New Ways of Thinking About Data Governance

Os Keyes, University of Massachusetts Lowell
Abraham D. Flaxman, University of Washington

Insights

- Data governance is an important problem, one where our linear conceptual models are allowing harms to go unobserved and unaddressed.
- A *relational* view of data governance can correct these issues.
- Such a view can be easily integrated into the structure of data-generating technologies and projects.

Regular readers of *Interactions*—and anyone tangentially involved with automation and data—will be familiar with debates over the need to improve how datasets and algorithmic systems are governed and overseen [1]. The HCI community, with its interests and expertise in applying a sociotechnical lens to the design of technology, has played an important role in this by designing and deploying new structures for governing both data and algorithms [2].

But these new structures depend on theories of what data is, how people relate to it, and what these relationships of trust and accountability should look like. In this article, we argue that the conventional theories for thinking about these questions are too limited, and we want to encourage HCI practitioners to adopt a new one. We see Salomé Viljoen's idea of data as a relational (and relationally governed) concept as the appropriate theory [3]. We also briefly explore our own attempts at implementing relational data governance. Taking a relational approach will strengthen the applicability of novel structures for data governance and enhance our ability to imagine better ways of approaching these problems.

THE TROUBLE WITH THEORIES OF DATA GOVERNANCE

Any intervention into data governance—whether it takes the form of legislation, voluntary proposals, or internal organization practices—

IMAGE BY HOBBITART / SHUTTERSTOCK.COM

ultimately depends on having a theory (however tacit) of what data is, which entities are party to it, and the interplay between those parties. When practitioners and researchers in HCI and other data-related fields think about relationships around data, we often imagine two or potentially three parties. These include the *data subject* (the source of the data), the *data processor* (who collects and organizes it), and, if data is not entirely organizationally self-contained, with the *user* also being the *data processor*, then perhaps also the *data reuser* (who takes the data and uses it to inform actions). Furthermore, we often imagine the relationships between these parties as somewhat linear, as seen in Figure 1. Relationships pass from subject to processor and then from processor to reuser.

But this way of tracking the consequences of data is inaccurate. Empirically, we can say with some certainty that there are more parties to data than a three-party linear model encompasses. For one, there are more *subjects*: Data that is collected about a person is not only used for decision-making about that individual but also about people who are categorized as similar to them in some relevant way. Additionally, there are more *reusers*: Data usage in practice often involves an entire string of reusers and processors. As David Gray Widder and Dawn Nafus have documented, the AI supply chain involves data moving between—and being modified and reused by—a multitude of reusers, sometimes very distant from the original context of collection [4].

The reason why only a few of these reusers are visible is very simple. Although we are using the language of relationships here, conceptualizations of relationships around data are often based on a *contractual* understanding of it: Parties need to have a direct and formal connection to be in relation to each other. In practice, this means that models of data relationships and governance fail to consider situations where different parties might never directly encounter each other, but nonetheless affect each other. Mary F.E. Ebeling provides a beautiful illustration of this problem in her book *Healthcare and Big Data: Digital*

Specters and Phantom Objects, which documents her efforts to identify the path that data about her and her pregnancy took—traveling from her doctor to insurers, advertisers, and even magazine publishers [5]. Ebeling struggled to connect these dots, and the length of the chain meant that many of the intermediaries, such as her doctor, were also unaware of where the data was going.

As this example suggests, our view of who is party to data isn't just theoretical pedantry; it's something that shapes what forms accountability can take and what harms are and aren't subject to accountability. A contractual view is limiting in what parties and actions it considers, and it leaves many problems unaddressed. But there are alternatives, one of which—the focus of this piece—is Salomé Viljoen's idea of *relational data governance* [3].

RELATIONAL GOVERNANCE

Drawing on some of the same concerns we have, Viljoen proposes reconceptualizing data governance so that it isn't structured on a contractual basis, but rather by looking for the actual relations between entities—the ways different bodies involved with the data affect each other. A central part of relational data governance is that reusers are understood to inherently be in relation not only to data processors but data subjects too, resulting in an image of data governance such as the one depicted in Figure 2. This extends all the way down the chain of parties that use data.

To demonstrate what this looks like in practice, we return to Ebeling, who couldn't trace her data and seek accountability for how it was used. This happened due to the lack of a formal relationship between her and many of the reusers who stored and relied on her data. In a world where data governance would be conceptualized as relational, these problems would go away because of the differences in how data would travel and be understood. Every body Ebeling had struggled to reach—from the electronic health record organizations to the marketing companies—would be far easier to find. By accepting her data, they would have to agree to make themselves known to and in relation to her. As a consequence of this being-in-relation, they would also have to respond to her requests and provide, at the very minimum, transparency about the use of her data.

IMPLEMENTING RELATIONAL GOVERNANCE

Relational governance work tends to assume a legal framework, which poses a problem. Viljoen's work imagines a particular kind of subjectivity and orientation, backed and made present in the force of law. But much of what HCI and AI researchers do is in a more informal or practice-oriented space—one where the law is often silent or, in international collaborations, somewhat moot. Governance is instead a matter of the informal relationships and expectations that come with access to data. What does it look like to try to implement relational governance in such a space? To try to enculturate a new governance approach with informal tools?

This is precisely what we tried to find out, in collaboration with the U.S. Census Bureau, as part of a larger project called Pseudopeople [6]. The goal of Pseudopeople is to generate a simulated version of key census datasets, which can be used to test privacy protection and record-linking techniques without requiring researchers to access the real, confidential census data. Although the data is synthetic, it still aims to represent the entire U.S. population, making them—under Viljoen's conception of governance—*data subjects* in the sense that they are subject to the data's use. This makes the involvement of American publics in governance—and in relation to the researchers expected to reuse the data—central. Such urgency is not new to the Census Bureau, making Pseudopeople an additionally attractive site of experimentation. As a result of both general government expectations, such as the Evidence Act and the work of the Council of Professional Associations on Federal Statistics, and the Census's own specific imperatives, the organization has a long and unique history of prioritizing subject protection and statistical innovation.

To provide meaningful opportunities for this involvement, we began by ensuring that reuser access to the data was constrained in very particular ways. Rather than making data publicly available or accessible only through private requests to the Pseudopeople developers, reusers who seek access have to make a public request via GitHub. This requires them to describe the project they are using the data for, who will have access to it and under what conditions, and how data will be disposed of or updated.

The point of making data public is not simply "transparency," in the sense of some abstract virtue. Instead, the goal is to enable actual discussion of reusers' research and data needs with members of the public. We aim not only to inform decisions about whether data access should be granted but also to create ongoing and long-term relationships of responsiveness and accountability between members of the public, who constitute data subjects, and the data reusers seeking access to Pseudopeople. It is for this reason that reusers are additionally asked, as part of their initial request, to commit to being responsive to those publics and to treat the conversation around the request as something ongoing throughout the life of the project. This approach to reuser access also means that reusers are by necessity in contact and relation with subjects. We hope this results in relationships between reusers and data subjects that are thicker than those in "click to download" approaches to dataset access. Rather than abstract data entries, data subjects who participate in governing access requests and the terms of reuse are concrete, immediate people to reusers.

Once submitted, GitHub requests remain public to act both as a log of any access that was granted and any discussion that surrounded it, as well as a site for that discussion—during and after any permission is granted or denied. Public members who are interested or relationally involved in the dataset can thus ask questions, raise concerns, and articulate their own approval or disapproval of proposed plans in perpetuity. In practice, this

There are more parties to data than a three-party linear model encompasses.

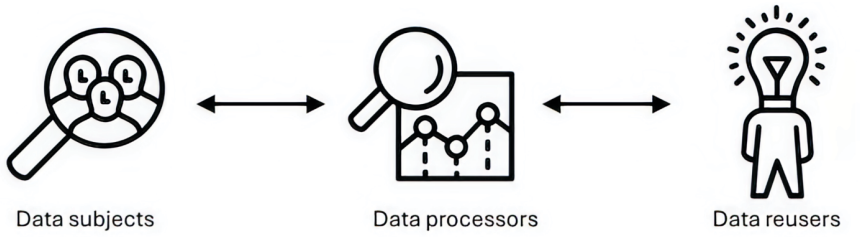

Figure 1. Data governance relations are often conceived as passing from the data subject (the source of the data) to the data processor (who collects and organizes it) to the data reuser (who uses it to inform actions).

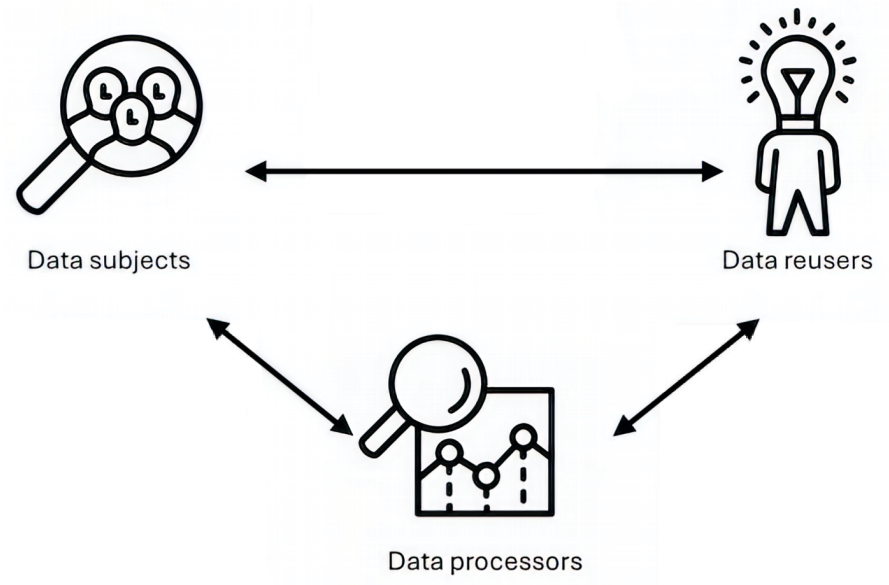

Figure 2. An alternative relational structure for data governance is achieved when the data reusers are understood to inherently be in relation with the data processors, as well as with the data subjects.

mechanism also demonstrates some of the limitations and difficulties of integrating relational governance with conventional platforms for code and data development.

Although GitHub issues and discussion threads are the most accessible option for public discussion, "most accessible" says more about the overall ecosystem of development options than it does about GitHub. In practice, it is still a jury-rigged remedy that has its own barriers. GitHub, as a platform for developing and distributing code, is not a natural place for members of the general public to have accounts or necessarily feel comfortable. Selecting it was largely based on the fact that every other option was worse. Most platforms for code and data storage are even more siloed, with data platforms in particular, such as Data Dryad or Kaggle, designed primarily for data preservation. They prioritize keeping data in a static and fixed form rather than treating data and the decisions around it as a site of active movement. Relational governance also offers an opportunity for HCI researchers interested in the more traditional design-oriented aspects of the field to prototype and develop platforms that treat this form of governance as a priority.

CONCLUSION

Data governance is difficult, but it is also urgently important and a vital site of possibility for the creation of new and better futures. Frameworks for thinking about governance often replicate existing and failed ways of thinking about power and relations. In this article, we have documented efforts to implement a different way of thinking, which focuses on treating data and its consequences as relationally tied to the parties affected by data. Instead of repeating the status quo, implementing this new conceptualization of data and its governance offers an opportunity to experiment with new ways of thinking about and managing the consequences of data-centric systems.

References

1. Wong, J. Data practices and data stewardship. *Interactions 30*, 3 (2023), 60–63.
2. Delgado, F., Yang, S., Madaio, M., and Yang, Q. The participatory turn in AI design: Theoretical foundations and the current state of practice. *Proc. of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, 2023, Article 37, 1–23.
3. Viljoen, S. A relational theory of data governance. *The Yale Law Journal 131*, 2 (2021), 573–654.
4. Widder, D.G. and Nafus, D. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society 10*, 1 (2023); https://doi.org/10.1177/20539517231177620
5. Ebeling, M.F.E. *Healthcare and Big Data: Digital Specters and Phantom Objects*. Palgrave Macmillan, 2016.
6. Haddock, B. et al. Simulated data for census-scale entity resolution research without privacy restrictions: A large-scale dataset generated by individual-based modeling [version 2; peer review: 1 approved, 2 approved with reservations]. Gates Open Research 2024, 8:36; https://doi.org/10.12688/gatesopenres.15418.2

Os Keyes is a science and technology studies researcher currently working as a postdoctoral fellow at the Evans Lab at the University of Massachusetts Lowell. They have published at venues including CHI, CSCW, and ASSETS, and written for *Vice*, *Logic(s)*, and *Scientific American*, among others. They are the inaugural recipient of an Ada Lovelace Fellowship.
→ os_keyes@uml.edu

Abraham D. Flaxman is an associate professor of global health at the Institute for Health Metrics and Evaluation at the University of Washington. He is currently leading the development of new methods for cost-effective analysis with microsimulation and is engaged in methodological and operational research on verbal autopsy.
→ abie@uw.edu

FIGURES COURTESY OF THE AUTHORS

DOI: 10.1145/3774948