Opacity and reproducibility in data processing

Reflections on the dependence of AI on the data ecosystem

Sabina Leonelli

1. Introduction

It is sometimes argued that AI tools, though strongly dependent on the availability of large volumes of training data for their accuracy and effectiveness, are becoming increasingly less constrained by the scope and biases of the data themselves – both because the quantity and variety of data used to train algorithms grows at vertiginous speed, and because AI gets exponentially better at correcting bias and calibrating results towards specific, accurate solutions. Without wishing to deny such advancements and the resulting increase in potential for these technologies, I here maintain that AI is still strongly tied to the quality and representativeness of training data and that existing data gaps are not credibly filled by data produced for that very purpose, given that such production is strongly informed by expectations around the outputs and the focus on algorithmic outputs is taking attention away from the decision-making happening at various stages of data elaboration. Indeed, simulated, augmented, or synthetic data, which are supposedly 'artificial' insofar as they are created by humans for training algorithms and are not meant to faithfully document a specific aspect of the world, are produced and processed through specific assumptions about what the world may be like or what characteristics of the world one may be interested in. Whether or not these assumptions are explicitly identified and debated, they play an important role in framing the ways in which algorithms are developed to mine, model and visualize data, and thus directly affect the goals, methods and tools of AI. In what follows, I reflect on these concerns and on their implications for how we may understand the notion of opacity, so often identified as a major concern in the use

https://doi.org/10.14361/9783839467664-017 https://www.inlibra.com/de/agb - Open Access -

of AI for research purposes, and its relation to the reproducibility of research, that is the idea that it is possible to ascertain the credibility of specific outputs through success in re-creating them, which in turn involves some understanding of how they were produced in the first place.

2. Investigating research data journeys

My research concerns knowledge production through AI, particularly in the biological, biomedical and environmental domains. In that context I am interested in the extent to which insights derived from existing knowledge and research shape AI-powered data analytics and how/if such analytics are themselves capable of producing novel insights. As a window towards that problem, I have investigated not just what data collections exist – what people can actually source as input for their analysis – but also *how data are mobilized* once they have been generated and/or collected, garnered into digital infrastructures, and eventually re-used. I have traced and theorized such processes as "data journeys" (Leonelli/Tempini 2020), with a particular interest in data sets that get repurposed several times by people with different expertises. One example is data collected from social media (tweets, comments, 'likes') being reused to track public health concerns – as for instance happened during the COVID-19 pandemic – as well as mobility trends, such as how often people use public transport following periods of lock-down (e.g. Leonelli et al. 2021; Leonelli 2021). Another example is data acquired from detailed satellite imaging of specific territories, which are used to study phenomena as wide-ranging as deforestation trends, farming habits, urban planning and migration patterns, depending on how the images are processed and what other datasets they are combined with (Leonelli/Williamson 2023). Such situations are prime instances of what AI tools are supposed to achieve: That is, to enable researchers to recombine and reanalyse existing datasets for a variety of purposes, thereby extracting maximum value from the data as evidence for knowledge claims and related interventions.

The major challenge in tracking data journeys has been thinking about what happens when you have a very large, heterogeneous set of data and people need to rely on that dataset to do certain kinds of work, but at the same time have to make decisions about what part of that data they can trust.

https://doi.org/10.14361/9783839467664-017 https://www.inlibra.com/de/agb - Open Access -

How should/can the reliability of data and the quality of the information that is to be extracted from it be assessed? Who do you collaborate with when you're trying to do this kind of work, and how do you make such decisions? How is expertise distributed across data journeys, including the employment of data within AI, and which of the experts involved are accountable for the overarching outcomes of that complex system? The moment we are plunged into a large data ecosystem, we are often looking at thousands of people who have been working on that ecosystem and changing it to fit their aspirations, assumptions and goals. How to trust such a distributed system – does it mean verifying whether each individual contributor has done a good job, and if so, how can this be done? Are there ways to verify the quality and reliability of data ecosystems beyond the reconstruction of individual contributions, and if so, what are they?

I have explored these questions in collaboration with Niccolò Tempini and several collaborators from the natural sciences through DATA_SCIENCE ("The Epistemology of Data-Intensive Science"), a project sponsored by the European Research Council which ran from 2014 to 2019 and focused on the epistemology of data science and its applications in biology and biomedicine. We attempted to follow some datasets from the moment they were created to the moment they were organized into data infrastructures and further reused in a variety of projects. In an approach closely aligned with the infrastructural inversion pioneered by Geoffrey Bowker and Susan Leigh Star (1999), the starting point typically was data infrastructures, because this was a moment in the history of data when we began witnessing different perspectives on the conditions under which data could be used – intelligibly and actionably. From there, the next step was to find out where data were originally sourced and investigate how they were deployed and interpreted by database users. This was a difficult enterprise because you cannot tag data – it has been tried and found to be too difficult to implement. It is a form of detective work to try and track what happens to particular data sets, how they get modified and reshaped to fit different purposes and what the consequences for knowledge production are, particularly in cases where there are some very substantive disagreements between people who produce or collect data in the first place, and people who end up reusing them in a different environment and giving them a completely different meaning and frame of reference, which is where we saw many of these kinds of conflicts.

https://doi.org/10.14361/9783839467664-017 https://www.inlibra.com/de/agb - Open Access -

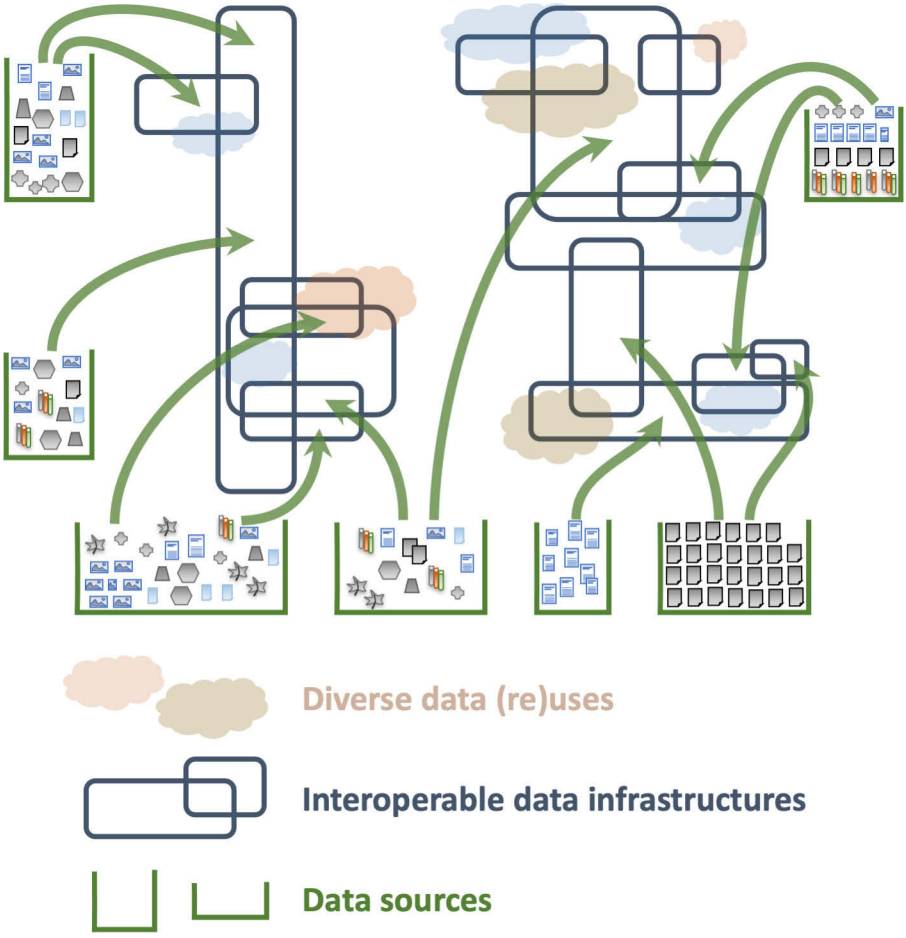
3. In-practice opacity within data ecosystems

Here is one potential representation of the research landscape viewed from the perspective of data movements and reanalysis (see fig. 1). The blue boxes in the middle of this figure are various databases. Sometimes they overlap, sometimes not: They are haphazardly overlapping. They tend to be funded in different ways and for different purposes by different institutions. They have different objectives. They have different lifespans and different types of data intersect with these data infrastructures, which different audiences use in different ways. A noteworthy aspect when considering data ecosystems as a serendipitous, organically growing ensemble is the fact that people who end up using data very often not only do not have a clue how data were processed or what the underlying structure of the organizations that are caring for, maintaining and stewarding the data, are. Even in the rare cases when there is a way to track data processing within a given database, with detailed information about where data comes from and how they have been manipulated, it would take too long to understand this narrative and its implications for one's work. Thus, effectively these systems become black boxes. This is not in-principle opacity of the kind sometimes encountered in AI tools, where we simply do not know – and cannot explain – how machines are generating a given output. This is in-practice opacity, emerging from pragmatic issues of tractability and intelligibility of large data structures. Even in a situation where there are enough metadata and contextual information that you could try and reconstruct the whole history of the data, thereby better understanding what decisions have shaped its processing and why, such an enterprise becomes undoable for lack of time.

All the cases we examined kept showing us that the bigger the exercise in data linkage and reuse, the bigger the effort to calibrate, process,reprocess and reanalyze the data that went into the system, in the attempt to make sure that the results were reliable. There is a constant and growing tension between the need to consider the history of the data to understand which of these correlations you could even set up, let alone trust for further work, and the imperative of feeding data like this to AI systems and accelerate the production of potential inferences by using some of these objects as training data for a variety of algorithms.

https://doi.org/10.14361/9783839467664-017 https://www.inlibra.com/de/agb - Open Access -

Figure 1: A schematic representation of the research data ecosystem. Translated in English from Leonelli 2018a.



My perspective on the epistemology of data originates in the consideration of the multitude of ways in which people interact with the world and generate artifacts (images, numbers, textual descriptions) that are meant to capture or document these interactions in some way. Many interactions with the world produce some kind of object or atrace of some sort, and those objects may or may not be processed as data. In my view, data does not become a representation of the world until it gets clustered, ordered and interpreted in a particular

https://doi.org/10.14361/9783839467664-017 https://www.inlibra.com/de/agb - Open Access -

kind of way. In other words, data models represent specific phenomena; data represent objects that are processed and stewarded for their potential to serve such representational purposes. Once a decision has been made about what data may be evidence for, the resulting models are used to interpret the data and acquire knowledge, which in turn informs further interactions with the world (Leonelli 2016).

There is a fragility and unreliability to the current data system, since it is hard to distinguish datasets that have been well-maintained and updated from those that have not been checked and adequately curated (Floridi/Illari 2014). Datasets available online are limited and biased, and there is a multitude of vested interests around which types of data become easier to access or more valuable to trade (Kitchin 2014; Mackenzie 2017). All these considerations contribute towards enhancing the in-practice opacity of data ecosystems, making it often near-impossible to unravel such opacity in a way that fosters intelligibility.

4. Reproducibility and the illusion of transparency

Situating data movements within a broad landscape which includes AI technology, as well as research institutions, industry, policy-making and various other publics and stakeholders lead to the investigation of the idea, which is common among supporters of Open Science, that increasing the transparency through which data processing is documented and explained may contribute to lessening the opacity characterizing large data ecosystems (Leonelli 2023).

One example of this approach is the discussion of reproducibility, which includes the application/consideration of a scientific method but also that of the priorities, goals and interests of the various institutions engaged in science. In particular, it interrogates what it means for data-intensive analyses to be scrutinized, reenacted and understood, no matter how complex the relevant sources, processes and analytics may be. The debate on reproducibility is a good representation of how the use of data-hungry AI in research raises issues beyond the traditional questions asked of the statistical methods used to validate datasets and analyses. While we witness a large increase of integrated research efforts and the application of algorithms across large domains, there are also increasing problems in getting people who are specialists in different parts of the research ecosystem to interact with each other and assess the value and significance of each other's work. Lots of confusion is generated

https://doi.org/10.14361/9783839467664-017 https://www.inlibra.com/de/agb - Open Access -

by questions around scales and who can be trusted in this kind of landscape. Peer review is increasingly acknowledged not to work well when attempting to check data quality and incentives for researchers to engage in careful scrutiny of peers remain scarce. A strong reliance on automated research systems complicates matters further. Within such a landscape, reliance on AI creates even more a sense of research processes increasingly being impenetrable black boxes, whose inner mechanisms and functions remain invisible and unreachable to observers. There is a growing mistrust of scientific results even by actual scientists, let alone members of the public. The moral economy of science, strongly grounded on trust among peers, is being disrupted. It is in this climate of mistrust and uncertainty that the question of opacity associated with the use of AI in research has acquired poignance and prominence, prompting calls for explainable and transparent uses of AI for discovery and warnings against the reliability of systems that do not seem accessible for scrutiny (Council of Canadian Academies 2022).

There is little doubt that we are witnessing a real challenge in contemporary applications of AI to research processes and that questions around how such applications should be scrutinized and integrated into existing methods are urgent and unresolved. I do not think, however, that the main problem lies with the opacity of research systems per se. To an extent, research processes have always been and will always be opaque. It is simply impossible to account for every aspect of a research process, including the tacit knowledge used to calibrate instruments, set-up experiments, adapt methods to the specific situation and materials on which research is being carried out. The question is, rather, what forms of opacity end up being damaging to research and its role in society.

Reproducibility is often evoked as a solution to the problem of opacity in research, including in AI applications. You want to try and make sure that when you repeat a piece of research, there are some consistent results obtained. This seems like a fair requirement – a good thing for scientists to try and strive for. Consequently, there is a push to try and have more transparent sharing of information, particularly meta and para information around data sets, so it is easier to evaluate how data have been created and processed, with the aim to reproduce these conditions. Some even argue that the more we know about the process of research – the more we can capture, publish, debate and the more we may be able to automate some of those processes in interesting ways that can complement and sometimes even substitute humans who are involved in

https://doi.org/10.14361/9783839467664-017 https://www.inlibra.com/de/agb - Open Access -

a discovery (for a depiction of the debate, see for instance The Royal Society 2019).

Despite its promise, reproducibility however is not a silver bullet. To begin with, there are many different types of reproducibility (Leonelli 2018b; Leonelli/Lewandoswky 2023) that range from the more classical computational reproducibility, which assumes total control in the system, to reproducible observations that assume very low controls in terms of statistics, goals and judgments. There is a big discrepancy in how different domains depend on statistics and computation, not just as a tool to get the research done, but as a reasoning tool to make inferences. Clinical trials are typical examples of hypothesis testing situations where methods and results are expected to conform to detailed and sophisticated advance plans, but there isa lot of exploratory research that operates differently. How stable you assume your background knowledge to be also makes a difference, as well as whether or not you think it is acceptable for researchers to declare that they've exercised their subjective judgment in setting up their technical system. In evidence-based medicine this is something that people are not comfortable admitting, because the idea that expert judgment is used in someone's work is regarded as making research subjective and potentially unreliable. There is a desire to reach conclusions in ways that do not depend on the specific circumstances of the researcher's judgment. Nevertheless, such independence is yet to be found (Leonelli forthcoming).

I am worried about the fact that we are often confronted with a very narrow interpretation of reproducibility when thinking about how this principle operates in research practice. Highly controlled experiments which have pre-specified goals have come to exemplify best practice for some reason, and rigorous research, partly because they tend to adhere more easily to potentially misguided ideas about objectivity in science. This ends up doing no justice to other research methods that are accused of being unscientific. We are losing important expertise by creating priorities and rankings over what kind of methods should be prioritized in research. Qualitative research traditions get put aside and there is a strong emphasis on hypothesis-driven research to the expense of data mining, where in many cases hypotheses are not specified in advance. A narrow interpretation of reproducibility sets up a false dichotomy between quantitative approaches and more hermeneutic, judgment-based approaches, which devalues the role of expertise and embodied knowledge in dealing with data, but also the very significant social context in which research is happening. This does not resolve at all the problem of reproducibility to start with, because it really doesn't necessarily help to distinguish between what may be an

https://doi.org/10.14361/9783839467664-017 https://www.inlibra.com/de/agb - Open Access -

unintentional mistake, what may be an actual case of cheating, or what may be a variation which is due to differences in research conditions, which may be actually quite interesting, and the situations where the best guess is to constructively poke at accepted facts. This pursuit of reproducibility as an overarching epistemic value, particularly when focused on increasing transparency in documenting research methods as a key solution, is not some sort of magic trick or a magic formula for what might constitute good science. It doesn't necessarily fix concerns around research quality, since simply providing more information about data processing does not necessarily help evaluate such processes – especially in situations where the processes in question are so vast and complex that they cannot be synthesized or comprehended. Nor does it provide some universal solution, particularly because there are all these different ways in which you can interpret the possibility, which are active and useful in different ways, depending on what kind of domain and what kind of practices you're adopting.

To continue, it does not necessarily help to address systemic issues with who is incentivized to make their data available, who is incentivized to curate data properly, and how people are rewarded for documenting their data management decisions – issues that are at the root of many of the problems prompting calls for reproducibility. Attention should be redirected towards the thinking of existing assumptions about hierarchies of evidence, where they come from and what their effects are likely to be when they become part of the research infrastructures, including algorithms and machine learning applications. More reflection also needs to go into what kinds of data should be preserved for long term storage, dissemination and sharing, and under which conditions, and how, such choices may be made accountable within expansive data ecosystems (Zook et al. 2017; Elliott et al. 2021). Most of our digital data ecology is ephemeral, with few attempts to think about data collection and data storage online for more than 10 years. Algorithms are currently trained on a rather serendipitous collection of data, whose availability depends on who gets funding at a particular point in time and how tractable data are digitally. There is a significant skew in the kind of machine-readable data that can be utilized for algorithmic elaboration. Finally, there is a sidelining of research geared towards involving transdisciplinary communities and expertise, accompanied by an emphasis on short-term outcomes and low-hanging fruit that stays away from complex, heterogeneous datasets in favor of homogenous, easy-to-handle ones. All this creates skews in the data system feeding AI, which is sure

https://doi.org/10.14361/9783839467664-017 https://www.inlibra.com/de/agb - Open Access -

to have significant implications for the kinds of questions AI can help answer more accurately, as well as for the content of those answers.

5. Cracks in the looking glass: AI and the data ecosystem

What are the implications of these reflections for AI? Narrow interpretations of reproducibility tend to go hand-in-hand with an insistence on computational tools to automate research processes, with the hope that AI can provide a quick fix for problems around the quality of research – perhaps even help researchers to replicate experiments and methods without effort. This constitutes, in my view, a vicious circle. There is insistence on narrow, computational understandings of reproducibility because this seems to be a watertight way of thinking about checking the quality of a particular set of algorithms. However, this disregards the problems that arise through systems that are difficult to automate, such as quality checks for domain specific data obtained from complex experiments and observational methods, as well as the limits and histories entrenched in the current ecosystem of widely accessible, machine-readable data useable for training AI tools.

There is a gulf opening between discussions on reproducibility and what constitutes reliable training of data, reliable methods and reliable algorithms, which can be evaluated through those particular tools and others that are seen to be much less reliable because they just don't fit this kind of more automated, quick, computational check. It is crucial to address how one ought to formulate, assess and acknowledge the qualitative judgments that accompany data driven methods. In many AI discussions there is a tendency to think that judgements made around data – in calibrating data, in thinking about what is actually being processed, in picking training data, in creating artificial data that may fit new analytic tools – are important, but will be superseded by the emergence of better and better AI technology and more and more data sets. The hope is that the biases and the kind of externalities produced by judgments in those respects will disappear within a beautifully irrefutable and increasingly objective system. By contrast, I and many other scholars interested in data-intensive AI are seeing it as something quite different. On the one hand, there is reluctance to acknowledge the methodological choices and assumptions made at different points in time within the research process, since those are seemingly in tension with such promises of progress. On the other hand, the power exercised by few corporate platforms with the resources to garner,

https://doi.org/10.14361/9783839467664-017 https://www.inlibra.com/de/agb - Open Access -

mobilize and analyze data – thereby deciding which data are valuable, how and for which purposes – is exasperating the bias, serendipity and digital divides already thriving in data-intensive systems, thereby increasing the risk of losing perspective on what data are reliable, representative and fit for purpose, and under which circumstances. We are making tremendous strides in developing large language models for translating between English, Mandarin, German or French, but could there be a comparable data processing effort to do the same for minority languages? Genomic sequencing is increasingly cheap and done on a scale that was unimaginable ten years ago, but how can we ensure that comparable attention is devoted to collecting, mining and interpreting data about metabolism, development and morphology, thereby probing alternatives to genetic determinism? Investment in clinical data on specific pharmaceutical treatments drives medical advancements, but how can the development of a comparable data ecosystem to support research on lifestyle and social interventions, which may have an equal or better chance to improve individual health and wellbeing, be ensured? Making AI less opaque and more accountable includes interrogating the make-up, evolution and future directors of the data ecosystem, taking into account the multiple goals which AI – and the underpinning data resources – are meant to serve.

List of references

Bowker, Geoffrey C./Star, Susan Leigh (1999): Sorting Things Out: Classification and its Consequences, Cambridge, MA: The MIT Press.

Council of Canadian Academies (2022): Leaps and Boundaries. The Expert Panel on Artificial Intelligence for Science and Engineering, Council of Canadian Academies, Ottawa, ON: Council of Canadian Academies (https://www.cca-reports.ca/wp-content/uploads/2022/05/Leaps-and-Boundaries_FINAL-DIGITAL.pdf).

Elliott, Kevin C./Cheruvelil, Kendra S./Montgomery, Georgina M./Soranno, Patricia A. (2016): "Conceptions of Good Science in Our Data-Rich World." In: BioScience 66/10, pp. 880–889.

Floridi, Luciano/Illari, Phyllis (eds.) (2014): The Philosophy of Information Quality, Cham: Springer.

Kitchin, Rob (2014): The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences, London: SAGE.

https://doi.org/10.14361/9783839467664-017 https://www.inlibra.com/de/agb - Open Access -

Leonelli, Sabina (2016): Data-Centric Biology: A Philosophical Study, Chicago and London: Chicago University Press.

Leonelli, Sabina (2018a): La Ricerca Scientifica nell'Era dei Big Data, Milan: Meltemi Editore.

Leonelli, Sabina (2018b): "Re-Thinking Reproducibility as a Criterion for Research Quality." In: Research in the History of Economic Thought and Methodology 36B, pp. 129–146.

Leonelli, Sabina (2021): "Data Science in Times of Pan(dem)ic." In: Harvard Data Science Review 3/1, (https://doi.org/10.1162/99608f92.fbb1bdd6).

Leonelli, Sabina (2023): Philosophy of Open Science (Elements in the Philosophy of Science Series), Cambridge: Cambridge University Press.

Leonelli, Sabina (forthcoming): "Is Data Science Transforming Biomedical Research? Evidence, Expertise and Experiments in COVID-19 Science." In: Philosophy of Science.

Leonelli, Sabina/Lewandowsky, Stephan (2023): The Reproducibility of Research in Flanders: Fact finding and Recommendations – KVAB Thinkers' Report 2022, KVAB Standpunten 81, Brussels: Royal Flemish Academy of Belgium for Science and the Arts.

Leonelli, Sabina/Lovell, Rebecca/Wheeler, Benedict W./Fleming, Lora/Williams, Hywel (2021): "From FAIR Data to Fair Data Use: Methodological Data Fairness in Health-related Social Media Research" In: Big Data & Society 8/1 (https://doi.org/10.1177/20539517211010310).

Leonelli, Sabina/Tempini, Niccolò (eds) (2020): Data Journeys in the Sciences, Cham: Springer.

Leonelli, Sabina/Williamson, Hugh F. (2023): "Artificial Intelligence in Plant and Agricultural Research." In: Alok Choudhary/Geoffrey Fox/Tony Hey (eds.), Artificial Intelligence for Science. A Deep Learning Revolution, New Jersey et al.: World Scientific Publishers, pp. 319–333.

Mackenzie, Adrian (2017): Machine Learners: Archaeology of a Data Practice. Cambridge, MA: The MIT Press.

The Royal Society (2019): The AI Revolution in Scientific Research, London: The Royal Society.

Zook, Matthew/Barocas, Solon/boyd, danah/Crawford, Kate/Keller, Emily/Gangadharan, Seeta Peña/ Goodman, Alyssa/et al. (2017): "Ten Simple Rules for Responsible Big Data Research." In: PLoS Computational Biology 13/3, e1005399.

https://doi.org/10.14361/9783839467664-017 https://www.inlibra.com/de/agb - Open Access -