two topic experts – a criminologist and medical doctor. The host department, CWTS, which mainly deals with research about science, does not require ethical approval and we deemed the research to be low risk but even so, due to the sometimes heated nature of AI Ethics debates, we conducted the focus groups with informed consent, participants were given the option of anonymisation and Chatham House rules were followed.[7] The discussion was centred

around a series of visualizations, one after the other, with the moderator asking 'what participants recognized about the field from the visualizations' and 'what was missing.'

At first, we were quite disillusioned about the results of our conversations. Although Moats, as the moderator, had emphasized that the maps were in-progress and partial, we felt that the participants had focused too much on criticizing our visualizations for not representing the AI Ethics debate well. Some found *their* discipline missing, others missed references they felt were seminal to the debate.

> … in the map, or my personal mental map of the field, I would see another cluster, and that would be the whole discourse on sustainability. I mean, which, at least in the centre, where I work there is a very big research field in AI and sustainability, and then the whole post-humanism discourse is not here.. and justice.

But listening back to the recordings, we realised that these criticisms also raised important points about the politics of representations, which disciplines are favoured by certain data sets and certain search terms. In fact, the score-based query prompted discussions about the meaning of these key terms. One participant told the following story:

> I asked people how they differently defined certain terms, because these were terms that I had experienced as very overloaded. […] a social scientist said, we shouldn't define the terms because we have to leave them open. So everybody can feel that they can connect. And at that point, I thought, Okay, this is the point where the discussion ends.

To which another participant responded:

> I suppose I'm a little annoyed. The characterization that [anonymised] makes about what the social scientist is doing, not least because I'm a social scientist, but moreover, because I think […] that the social scientist or someone who is more focused on quote, unquote, critique, wanted to use the un-clarity as a moment of productivity […] and understanding where the unclarity is, is part of the process.

In this exchange, the first participant advocates starting with clear, agreed and shared definitions, while the other participant advocates accepting the ambiguity and multiplicity of definitions to allow for interdisciplinary conversations to come to a shared understanding of what needs to be visible and defined. For the second participant, understanding difference and discord is part of the process of translation. This awkward, yet productive moment showed us that even the problem of the different meaning of words is seen differently by different disciplines: people from different disciplines favour different ways of addressing it. Although we were hoping to provoke exactly these sorts of disagreements, we hope that participants felt comfortable disagreeing because (1) of their shared stake in the future of AI and (2) the convivial conversational space we attempted to create through our moderation and selection of participants.

Like our own experiences, this highlighted the importance of creating an atmosphere where differences can be engaged with care-fully.

Revisiting the focus group transcripts, we saw that despite being the object of ridicule, the visualizations had nonetheless prompted a somewhat productive discussion about disciplinary differences. This is at least what Dechesne, a philosopher of technology who was originally one of the focus group participants and later joined the team, saw it as. She had previously tried to run interdisciplinary workshops around reading articles together but the discussion quickly ground to a halt, stuck in the weeds.

So the problem for the project then became: how to use the properties of visualizations to facilitate interesting and productive conversations about interdisciplinarity in AI Ethics? What shapes whether visualisations can help interdisciplinary interactions? We realized that, although we had synthesized and made compromises between our different approaches, the visualizations we presented appeared as cohesive wholes to our audiences. Would the disjunctures be more productive if we made visualizations which make available these fissures in representation – in the meaning of words for example? Was it possible to facilitate a discussion of the maps in a way that multiple interpretations were encouraged and misalignments became opportunities for productive interdisciplinarity? These are the questions and the problems we are continuing to work with.

Discussion

These four moments of the project interestingly correspond to what are often key phases of the research process: articulating the problem, delimiting the object of study, designing methods, and presenting results to specific audiences. As we hope to have shown, each moment was prompted by an unforeseen disjuncture between ways of knowing. The first was initiated by different understandings of a term ('interdisciplinarity'). The second arose when we had to decide between different approaches for delineating the boundaries of the study. In the third, different methods and tools were decomposed and contrasted. In the fourth moment, our approach was subjected to the demands of different audiences. Importantly, whilst these disjunctures were between ways of knowing, they did not fall along obvious disciplinary lines or traditional divides: technical and humanistic, quantitative and qualitative. In some cases, all of us broadly agreed about certain methodological points while at other times we were divided *within* similar approaches, not to mention the fact that some of us found ourselves pulled between fields. We suggest that having this balance might be key to these disjunctures being productive rather than alienating: there must be enough common ground (commitment to the empirical, data literacy) to have a conversation but enough diversity in the team for profound questions about the nature of

research, and the object of study, to be thrown up. In this way, something about Latimer's 'being alongside' concept may be necessary here. If we are too different / at odds, then nothing interesting will happen but if we are too close and comfortable, disconcertment and thus novel reframings of the problem may be impossible.

In each moment, our disciplinary positions were tested and probed, and, in the process, our ambitions for the project also developed. This did not however happen in a linear way: we did not encounter a sequence of problems, solve them and move on, rather the problem (and questions) which drove the project were further developed. The initial problem – an ambiguity around the status of interdisciplinarity in AI ethics which needed to be clarified – called for descriptive analyses. By the end, the problem – the lack of productive confrontations between disciplines – demanded something more interventionist (provoking such confrontations). The point here is that problems, and by extension grand challenges are not fixed, they develop and are shaped by the interactions of disciplines in research projects.

So how do we provoke these encounters care-fully? Maniglier and other theorists cannot help with these more practical matters, which vary considerably not just based on the different types of knowledges in play but also the particular types of stalemates, misunderstandings and missed opportunities which characterise a certain constellation of disciplines. For example, social science and computer science disciplines may be held apart by different technical capabilities (coding skills) more than epistemological disagreements. We found that it was important to invite everyone to participate in every discussion even if it did not pertain to their disciplinary specialisms. This was both to foster a sense of shared purpose but also to provoke disagreements. The social scientists got involved in discussing the query design just as much as the computer scientists were invited to qualitatively analyse texts and interview transcripts. The key was not to allow a normal division of academic labour to ossify.

We also found that objects like data visualizations, might be one way to bridge gaps, because they allow a group with variable technical abilities to have a conversation around them. This is possible in AI Ethics at least because there is a certain amount of shared data-literacy in the community. Yet as we saw, visualizations do not inherently lead to productive dialogue. Within our group, the visualizations provoked deep reflections about the nature of research, while in the focus groups they mostly surfaced disagreements. Perhaps it was the fact that our team simply had more time for reflection and had more flexible identities which made the difference, or perhaps we had more stake in the visualizations working because we were using them as part of a research project. In any case more work is needed to understand the role of objects such as visualizations, which are simultaneously uniting and dividing, in interdisciplinary encounters.

Conclusion

In this paper we asked about the role that disciplinary tensions play in interdisciplinary collaborations: how they are negotiated in practice and how they relate to research problems, including grand challenges. These tensions are often seen as a barrier to collaboration and thus the addressing of complex grand challenges, but we have argued that they might also be thought of as an opportunity to generate new ideas, specifically new problem-formulations. This might be especially true in the case of grand challenges, which are often formulated in narrow, tech-solutionist or vague and indeterminate ways. While it is crucial, as the Special Issue editors point out, to interrogate the policy framings of societal challenges and the implications for interdisciplinary research, it is also important to recognise that problem-formulations are actively transformed as part of interdisciplinary research.

To demonstrate this point, we used the example of our own interdisciplinary struggles in a project attempting to map the field of AI Ethics using scientometric visualizations and qualitative methods. Building on work in STS, we employed a technique for analysing the process by collecting our reflections about transitions in the project which we found were characterized by feelings of disconcertment. Drawing on these, we described how the project developed from a more descriptive enterprise (to describe the AI Ethics field) to a more proactive intervention into the field (to provoke disciplinary tensions in AI Ethics) and that these transitions were prompted by several key moments in which differences between our disciplinary protocols and terminology came into sharp relief.

This account helps to flesh out some theoretical work by Maniglier (2021) which proposes that interdisciplinary work (indeed, thinking in general) might hinge on such moments of disjuncture between ways of knowing. We also reviewed STS literature on interdisciplinarity in practice, specifically involving care, provides us with many examples of how disciplinary tensions are practically navigated (Galison, 1997) and how they may be inevitable and productive aspects of caring relationships (Puig de la Bellacasa, 2011). What we want to contribute to this literature is the observation that these tensions might need to be proactively induced, not just tolerated or managed. Indeed, rather than there being a conflict or trade-off between care-ful management of relations between researchers and provoking tensions, productive disagreements might actually *contribute* to greater care in the long run by revealing not just what is different but also what is shared. In this sense, 'being alongside' may suggest too static and separate a relationship, mutual prods and transgressions might be necessary as well. These prods could be accomplished by purposefully including a mix of disciplines when undertaking tasks (like data sourcing or qualitative analysis) and using objects like data visualizations which can be viewed collectively yet experienced quite differently but would need to be

balanced with some measure of shared common ground to ensure productive good faith discussions.

So what does this imply for the AI Ethics community? It has long been acknowledged that AI and Machine Learning problem formulations, with different normative consequences, develop over the course of a project (Passi and Barocas, 2019). For example, a presumed problem of improving access to government services might be unwittingly transformed into the more narrow problem of maximising click-throughs on a government webpage. But these shifts often seem to happen without much awareness and are often un-reflected upon (Berendt, 2019). We argue for more attention to these shifts in problem articulations and specifically in how they might be proactively encouraged through the purposeful juxtaposition of different ways of knowing. We also suggest that our process of collective reflection, based around moments of disconcertment, might be useful in this space.

How can policy help facilitate these interdisciplinary disjunctures? The fact that in AI Ethics, different concepts (like fairness, accountability and transparency) have potentially different usages in relation to other concepts, is perhaps not an obstacle, because is precisely the situation Maniglier is calling for. Yet something about the current field, it seems, prevents more interesting discussions from happening. Policy tends to frame this problem as one which can be solved with a process of translation but, as we learned: to suppose that meanings must be settled in advance may benefit certain disciplines as opposed to others (those that are analytical or start from first principles as opposed to those which are more empirical or inductive). Indeed, settling these matters in advance may preclude some of these productive disjunctures from arising over the course of a collaboration.

Instead, perhaps policy and funding bodies can facilitate more open forums or more opportunities for meaningful discussion and recognition of these disciplinary differences, rather than sweeping them under the carpet or trying to wish them away. There are, for example, a growing number of proposals for workshops and exercises to surface and settle the specific sort of tensions between the disciplines involved in AI Ethics (Berendt, 2019; Mulligan et al., 2019) but we think that further work needs to be done to productively *dwell within* these tensions, rather than attempt to settle them – to purposely be alongside as much as being with. Policy should, of course, foster some shared goods (goals, needs, literacies) between researchers but perhaps they should not promote shared terminologies or methodologies. The work of exploring, nurturing and provoking disciplinary differences should be valued by funders and recognized as a contribution in and of itself. Funders should also understand that the grand challenges or problems they frame in interdisciplinary research calls may be actively transformed by the productive interactions of different disciplines – they might not end up addressing the same problem they started with.

How can we create situations in which the differences between disciplines are not sublimated or taken as given but actively interrogated? Maybe if there are more of these interdisciplinary moments, we will be able to create AI which is not only more ethical or 'fair' but also more inventive and boundary pushing – to not only ask if AI is ethical or not, but: 'what is AI for in the first place?' Addressing the grand societal challenge of AI might mean rethinking what kind of problem it is.

Notes

1. Note however that not all discussions in Data Ethics are necessarily AI Ethics issues, such as for example more general issues like data sovereignty.
2. Callard and Fitzgerald (2015) list many of these different arrangements.
3. We chose 'artificial intelligence' AND 'ethic*' because we found that publications rarely use the 'AI' or 'A.I.' abbreviations on their own. 'Ai', also, is a word in languages such as Italian.
4. However, we noticed that many documents hardly mentioned ethics or socio-political concerns at all. Often the term ethics referred to the more procedural 'ethics' involved in institutional review boards for human subjects research (e.g. when studying AI system users) but not in relation to the automated system itself.
5. According to the total link strength metric in VOSviewer.
6. By default, items in a VOSviewer visualization are grouped into clusters using a clustering algorithm, and the colour of an item shows the cluster to which the item has been assigned. For example, in Figure 3, publications are grouped into six clusters, each depicted by a colour.
7. Participants can repeat what was said but not who said it.
8. The map shows 4134 clusters of publications. The size of a cluster reflects the total number of publications in the cluster (left) or the number of publications in the cluster that contain the terms 'artificial intelligence' and 'ethic*' in their title or abstract (right). Each cluster belongs to one of the five macro clusters, as indicated by the colour of a cluster. An interactive version of this map is available online at https://tinyurl.com/y6clmebo.
9. The size of a publication reflects the number of citations received by the publication. An interactive version of this map is available online at https://tinyurl.com/y4x9e973.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Nederlandse Organisatie voor Wetenschappelijk Onderzoek. This work was completed as part of the project REIMAGINE ADM (https://chanse.org/reimagine-adm/) supported by the CHANSE ERA-NET co-fund program, which has received funding from the European Union's Horizon 2020 Research and Innovation Program, under grant agreement no. 101004509.

Notes on contributors

David Moats is a lecturer in Digital Methods at the Department of Digital Humanities, King's College London. Specializing in Science and Technology Studies (STS), his research

is mainly about digitization and the role of machine learning and artificial intelligence in transforming various industries, including media, healthcare, politics and academia. He is also interested in the methodological implications of these new sources of digital data for the humanities and social sciences. He regularly works with computer scientists and designers to develop new research tools and has written about the role of interdisciplinarity in digital research. He previously held posts at the University of Helsinki and Linköping University.

Tjitske Holtrop is a senior researcher and director of culture and inclusion at the Centre for Science and Technology Studies (CWTS) at Leiden University. Tjitske is an anthropologist and obtained her PhD degree in 2017 from the University of Amsterdam with a study on the complex coming-into-being of the evaluation of the Dutch civil-military mission in Afghanistan. She is currently involved in projects on recognition and reward, inter- and transdisciplinarity, responsible research and innovation, and self-assessment.

Nees Jan van Eck is a senior researcher and head of data science at the Centre for Science and Technology Studies (CWTS) at Leiden University in the Netherlands, where he leads data infrastructure projects and coordinates the Information & Openness focal area. His research focuses on studying and developing infrastructures, algorithms, and tools to support research assessment, science policy, and scholarly communication.

Judit Varga is a postdoctoral researcher at the Centre for Science and Technology Studies (CWTS) at Leiden University in the Netherlands. She focuses on exploring values in science using mixed methods. Her doctoral research explored how STS and scientometrics can be combined to study the diversification of social scientific knowledge produced with computational methods, through a case study about geotagged social media research. Judit is interested in how ways of knowing converge, diverge and get coordinated in science, both as an object of study and in her own scientific practice.

Ludo Waltman is a scientific director and professor of Quantitative Science Studies at the Centre for Science and Technology Studies (CWTS) at Leiden University. He is co-chair of the Research on Research Institute (RoRI). Ludo's work focuses on studying and developing infrastructures, algorithms, and tools to support research assessment, science policy, and scholarly communication. Together with his colleague Nees Jan van Eck, Ludo has developed the well-known VOSviewer software for bibliometric visualization.

Francien Dechesne is an sAssociate Professor of Ethics and Digital Technologies at the Center for Law and Digital Technologies (eLaw) of the Leiden Law School. With a background in computer science, mathematics and philosophy, her research as well as her responsibilities in the education center around interdisciplinary perspectives on Artificial Intelligence, and its interactions with society: demystification of the technology, identifying and explaining societal impacts (to those affected as well as those involved in the development), and investigating different strategies towards responsible AI.

ORCID

David Moats http://orcid.org/0000-0001-9622-9915
Tjitske Holtrop http://orcid.org/0000-0002-1413-8591
Nees Jan van Eck http://orcid.org/0000-0001-8448-4521
Judit Varga http://orcid.org/0000-0003-3988-8820
Ludo Waltman http://orcid.org/0000-0001-8249-1752
Francien Dechesne http://orcid.org/0000-0002-3511-9103

References

Association of Nordic Engineers. (2020). Addressing ethical dilemmas in AI: Listening to engineers. Available at: https://standards.ieee.org/initiatives/autonomous-intelligence-systems/ethical-dilemmas-ai-report/ (accessed 15 May 2023).

Berendt, B. (2019) AI for the common good?! Pitfalls, challenges, and ethics pen-testing, *Paladyn, Journal of Behavioral Robotics*, 10(1), pp. 44–65. DOI: 10.1515/pjbr-2019-0004

Bietti, E. (2020) From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 210–219 (Barcelona: ACM)).

Binns, R. (2018) Fairness in machine learning: Lessons from political philosophy, in: *Conference on Fairness, Accountability and Transparency*, pp. 149–159 (New York: PMLR)).

Blok, A., Carlsen, H. B., Jørgensen, T. B., Madsen, M. M., Ralund, S. and Pedersen, M. (2017) Stitching together the heterogeneous party: A complementary social data science experiment. *Big Data & Society*, 4(2), pp. 1–15.

Bowker, G. and Star, S. L. (1999) *Sorting Things Out* (Cambridge, MA: MIT Press)).

Boyd, D. and Crawford, K. (2012) Critical questions for big data, *Information, Communication & Society*, 15(5), pp. 662–679. DOI: 10.1080/1369118X.2012.678878

Callard, F. and Fitzgerald, D. (2015) *Rethinking Interdisciplinarity Across the Social Sciences and Neurosciences* (Houndmills, UK: Palgrave Macmillan).

Callon, M. (1980) Struggles and negotiations to define what is problematic and what is not: The socio-logic of translation, in: K.D. Knorr, R. Krohn, and R Whitley (Eds) *The Social Process of Scientific Investigation*, pp. 197–219 (Dordrecht, Netherlands: D Reidel Publishing Company).

Coeckelbergh, M. (2020) *AI Ethics* (Cambridge, MA: The MIT Press)).

Collins, H., Evans, R. and Gorman, M. (2019) Trading zones revisited, in: D. Caudill, S. Conley, M. Gorman, and M. Weinel (Eds) *The Third Wave in Science and Technology Studies: Future Research Directions on Expertise and Experience*, pp. 275–281 (Houndmills, UK: Palgrave Macmillan)).

Davies, S. R. and Horst, M. (2015) Crafting the group: Care in research management, *Social Studies of Science*, 45(3), pp. 371–393.

Diedrich, A., Upham, P., Levidow, L. and van den Hove, S. (2011) Framing environmental sustainability challenges for research and innovation in European policy agendas, *Environmental Science & Policy*, 14(8), pp. 935–939.

Dwork, C., Hardt, M., Pitassi, T., et al. (2012) Fairness through awareness, in: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226 (New York, NY: ACM). DOI: 10.1145/2090236.2090255

European Commission. 2020. AI watch: Artificial Intelligence in public services. *European Commission.* Available at: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC120399/jrc120399_misuraca-ai-watch_public-services_30062020_def.pdf (accessed 15 May 2023).

Eubanks, V. (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York, NY: St. Martin's Press)).

Fitzgerald, D. and Callard, F. (2015) Social science and neuroscience beyond interdisciplinarity: Experimental entanglements, *Theory, Culture & Society*, 32(1), pp. 3–32.

Fjeld, J., Achten, N., Hilligoss, H., et al. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* Available at: http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420 (accessed 15 May 2023).

Floridi, L. and Cowls, J. (2019) A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L. and Taddeo, M. (2016) What is data ethics? *Philosophical Transactions of the Royal Society*, 374(2083), pp. 1–5. DOI:10.1098/rsta.2016.0360

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

Galison, P. (1997) *Image and Logic: A Material Culture of Microphysics* (Chicago: University of Chicago Press)).

Gardam, T. (2019). Data science and the case for ethical responsibility. *Ada Lovelace Institute*. Available at: https://www.adalovelaceinstitute.org/blog/the-culture-of-computing-and-the-case-for-ethical-responsibility/ (accessed 15 May 2023).

Gilligan, C. (1993) *In a Different Voice: Psychological Theory and Women's Development* (Cambridge, MA: Harvard University Press)).

Groot, B. C., Vink, M., Haveman, A., et al. (2019) Ethics of care in participatory health research: mutual responsibility in collaboration with co-researchers, *Educational Action Research*, 27(2), pp. 286–302.

Gitelman, L. (2013) *Raw Data Is an Oxymoron* (Cambridge, MA: MIT Press)).

Hagendorff, T. (2020) The ethics of AI ethics: An evaluation of guidelines, *Minds and Machines*, 30, pp. 1–22.

Hansson, S. O. (2017) *Introduction*, in: S. O. Hansson (Eds) *The Ethics of Technology: Methods and Approaches*, pp. 1–14 (London: Rowman & Littlefield).

Harper, D. (2002) Talking about pictures: A case for photo elicitation, *Visual Studies*, 17(1), pp. 13–26.

HM Government (UK). 2017. Industrial strategy: Building a Britain fit for the future. *Gov.uk*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/664563/industrial-strategy-white-paper-web-ready-version.pdf (accessed 15 May 2023.

Institute of Electrical and Electronics Engineers. 2020. Ethically aligned design. *IEEE.org*. Available at https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf (accessed 15 May 2023).

Introna, L. and Nissenbaum, H. (2000) Shaping the Web: Why the politics of search engines matters, *The Information Society*, 16(3), pp. 169–185.

Jacomy, M. (2020) Epistemic clashes in network science: Mapping the tensions between idiographic and nomothetic subcultures, *Big Data & Society*, 7(2), pp. 1–21.

Jensenius, A. R. (2012) Disciplinarities: intra, cross, multi, inter, trans. In: *ARJ*. Available at: https://www.arj.no/2012/03/12/disciplinarities-2/ (accessed 7 November 2025).

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

Kaltenbrunner, W. (2014) Decomposition as practice and process: Creating boundary objects in computational humanities, *Interdisciplinary Science Reviews*, 39(2), pp. 143–161.

Kitchin, R. (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (London: Sage)).

Klein, J. T. (2004) Prospects for transdisciplinarity, *Futures*, 36, pp. 515–526.

Latimer, J. (2013) Being alongside: Rethinking relations amongst different kinds, *Theory, Culture & Society*, 30(7–8), pp. 77–104.

Latimer, J. and López Gómez, D. (2019) Intimate entanglements : Affects, more-than-human intimacies and the politics of relations in science and technology, *The Sociological Review*, 67(2), pp. 247–263.

Law, J. and Lin, W. (2010) Cultivating disconcertment, *The Sociological Review*, 58(2_suppl), pp. 135–153.

Mahieu, R., van Eck, N. J., Van Putten, D. and Van den Hoven, J. (2018) From dignity to security protocols: A scientometric analysis of digital ethics, *Ethics and Information Technology*, 20(3), pp. 175–187. DOI:10.1007/s10676-018-9457-5

Maniglier, P. (2021) Problem and structure: Bachelard, Deleuze and transdisciplinarity, *Theory, Culture & Society*, 38(2), pp. 25–45.

Mittelstadt, B. (2019) Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1(11). Nature Publishing Group UK London, pp. 501–507.

Moats, D. (2021) Rethinking the 'Great Divide': Approaching interdisciplinary collaborations around digital data with humour and irony. *Science & Technology Studies* 34(1), pp. 19–42.

Mol, A., Moser, I. and Pols, J. (eds) (2010) *Care in Practice: On Tinkering in Clinics, Homes and Farms* (Bielefeld, Germany: transcript Verlag)).

Mulligan, D. K., Kroll, J. A., Kohli, N., et al. (2019) This thing called fairness: Disciplinary confusion realizing a value in technology, in: A. Lampinen, D. Gergle, and D. A. Shamma (Eds.), *Proceedings of the ACM on Human-Computer Interaction 3(CSCW)*, pp. 1–36 (New York, NY: ACM)).

NAIRR. 2023. Strengthening and democratizing the U.S. Artificial intelligence innovation ecosystem: An implementation plan for a national artificial intelligence research resource. *AI.Gov*. Available at: https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf (accessed 15 May 2023).

Neff, G., Tanweer, A., Fiore-Gartland, B. Osburn, L. (2017) Critique and contribute: A practice-based framework for improving critical data studies and data science. *Big Data*, 5(2), pp. 85–97.

NITRD. 2019. 2016-2019 Progress Report: Advancing Artificial Intelligence R&D. NITRD.gov. Available at: https://www.nitrd.gov/pubs/AI-Research-and-Development-Progress-Report-2016-2019.pdf (accessed 15 May 2023).

Nowotny, H., Scott, P. and Gibbons, M. T. (2001) *Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty* (Hoboken, NJ: Wiley).

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

Osborne, P. (2015) Problematizing disciplinarity, transdisciplinary problematics, *Theory, Culture & Society*, 32(5–6), pp. 3–35.

Pasquale, F. (2015) *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, MA: Harvard University Press)).

Passi, S. and Barocas, S. (2019) Problem formulation and fairness, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 39–48 (New York: ACM).

Puig de la Bellacasa, M. (2011) Matters of care in technoscience: Assembling neglected things, *Social Studies of Science*, 41(1), pp. 85–106.

Puig de la Bellacasa, M. (2012) Nothing comes without its world': thinking with care, *The Sociological Review*, 60(2), pp. 197–216. DOI:10.1111/j.1467-954X.2012.02070.x

Strathern, M. (2004) *Commons and Borderlands: Working Papers on Interdisciplinarity, Accountability and the Flow of Knowledge*, (Wantage, Oxon, UK: Sean Kingston Publishing).

van den Hoven, M. J. (1997) Privacy and the varieties of moral wrong-doing in the information age, *Computers & Society*, 27(3), pp. 33–37.

Stember, M. (1991) Advancing the social sciences through the interdisciplinary enterprise, *The Social Science Journal*, 28(1), pp. 1–14. DOI:10.1016/0362-3319(91)90040-B

Tronto, J. C. (1987) Beyond gender difference to a theory of care, *Signs: Journal of Women in Culture and Society*, 12(4), pp. 644–663.

van Eck, N. J. and Waltman, L. (2010) Software survey: VOSviewer, a computer program for bibliometric mapping, *Scientometrics*, 84(2), pp. 523–538. DOI:10.1007/s11192-009-0146-3

van Eck, N. J. and Waltman, L. (2014) Visualizing bibliometric networks, in: Y. Ding, R. Rousseau, and D. Wolfram (Eds) *Measuring Scholarly Impact: Methods and Practice*, pp. 285–320 (Berlin: Springer). DOI:10.1007/978-3-319-10377-8_13

Venturini, T and Munk, A. K. (2021) *Controversy Mapping: A Field Guide*. (Hoboken, NJ: John Wiley & Sons).

Verran, H. (2001) *Science and an African Logic* (Chicago: University of Chicago Press)).

Waltman, L. and van Eck, N. J. (2012) A new method for constructing a publication-level classification system of science, *Journal of the American Society for Information Science and Technology*, 63(12), pp. 2378–2392.

Wang, Q. and Schneider, J. W. (2020) Consistency and validity of interdisciplinarity measures, *Quantitative Science Studies*, 1(1), pp. 239–263.

Wang, K., Shen, Z., Huang, C., et al. (2020) Microsoft academic graph: When experts are not enough, *Quantitative Science Studies*, 1(1), pp. 396–413.

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K. and Cave, S. 2019. Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research. *Nufield Foundation*. Available at https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf (accessed 15 May 2013).

Washington, A. L. (2018) How to argue with an algorithm: Lessons from the compas-propublica debate, *Colorado Technology Law Journal*, 17(1), pp. 131–160.

West, S. and Schill, C. (2022) Negotiating the ethical-political dimensions of research methods: A key competency in mixed methods, inter – and transdisciplinary, and co-production research, *Humanities & Social Science Communications*, 9(294), pp. 1–13. DOI:10.1057/s41599-022-01297-z

Weszkalnys, G. and Barry, A. (2013) Multiple environments: Accountability, integration and ontology, in: A. Barry, and G. Born (Eds) *Interdisciplinarity: Reconfigurations of the Social and Natural Sciences*, pp. 178–208 (London: Routledge)).

Wilf, E. (2013) Toward an anthropology of computer-mediated, algorithmic forms of sociality, *Current Anthropology*, 54(6), pp. 716–739.

Zuiderent-Jerak, T. (2021) STS as a third space between evidence-based medicine and the human sciences, in: L. Downey, and T. Zuiderent-Jerak (Eds) *Making & Doing: Activating STS Through Knowledge Expression and Travel*, pp. 197–218 (Cambridge, MA: MIT Press).