Outlook

Scientific Data, Artificial Intelligence, and People

My aim in this book has been to examine how scientific data not only represent information but are also implicated in social accountabilities and social action. I argued that, more than ever before, data have become a peculiarly epistemic and social stuff that scientists make, use, and circulate to learn about nature, instruct students, build connections, transmit knowledge, grow trust, and evaluate the work of others. I have considered scientists, students, and technicians as people: embodied, cognizant, social, and cultural human beings. By attending to participants' accounting practices and their resources I focused on matters that are elementary and common and took them very seriously (cf. Munger 2023). I have aimed to stay true to Peter Winch's (1958, 84) observation that "to understand the activities of an individual scientific investigator we must take account of two sets of relations: first, his relation to the phenomena which he investigates; second, his relation to his fellow-scientists." I have refrained from separating "the epistemic" and "the social" and argued that proceeding this way promises insights that are both foundational and transferable beyond the settings that I have examined. But how can this approach contribute to make sense of transformations in data-rich science due to machine learning (ML) and generative artificial intelligence (AI)? In the following, I make the case for the benefits of considering this question at the refined granularity of social interaction, consider how scientists themselves are social inquirers who do ethnography, and attend to emergent socialities and their ethics.

Speculations about, and expectations of, the impact of ML and generative AI in the sciences reach from novel possibilities of discovery and changes to the division of labor, to concerns over the trustworthiness of results, the unsettling nature of AI systems as "black boxes," humans' potential marginalization in creative work, illusions of understanding, and problems of responsible uses, as well as concerns over the credit and authorship of data and results. Although ML and generative AI are related

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 04:02:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.011

conceptually (such as in automatizing probabilistic inference), many scientists regard them as distinct technologies. Diverse forms of ML have been part of scientific work for more than a decade and include widely accepted methods like artificial neural networks and Gaussian processes (Ting 2025). By contrast, uses of generative AI, conceived as Large Language Models (LLMs) and related technologies, have attracted scientists' attention since the publication of the transformer algorithm in 2017 and its uses in chatbots. Many forms of ML and AI have humans "in the loop" to improve their performance, such as in supervised learning or reinforcement learning (Ting 2025). These uses add a contemporary example to *How Data Need People*. But this is not what I address in the following.

Continuing my naturalistic stance, I propose to consider researchers as people – embodied, social human beings – who use ML and generative AI in data-rich science. Some authors focus on ML and AI applications in the sciences without explicitly addressing the sociality of their participants (Wang et al. 2023; Hogg and Villar 2024), others attend to researchers as individuals that interact with machines (Krenn et al. 2022; Collins et al. 2024),[1] and still others consider communal aspects in abstract terms (Messeri and Crockett 2024).[2] From the perspective that I adopt in this book, doing science is unavoidably both an individual and a collective endeavor. Much of what is interesting and consequential happens between the individual and the collective and comes into focus only when we attend to situated interactions. Doing so may also be our best way to identify backgrounded assumptions and to avoid abstractions that may easily turn out to be misplaced. Furthermore, a naturalistic ethnographic approach attends not to fixed relations and situations, but to members' methods that are by necessity open to novel and likely unexpected developments that lie ahead. Communities may be the sites where social norms are enacted, but their members are social actors – users of methods who live lives and thus have biographies.

Like anyone in society, scientists use accounting practices in their social and professional lives. In this book two distinct notions of accountability have been in play. Both are described in the Introduction. One is ethnomethodological accountability. Situated in lived practice, it is reflexive

[1] By invoking a sort of Turing test to examine the scientific understanding of AI systems, Krenn et al. (2022) go beyond individualist assumptions.

[2] For example, Messeri and Crockett (2024, 50) ponder illusions of understanding, due to uses of AI, of "communities of knowledge," defined as "groups of individuals with distributed knowledge and understanding that allow individual community members to benefit from expertise held by others."

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 04:02:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.011

in the sense that "the describing of social activities is part and parcel of the activities so described" (Sharrock and Anderson 1986, 57). The point is to "study the ways in which [people] organize themselves so that they can tell us about the things they do" (Sharrock and Anderson 1986, 57). In Chapters 2 to 8, I have traced ethnomethodological accountability through an observatory's control room, graduate student training, uses of diagrams and mundane reason in data analyses and interpretations, organizing collaborative work, achieving proper membership in "open science," and encoding the team's knowledge of its data in the dataset itself. These episodes show how epistemic and social orders are entangled in scientific work with large datasets. As scientists seek to agree on uses of ML and generative AI, they are unavoidably going to use accounting practices, but what these are, and which resources they involve, are questions for empirical study. Scientists will remain evaluative, accounting practices will shape the achievement of membership, diagrammatic practices and mundane reasoning are bound to remain fundamentally important, and normative orders will emerge.

The other accountability, of second order, that this book has considered is the "tyranny of accountability," the "ever-present possibility of being noticed, praised, blamed, questioned, called out, and judged" (Enfield and Sidnell 2022, 21), a reflective stance that requires language to become socially salient. Its relevance in view of the emergence of ML and generative AI may be recognizable more readily than ethnomethodological accountability. In science, peer review is a central "system of accountability" (Douglas 1980, 35) where authorship and its balancing of credit and responsibility come to the fore. This is where communal standards and norms, legitimate practices and resources, and issues of membership are adopted and sanctioned. Its changes deserve close attention. Journal articles are units of the scientific literature and, as such, also units of accountability. As Chapters 3, 4, 5, and 8 show, the accountability of authorship is reflexive throughout on the research process, including data analyses, which leads us back to the accounting practices that ethnographic studies can identify and describe.

During my study I found that it was not just I, the visiting anthropologist, who worked ethnographically, but that scientists and technicians, too, were social inquirers who did a sort of ethnography themselves, albeit one that is distinct from the anthropologist's (cf. Chapters 2, 3, 7, and 8). It is no surprise that rapid technological and conceptual changes prompt people to inquire into how fellow members adopt them, ponder the limits of acceptable or agreeable uses, and shape new norms. Anthropologists have

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 04:02:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.011

documented how swiftly uses of new media lead to the adoption of new norms (Miller and Horst 2012). Astronomers' reflections on the impact of ML and generative AI have many sites, from lunch conversations and seminar discussions to conferences and hack weeks. Some astronomers have probed into their community's practices, composed reports of their findings, and posted them on the arXiv preprint server, making them widely accessible.

Considering ML, Daniela Huppenkothen and her coauthors (2023) assembled a team of astronomers at a hack week to formulate a "primer to the astronomical community, including authors, reviewers, and editors, on how to implement machine learning models and report their results in a way that ensures the accuracy of the results, reproducibility of the findings, and usefulness of the method." Examining uses of LLMs, Morgan Fouesneau and his coauthors (2024) "conducted a study involving 13 astronomers at different career stages and research fields to explore LLM applications across diverse tasks over several months and to evaluate their performance in research-related activities." This was followed up by a survey and led to the formulation of recommendations to ensure that "these tools serve as aids rather than substitutes for rigorous scientific inquiry." In proposing best practices and formulating recommendations, both studies have a normative bent.[3]

These "members' studies" are not ethnographic, but one may consider them as one part of a spectrum of scientists' social inquiries, of which a collaboration's "ethnographic" exploration of how to act properly in the uncommon domain of "open science" is another (Chapter 7). The latter translated elements of a shared lifeworld (uses of credit cards, contracts, lawsuits, and other means of sanctioning) into the domain of open data access. Exploring a novel domain and describing it with a certain vocabulary may be consequential in that the terms used to describe it may end up defining its practices. Given how easily and swiftly generative AI, in the form of chatbots, is integrated into diverse scientific workflows and is designed to be interactive, an interactional ethnographic perspective appears almost inevitable. But an ethnographic approach is also called for to discover ethical issues that can remain hidden from analysts otherwise.

My notion of data-centrism – that making, using, and publishing data causes humans to interact and engage in specifiable ways – is inspired by Georg Simmel's explorations of the diversity of social forms in the early

[3] Related earlier studies include surveys of astronomers' use of software (like Momcheva and Tollerud 2015) that informed the work of the Astropy Collaboration (2013, 2018, 2022).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 04:02:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.011

twentieth century. As I wrote in the Introduction, Simmel realized when writing his *Sociology* (1992 [1908]) that many forms of sociation awaited discovery – not least by means of more detailed, "microscopic" studies –, but also because new forms keep coming-to-be while others pass away. And so it will be in the twenty-first century. The large-scale impacts of ML and generative AI will unavoidably be tied to microscopic, granular interactions, whose study, I hope, will continue to be a reminder of human agency and responsibility.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 04:02:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.011