Data & Policy (2021), 3: e6
doi:10.1017/dap.2021.4

COMMENTARY

Reimagining data responsibility: 10 new approaches toward a culture of trust in re-using data to address critical public needs

Stefaan G. Verhulst

Co-Founder, The GovLab at New York University
Corresponding author. E-mail: stefaan@thegovlab.org

Received: 27 January 2021; **Revised:** 10 April 2021; **Accepted:** 22 April 2021

Keywords: data responsibility; data4good; data stewardship; privacy; fair

Abstract

Data and data science offer tremendous potential to address some of our most intractable public problems (including the Covid-19 pandemic). At the same time, recent years have shown some of the risks of existing and emerging technologies. An updated framework is required to balance potential and risk, and to ensure that data is used responsibly. Data responsibility is not itself a new concept. However, amid a rapidly changing technology landscape, it has become increasingly clear that the concept may need updating, in order to keep up with new trends such as big data, open data, the Internet of things, and artificial intelligence, and machine learning. This paper seeks to outline 10 approaches and innovations for data responsibility in the 21st century. The 10 emerging concepts we have identified include:

End-to-end data responsibility
Decision provenance
Professionalizing data stewardship
From data science to question science
Contextual consent
Responsibility by design
Data asymmetries and data collaboratives
Personally identifiable inference
Group privacy
Data assemblies

Each of these is described at greater length in the paper, and illustrated with examples from around the world. Put together, they add up to a framework or outline for policy makers, scholars, and activists who seek to harness the potential of data to solve complex social problems and advance the public good. Needless to say, the 10 approaches outlined here represent just a start. We envision this paper more as an exercise in agenda-setting than a comprehensive survey.

Policy Significance Statement

Policy makers currently lack a framework to balance the potential of technology against the risks it also poses. This is perhaps especially true in the rapidly changing field of data science, where new horizons are being charted on a regular basis. The purpose of this paper is to provide policy makers—and others—with the outlines of an emerging framework. We outline 10 areas for data responsibility that policy makers can use to guide decisions at every stage of the data life cycle. The paper is exploratory. It is intended more as an exercise in agenda-setting than as a definitive statement. Nonetheless, these ten areas may prove useful in evaluating data uses and applications, and they may also help identify areas that merit further research or resources within the data ecology.

Data, data science, and artificial intelligence can help generate insights to inform decisions and solutions for some of our most intractable societal challenges, including climate change, global pandemics and health, food insecurity, and forced migration. The possibilities are immense—but only if the associated technologies, practices, and methodologies are used responsibly and appropriately.

While we have witnessed rapid innovation in how data are being leveraged, existing policies and tools for data responsibility have struggled to keep up. Most of the existing data governance frameworks and laws are still based upon the Fair Information Practice Principles (FIPPS) which were designed in the 1970s[1] for a different data environment. As a result, many consider the FIPPS as the necessary baseline but not sufficient or applicable when considering new data challenges.[2] As concern grows among policy makers, civil society and citizens about the harmful (and often unanticipated) effects of data and technology, it is imperative to reimagine data responsibility by developing and expanding new approaches and concepts for how data could be harnessed responsibly toward the public good.

Below we present 10 emerging approaches, grouped into three broad priority areas, that add up to a framework for data responsibility for policy makers, scholars, activists, and others working in the field. Taken together, these approaches and concepts can provide for more participatory processes, ensure greater equity and inclusion, and embed new practices and procedures in the way organizations across sectors collect, store, and use data. In arriving at these priorities, we begin from a recognition that data has both risks and potential. No technology is an unmitigated good; finding ways to maximize opportunity while limiting risks is key to generating new insights that can lead to societal benefits while simultaneously protecting individual and collective rights.

As indicated in Figure 1, the three priority clusters for responsible data that we identify are: Rethinking Processes and Systems, Rethinking Duties and Roles, and Rethinking Rights and Obligations.

While our discussion is preliminary and necessarily incomplete, these 3 priorities, along with the 10 components approaches we include, can help guide practices and policy-making surrounding the use of data, and they can also help identify areas for future research. The following discussion should be seen as an exercise in agenda-setting rather than a comprehensive survey; and is meant to give policy practitioners a rapid update of the emerging innovations. One of our primary goals is to identify which data responsibility approaches merit further attention (and resources) in order to develop or enhance a culture of trust in using and reusing data to address critical public needs.

1. Priority #1: Rethinking Processes and Systems

Data responsibility cannot be achieved in a piecemeal manner. Balancing risk and potential requires a systematic rethink of how organizations handle information across the data lifecycle. This systematic rethink can in turn be broken up into at least the following specific components.

1.1 End-to-end data responsibility

Using data and data science for the public good means promoting responsibility throughout the data lifecycle.

To generate value and effect change, data must be transformed into insight, and insight must be transformed into action. The sequences or stages of that enable value creation make up the data lifecycle(El Arass and Souiss, 2018). Each phase of the data lifecycle—planning, collecting, processing, sharing, analyzing, and using—plays a formative role in generating value for stakeholders(Stodden, 2020). The respective phases also pose a number of risks and potential harms. Proactively assessing and

[1] https://link.springer.com/chapter/10.1007/978-94-017-9385-8_12.

[2] https://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=3759&context=mlr.

https://doi.org/10.1017/dap.2021.4 Published online by Cambridge University Press

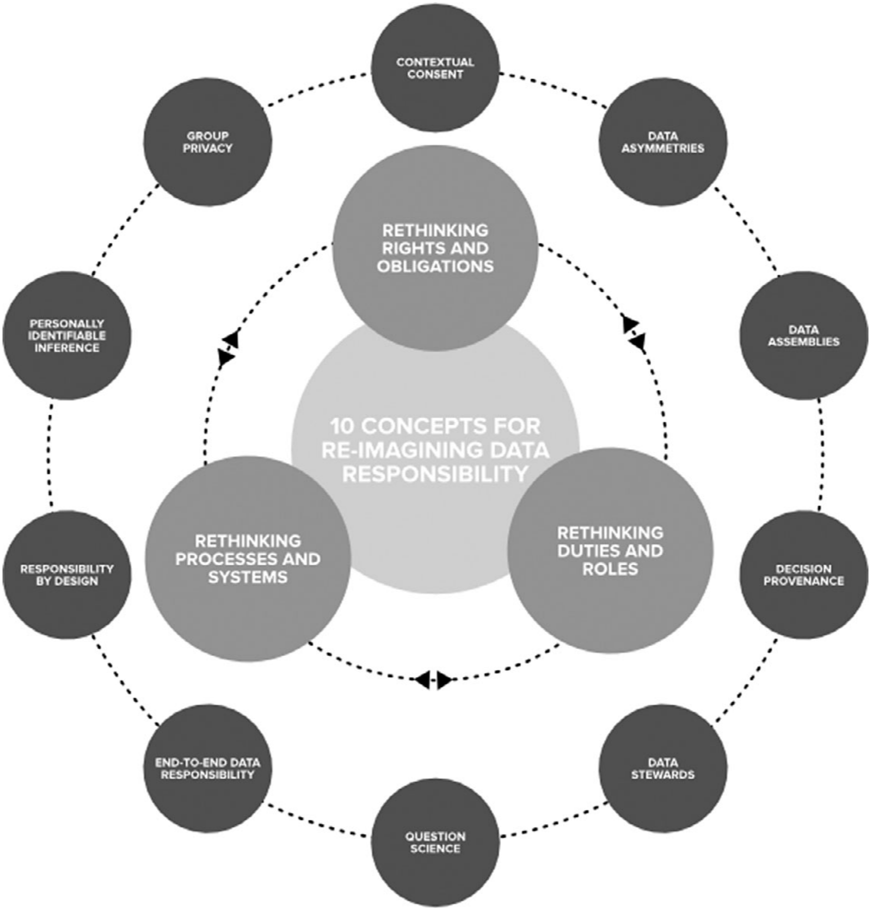
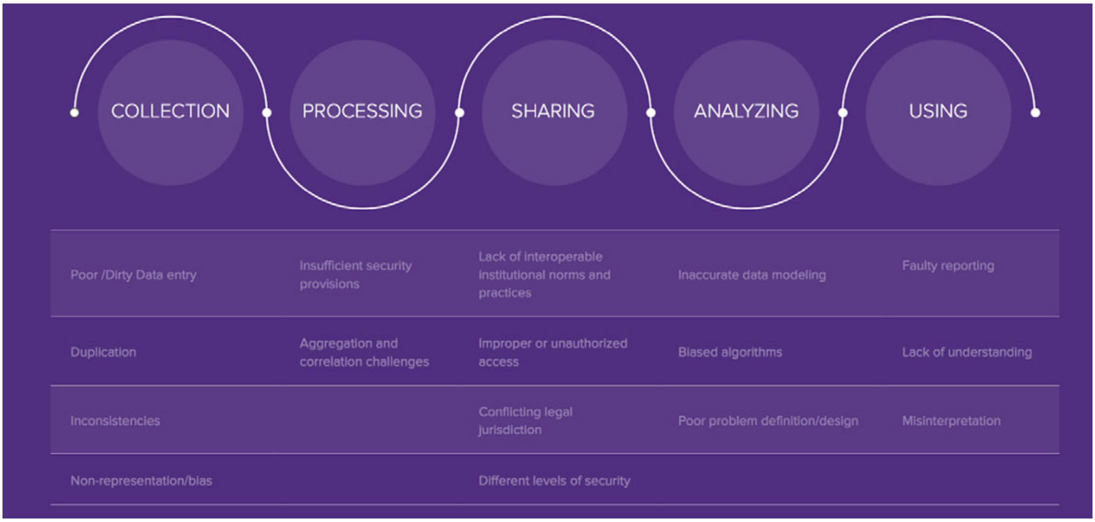

CONTEXTUAL CONSENT
GROUP PRIVACY
DATA ASYMMETRIES
RETHINKING RIGHTS AND OBLIGATIONS
PERSONALLY IDENTIFIABLE INFERENCE
DATA ASSEMBLIES
10 CONCEPTS FOR RE-IMAGINING DATA RESPONSIBILITY
RESPONSIBILITY BY DESIGN
RETHINKING PROCESSES AND SYSTEMS
RETHINKING DUTIES AND ROLES
DECISION PROVENANCE
END-TO-END DATA RESPONSIBILITY
DATA STEWARDS
QUESTION SCIENCE

Figure 1. *Re-imagining data responsibility.*



COLLECTION	PROCESSING	SHARING	ANALYZING	USING
Poor /Dirty Data entry	Insufficient security provisions	Lack of interoperable institutional norms and practices	Inaccurate data modeling	Faulty reporting
Duplication	Aggregation and correlation challenges	Improper or unauthorized access	Biased algorithms	Lack of understanding
Inconsistencies		Conflicting legal jurisdiction	Poor problem definition/design	Misinterpretation
Non-representation/bias		Different levels of security		

Figure 2. *End-to-end data responsibility.*

https://doi.org/10.1017/dap.2021.4 Published online by Cambridge University Press

mitigating risks (as well as capitalizing on opportunities) at each stage constitute the requirement of *End-to-end data responsibility. (see Figure 2)*

This approach is vital and goes beyond much recent thinking on data responsibility, which typically emphasizes risks at a single stage of the lifecycle—for example, algorithmic bias when analyzing data; secure cloud approaches when processing data or the need for consent at the collection stage. Mitigating risk at a single stage may be a good start, but it is not sufficient to accomplish true data responsibility. For example, we could envision a scenario in which collection stage risks are addressed by proactively collecting data on minority communities to mitigate bias; however, if acceptable sharing and use protocols are not put in place, there could be downcycle risk through exposure of sensitive demographic data. This is just one (hypothetical) illustration of why data responsibility must be prioritized more systematically, throughout the entire data lifecycle.

1.2 From data science to question science

> *Data-driven projects should arise as a result of genuine need, and not simply because an organization wants to access or use data that happens to be available. Toward that end, we need a new science of asking questions in order to identify priorities.*

When using data for the public good, organizations too often focus on the supply side of the data equation ("What data do we have or can access?") rather than the demand side ("What problems do we face, and what problems can data solve?"). As such, data initiatives often provide marginal insights and generate privacy risks while utilizing data that may not in fact be relevant to our most important societal problems (Auxier and Lee, 2019). What is required is a new approach that begins by asking the right questions about priorities and needs. In fact, we need a new science (and practice) of questions—one that identifies the most urgent needs in a participatory manner,[3] and that matches data supply and demand so as to create a more targeted approach to problem solving (Verhulst, 2019).

In order to identify the questions that really matter, domain knowledge and data science expertise must be integrated. This can be done by targeting "bilinguals": individuals who are proficient in a given domain and also skilled in the fields of data science and/or statistics (Meer, 2015). The expertise of these individuals can be combined with the inputs of a wide range of users and citizens, helping formulate questions that are both relevant to societal issues and technically feasible. By ensuring the relevance of data projects, such an approach increases the likelihood of their impact and also conserves human and financial resources.

The Governance Lab's 100 questions initiative seeks to do just this, developing a process that takes advantage of distributed and diverse expertise on a range of given topics or domains so as to identify and prioritize questions that are high impact, novel, and feasible, in the process ensuring that already scarce data resources are adequately used.[4] In this work, we have identified 10 domain fields of interest, such as migration, gender, and climate change, and identify and engage bilinguals who are tasked with identifying 10 important questions in their respective fields. The goal of this initiative is to create a template for future efforts that seek to reinvent how we ask questions about data.

1.3 Responsibility by design

> *Responsibility by design is a practice that seeks to introduces fairness, diversity, transparency, equality and data protection principles into all aspects and methods of data science.*

The tools and methods of data science—including dataset selection, cleaning, preprocessing, integration, sharing, algorithm development, and algorithm deployment—span the data lifecycle. However, the

[3] https://journals.aom.org/doi/abs/10.5465/AMBPP.2019.115

[4] "The 100 Questions Initiative." *The Governance Lab*, http://www.the100questions.org/.

https://doi.org/10.1017/dap.2021.4 Published online by Cambridge University Press

need to ensure responsibility in all these processes is often overlooked, leading to biases and inequities in the way data are collected, preprocessed and processed, correlated, and reproduced. Such problems are exacerbated by the rise of machine learning and artificial intelligence techniques, as well as by the era of big data. As Julia Stoyanovich, founder of the Data, Responsibly consortium[5] explains: "If not used responsibly, big data technology can propel economic inequality, destabilize global markets and affirm systemic bias."[6] More recently, she and her colleagues have called for the need to consider data equity more broadly—as it relates to representation equity; feature equity; access equity; and outcome equity.[7] In order to mitigate such problems, it is necessary to utilize the principles and practices of *responsibility by design*. Although similar in some respects, responsibility by design goes beyond the earlier fairness, accountability, and transparency (FAT) concerns that are now widespread in the data community and that seek to address algorithmic bias (Lepri et al., 2018). Responsibility by design focuses on the full data science lifecycle—not just the design and deployment of particular algorithms (Stoyanovich et al., 2017). It also borrows much of what is now known as privacy by design (Kingsmill and Cavoukian, n.d.; Cavoukian et al., n.d.) and embraces privacy enhancing technologies.[8]

Responsibility by design—and more recently the FAT community—suggests a "holistic consideration" of the sociotechnical systems that introduce the above concerns. As a framework, it ensures a well-planned and intentional approach to data science, and contributes to end-to-end data responsibility.

2. Priority #2: Rethinking Duties and Roles

The above priorities and steps require widescale and systematic transformation of organizations involved in the data lifecycle. But who will implement and oversee such change? In an era of data responsibility, we need to reconceptualize our notions of key actors and stakeholders, as well as the nature of their duties and roles. Here, we suggest three specific steps toward achieving the required transformation.

2.1 Decision provenance

To implement data projects with accountability, legitimacy and effectiveness, it is necessary to keep track of the decision flow - who makes decisions when - known as decision provenance, throughout the data lifecycle.

Data projects are complex undertakings that often involve disparate chains of authority and expertise applied to large, diffused problems. Often, the roles and responsibilities of different actors are poorly understood, and their actions poorly tracked; this complexity makes accountability difficult. Systematic data responsibility requires what is known as *decision provenance*—defined as a framework to "provide information exposing decision pipelines: chains of inputs to, the nature of, and the flow-on effects from the decisions and actions taken (at design and run-time) throughout systems" (Singh et al., 2019). More specifically, decision provenance requires that organizations log each decision, and that all involved stakeholders have a clear understanding of how decisions are made across the data lifecycle, by whom, and for what purposes. For example, in the Governance Lab's Responsible Data for Children Initiative (RD4C), we created a decision provenance mapping tool to support actors designing data investments for

[5] *Julia Stoyanovich—Projects*, stoyanovich.org/p rojects/.

[6] *Data, Responsibly*, https://dataresponsibly.github.io/.

[7] https://dl.acm.org/doi/fullHtml/10.1145/3440889.

[8] https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/privacy-enhancing-technologies-report.pdf?la=en-GB&hash=862C5DE7C8421CD36C105CAE8F812BD0.

https://doi.org/10.1017/dap.2021.4 Published online by Cambridge University Press

children to identify key decision points in the data ecosystem, advancing professionally accountable data practices. The tool requires users to identify specific data activities undertaken, note policies and laws impacting those activities and record all parties supporting decision making across the data lifecycle. And perhaps most importantly, the tool seeks to provide transparency into who is responsible and accountable at each stage, and who should be engaged when making decisions.[9]

Decision provenance bolsters coordination and effective governance because it lets organizations keep track of what happened, and who was involved in data-related decision making. Decision provenance also establishes accountability and creates a "paper trail" (often virtual) in case something goes wrong. This is essential because a lack of accountability results in the potential for data misuse. Insufficient documentation of decision making can also lead to "missed opportunity," or cases when data are not applied in a situation where they could contribute to problem solving. Finally, by exposing decisions that led to successes and failures, decision provenance can contribute to sustainability and scalability, helping identify ways to expand existing data projects and create new ones.

2.2 Professionalizing data stewardship

Organizations may consider creating a new role to establish data responsibility: the Chief Data Steward.

The complexity of data projects means that they typically involve large numbers of people. Often, the roles and responsibilities of these people are unclear, and projects suffer due to a lack of clear definition surrounding roles, hierarchies, and responsibilities. Responsible data handling requires building out a professionalized human infrastructure. In particular, it requires establishing a clear framework for what we call "data stewardship."

Data stewards, as the Governance Lab defines them, are "individuals or teams within data-holding organizations who are empowered to proactively initiative, facilitate, and coordinate data collaboratives toward the public interest." Stewards have three core responsibilities: *collaborate,* working with others to unlock the value of data; *protect* all actors in the data ecosystem from harm that might come from the sharing of data; and *act,* monitoring that data users are appropriately generating value and insights from the data shared. In addition, we have outlined five key roles a data steward must fill, to partner, coordinate, assess risk, disseminate findings, and nurture sustainability within collaboration efforts.[10] Data stewards need to be multidisciplinary to meet the evolving needs of the data ecosystem, as well as fulfill these roles and responsibilities in an agile manner. Data stewards are critical in making data collaboration efforts more systematic, sustainable, and responsible.

2.3 Greater use of data assemblies and public deliberation to establish legitimate re-use

Secondary re-use of data is an effective and efficient approach to data-driven problem solving, but must be guided by public deliberation and policies in order to establish legitimacy and abide by contextual consent.

One area of particular concern in the emerging field of data science and ethics concerns secondary use—that is, the use of data that was originally collected with another purpose in mind. Examples include the use of Geographical Information Systems and mobile data for crisis response or urban planning, and tax return data for civil sector transparency and accountability. Reusing data can save time and resources while providing useful insights, but serious consideration must be given to the specific circumstances

[9] "Responsible Data for Children." *The Governance Lab*, https://rd4c.org/.

[10] "Wanted: Data Stewards: (Re-)Defining The Roles and Responsibilities of Data Stewards for an Age of Data Collaboration." *The Governance Lab,* Mar. 2020, www.thegovlab.org/static/files/publications/wanted-data-stewards.pdf.

https://doi.org/10.1017/dap.2021.4 Published online by Cambridge University Press

under which it is acceptable to reuse data. Absent such consideration and clear guidelines, secondary use runs the risk of violating regulations, jeopardizing privacy, and de-legitimizing data initiatives by undermining citizen trust. Among the questions that need to be asked are what types of secondary use should be allowed (e.g., only with a clear public benefit?), who is permitted to use them, are there any types of data that should never be reused (e.g., medical data?) and what framework can allow us to weigh the potential benefits of unlocking data against the costs or risks.(Young et al., 2020)

Answering these questions requires not only new processes, but also rethinking our notion of stakeholdership (OECD, 2020) and arriving at a new understanding of what types of actors are involved in the data lifecycle. One emerging vehicle for balancing risk and opportunity is the use of working groups or symposia where thought leaders, public decision makers, representatives of industry and civil society, and citizens come together to take stock of existing approaches and design methodologies. The data assembly, for example, is an initiative from The Governance Lab, supported by the Henry Luce Foundation, that seeks to solicit public input on data re-use for crisis response in the United States.[11] In the summer 2020, the Assembly hosted three "mini-public" deliberations with data holders and policy makers, representatives of civic rights and advocacy organizations, and New Yorkers to address the needs of varying stakeholders and move toward a consensus on data-driven response to COVID-19 and other emerging threats. As a result of these deliberations, a data responsibility framework was co-designed indicating the conditions and principles that should guide the re-use of data for pandemic response and recovery.[12]

This data assembly offers just one possible model. Many others exist.[13] The broader point is that data responsibility requires a reconceptualization of what types of actors are involved, and how they are involved. Broadening our notions of stakeholdership and participation will not only lead to better policy, but also work toward flattening some of the inequities and asymmetries we highlight below.

3. Priority #3: Rethinking Rights and Obligations

New systems and new roles will inevitably result in new risks. Data responsibility seeks to address not only known and legacy challenges, but also more dynamic, emerging ones. As part of this process, we must work to secure new rights—and new types of rights--and define new obligations on the part of all actors in the data lifecycle. Here, we outline four such rights and obligations.

3.1 Contextual consent

Contextual consent aims to obtain and use consent in a meaningful way, and to give users a legitimate say in the data ecosystem.

Recent privacy scandals and the introduction of regulations such as GDPR and the Digital Services Act in Europe have brought consent to the center of conversations about data. While much (if not all) data are assumed to be collected with the consent of users, especially if this no other lawful basis, there remain many shortcomings in the way consent is understood and deployed (Richards and Hartzog, 2019). Helen Nissenbaum, a professor at Cornell Tech known for her work in the fields of online privacy and security, argues that "consent as currently deployed" is a "farce" because it gives the "misimpression of meaningful control" (Berinato, 2018). Problems arise when data are collected for one purpose and later used for another, or when users grant consent (e.g., to cookies) without fully understanding the implications of

[11] "The Data Assembly." *The Governance Lab*, https://thedataassembly.org/.

[12] https://medium.com/participo/how-can-stakeholder-engagement-and-mini-publics-better-inform-the-use-of-data-for-pandemic-response-ea1cc5b1ee68.

[13] "Citizens Biometric Council." *Ada Lovelace Institute*, https://www.adalovelaceinstitute.org/project/citizens-biometrics-council/.

https://doi.org/10.1017/dap.2021.4 Published online by Cambridge University Press

doing so. Users may also grant consent because they feel there is no alternative. A more responsible approach would rely on the notion of *contextual consent* (Barkhuus, 2012).

Contextual consent is based on the understanding that obtaining consent means more than simply soliciting a binary "yes" or "no." The context of the interaction must be taken into account.[14] In some cases, the benefits of utilizing data without explicit consent may outweigh the potential harms. For example, at the onset of the COVID-19 pandemic, user location data was important to help decision makers at the state and local levels understand how citizens were responding to social distancing policies and to perform robust contact tracing (Wetsman, 2020). This information also aided population cluster monitoring to inform the risk of COVID-19 spread in certain areas. It is important to note that GDPR does allow for organizations to (re)use personal data without consent if they can demonstrate a "legitimate purpose" and the "necessity" of processing personal data toward that end[15]—which in essence a version of contextual consent. In other situations, contextual consent means engaging users in the data ecosystem in a more meaningful way—at the design, implementation, and review stage of data initiatives. Ideally, data subjects and community partners can be involved in all planning stages of public initiatives driven by local government bodies. Likewise, for cases in which user data will be used for efforts beyond that which it was originally collected, data subjects should be informed about ongoing processes on the retention and analysis of these datasets.

It is important to emphasize that, while contextual consent may in some cases allow for reuse without explicit approval, moving away from the current limited notion of consent does not imply disempowering users or enabling the unfettered reuse of personal data (Dunn, 2016). The goal is to give users a greater say in the data ecosystem and, in doing so, to encourage a more responsible approach to gathering, storing, and using data.

3.2 Righting data asymmetries through data collaboratives

> *Power asymmetries often result from unequal access to data and data science expertise. These data asymmetries threaten to restrict rights and freedoms in a broader socio-economic context.*

A wide variety of imbalances exist when it comes to accessing data, and as well as in the ability to derive meaningful information and insights from it. This phenomenon is true across sectors, and equally applicable to individuals as organizations (Verhulst, 2018). Marginalized citizens and traditionally disempowered demographic groups have less access to data (and its benefit) than those who are better resourced and educated. Similarly, in the private sector, data monopolies—companies with access to vast, almost limitless amounts of data—pose clear threats to competition and innovation. Even in the public sector and civil society, we see large discrepancies in the ability of governments and nonprofit groups to identify data opportunities, and to act on them in furtherance of better public policy or other social goods.

Data responsibility—and, indeed, social responsibility—requires us to address these asymmetries. There are many possible avenues, including increased "data portability"[16] and "data sovereignty"[17] but one of the most powerful tools we have at our disposal is a greater use of data collaboration. In particular, the emerging notion of "data collaboratives," in which information is collectively accessed and acted upon across sectors (Susha et al., 2017), can help break down data silos and ensure that the right data gets to the people who can really benefit from it.[18] There is a growing body of literature, as well as several case studies (Verhulst et al., 2019), to support the use of data collaboratives in mitigating inequalities and data asymmetries (Young and Verhulst, 2020).

[14] "A Contextual Approach to Privacy Online", H. Nissenbaum, Source: https://www.mitpressjournals.org/doi/abs/10.1162/DAED_a_00113.

[15] https://medium.com/data-stewards-network/how-i-learned-to-stop-worrying-and-love-the-gdpr-aa39a12045b3.

[16] https://medium.com/data-policy/data-to-go-the-value-of-data-portability-as-a-means-to-data-liquidity-d8907565e515.

[17] https://medium.com/data-stewards-network/selected-readings-on-indigenous-data-sovereignty-59ffc0b36bfe.

[18] "An introduction to Data Collaboratives", *The Governance Lab,* https://datacollaboratives.org/.

https://doi.org/10.1017/dap.2021.4 Published online by Cambridge University Press

3.3 Transparency surrounding personally identifiable inference

There is a need for greater transparency regarding inferred personal data, or information that is interpreted about a person through multiple data points.

As we spend more time browsing the web, visiting social media platforms, and shopping online, we provide platforms and advertisers with growing amounts of personal data. Some of these data are explicitly disclosed, and some of them are observed (e.g., information about a user's location or browser, which they may not have explicitly meant to share). An emerging third category is inferred data—information that is neither explicitly nor implicitly revealed by users, but that is arrived at through multiple pieces of disclosed or observed personal data. For example, a company may be able to infer a user's gender, race, or religion based on content they are posting on social media platforms. They can then use this information to personalize marketing efforts toward certain users.(Barocas, n.d.).

The goal of inferred personal data, writes Katarzyna Szymielewicz, the co-founder of Panoptykon Foundation, is to "guess things that you are not likely to willingly reveal" such as "your weaknesses, psychometric profile, IQ level, family situation, addictions, illnesses…and series commitments" (Szymielewicz, 2019). The resulting profiles can be invasive and often incorrect, and users typically have little insight into how they are created and how to correct them (or even that they exist). A more responsible approach would require greater transparency, explainability, and human oversight over how such data are collected and used. Users should have more insight into and control over their inferred personal profiles, and they should be able to rectify or otherwise address inferred information about them. More generally, we need more conversations about acceptable uses of inferred personal data (Viljoen, 2020). While some uses (e.g., targeted public services) may be acceptable, others (e.g., determining loan or housing access) may cross a line and should be regulated or outright banned.

3.4 Ensuring group privacy

Data segmentation by demographic group (e.g., by gender or age) has the potential to offer useful insights but it also requires special considerations for group privacy.

Segmenting data by demographic groups can make insights derived from that data more specific and lead to better targeted policies and actions. In addition, Demographically Identifiable Information (DII) can also help address the problem of "data invisibles," in which certain groups are traditionally excluded from data collection and the resulting insights and actions. At the same time, there is growing recognition that DII poses certain risks (Floridi, 2014). The focus on individual data harms—and rights—"obfuscates and exacerbates the risks of data that could put groups of people at risk" (Young, 2020). We need to think of rights and obligations, too, at the group level (Taylor et al., 2017).

A key task facing any attempt to create a responsible data framework is to balance the benefits and threats to group privacy. Attention must be paid to avoid the so-called "mosaic effect" (Office of the Assistant Secretary for Planning and Evaluation (ASPE), 2014). This phenomenon occurs due to the re-identification of data (including anonymized data) by combining multiple datasets containing similar or complementary information. The mosaic effect can pose a threat to both individual and group privacy (e.g., in the case of a small minority demographic group). In addition, groups themselves are frequently established through data analytics and segmentation choices (Mittelstadt, 2016). Due to the algorithmically guided criteria for group sorting, it becomes unclear who can and should advocate for the groups and be empowered to seek redressal. Lastly, individuals are likely unaware their data are being included in the context of a particular group (Radaelli et al., 2018). Decisions made on behalf of this group can limit data holders' control and agency. Mitigation strategies include considering all possible points of intrusion, limiting analysis output details only to what is needed, and releasing aggregated information or graphs rather than granular data. In addition, limited access conditions can also be established to protect datasets that could potentially be combined (Green et al., 2017).

https://doi.org/10.1017/dap.2021.4 Published online by Cambridge University Press

Conclusion

Put together, the above 3 priorities and 10 steps add up to the outlines of a more comprehensive framework for data responsibility in the 21st century. Needless to say, our discussion is preliminary, and much work remains to be done in order to flesh out these concepts, as well as to understand nuances and variations in the way they may be applied. Indeed, we believe that one of the cross-cutting principles of data responsibility must be greater attention to variance. What works in one setting (geographic, cultural, or demographic) may not in another. A more comprehensive framework would drill down into these various requirements, allowing us to better understand how to tailor policies and principles to different contexts.

Funding Statement. Stefaan G. Verhulst has received funding to work on data responsibility by Luminate, UNICEF, the Rockefeller Foundation and FCDO (UK).

Competing Interests. Stefaan G. Verhulst is an Editor-Chief of Data & Policy. This article was accepted after an independent review process.

Data Availability Statement. Data availability is not applicable to this article as no new data were created or analyzed in this study.

Acknowledgments. The author would like to recognize and thank Akask Kapur, Aditi Ramesh, Andrew Young, Andrew Zahuranec, and Uma Kalkar, all at The GovLab (NYU), for their editorial support and review.

References

Auxier B and Lee R (2019) Key takeaways on Americans' views about privacy, surveillance and data-sharing. *Pew Research Center.* Available at https://www.pewresearch.org/fact-tank/2019/11/15/key-takeaways-on-americans-views-about-privacy-surveillance-and-data-sharing/?utm_source=Pew+Research+Center&utm_campaign=ff7e544bad-Internet-Science_2019_11_18&utm_medium=email&utm_term=0_3e953b9b70-ff7e544bad-399447189.

Barkhuus L (2012) *The Mismeasurement of Privacy: Using Contextual Integrity to Reconsider Privacy in HCI.* Mobile Life Centre, Stockholm University. Available at http://barkhu.us/barkhuus-privacy2012.pdf.

Barocas S (n.d.) *Data Mining and the Discourse of Discrimination*. Center for Information Technology Policy. Available at www.cs.yale.edu/homes/jf/Barocas-Taxonomy.pdf.

Berinato S (2018) Why data privacy based on consent is impossible. *Harvard Business Review.* Available at hbr.org/2018/09/stop-thinking-about-consent-it-isnt-possible-and-it-isnt-right.

Cavoukian A, **Taylor S and Abrams ME** (n.d.) Privacy by design: Essential for organizational accountability and strong business practices. Available at https://link.springer.com/article/10.1007/s12394-010-0053-z#citeas.

Dunn M (2016) Feb; Contextualising consent. *Journal of Medical Ethics*. *42*(2): 67–8.

El Arass M and Souiss N (2018) Data lifecycle: From Big Data to SmartData. *IEEE 5th International Congress on Information Science and Technology (CiSt), Marrakech, Morocco, 2018.* Available at https://ieeexplore.ieee.org/abstract/document/8596547.

Floridi L (2014) *Open Data, Data Protection, and Group Privacy*. Springer. Available at https://link.springer.com/article/10.1007/s13347-014-0157-8.

Green B, **Cunningham G**, **Ekblaw A**, **Kominers P**, **Linzer A and Crawford S** (2017) Open data privacy. *Berkman Klein Center for Internet & Society Research.* Available at https://dash.harvard.edu/bitstream/handle/1/30340010/OpenDataPrivacy.pdf.

Kingsmill S., **Cavoukian A** (n.d.) *Privacy by Design: Setting a New Standard for Privacy Certification*. Deloitte & Ryerson University, Toronto, Canada. Available at https://www2.deloitte.com/content/dam/Deloitte/ca/Documents/risk/ca-en-ers-privacy-by-design-brochure.PDF.

Lepri B, **Oliver N**, **Emmanuel L**, **Pentland A and Vinck P** (2018) Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology 31*, 611–627.

Meer D (2015) Overcoming Big Data's challenges. *Strategy + Business.* Available at https://www.strategy-business.com/blog/Overcoming-Big-Datas-Challenges.

Mittelstadt B (2016) *From Individual to Group Privacy in Big Data Analytics*. Springer. Available at https://www.researchgate.net/publication/313599812_From_Individual_to_Group_Privacy_in_Big_Data_Analytics

OECD (2020) *Innovative Citizen Participation and New Democratic Institutions: Catching the Deliberative Wave*. Paris: OECD Publishing. https://doi.org/10.1787/339306da-en.

Office of the Assistant Secretary for Planning and Evaluation (ASPE) (2014) Minimizing disclosure risk in HHS open data initiatives. C. The mosaic effect." *US Department of Health & Human Services.* Available at https://aspe.hhs.gov/report/minimizing-disclosure-risk-hhs-open-data-initiatives/c-mosaic-effect.

Radaelli L, *et al.* (2018) *Quantifying Surveillance in the Networked Age: Node-based Intrusions and Group Privacy.* Cornell University. Available at https://arxiv.org/abs/1803.09007.

https://doi.org/10.1017/dap.2021.4 Published online by Cambridge University Press

Richards NM and Hartzog W (2019) The pathologies of digital consent. *96 Washington University Law Review 1461*. Available at https://ssrn.com/abstract=3370433.

Singh J, *et al.* (2019) Decision provenance: Harnessing data flow for accountable systems. *IEEE Access*, *7*, 6562–6574. https://doi.org/10.1109/access.2018.2887201.

Stodden V (2020) *The Data Science Lifecycle: A Disciplined Approach to Advancing Data Science as a Science*. Association for Computing Machinery Digital Library. Available at https://dl.acm.org/doi/fullHtml/10.1145/3360646

Stoyanovich J, **Howe B**, **Abiteboul S**, **Miklau G**, **Sahuguet A**, **Weikum G** (2017) Fides: Towards a platform for responsible data science. *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 1 June 2017. Available at dl.acm.org/doi/10.1145/3085504.3085530.

Susha I, **Janssen M and Verhulst S** (2017) "Data collaboratives as "bazaars"? A review of coordination problems and mechanisms to match demand for data with supply", transforming government. *People, Process and Policy*, *11*(1), 157–172. https://doi.org/10.1108/TG-01-2017-0007.

Szymielewicz K (2019) Your digital identity has three layers, and you can only protect one of them. *Quartz.* Available at qz.com/1525661/your-digital-identity-has-three-layers-and-you-can-only-protect-one-of-them/.

Taylor L, **Floridi L**, **van der Sloot B** (eds) (2017). *Group Privacy: New Challenges of Data Technologies*. Dordrecht: Springer. Available at https://www.stiftung-nv.de/sites/default/files/group-privacy-2017-authors-draft-manuscript.pdf.

Verhulst S (2018) Information asymmetries, blockchain technologies, and social change. *Medium.* Available at https://sverhulst.medium.com/information-asymmetries-blockchain-technologies-and-social-change-148459b5ab1a.

Verhulst S (2019) Solving a problem starts by asking the right question. *Apolitical.* Available at apolitical.co/en/solution_article/raw-data-wont-solve-our-problems-asking-the-right-questions-will.

Verhulst SG, **Young A**, **Winowatan M and Zahuranec AJ** (2019) Leveraging private data for public good: a descriptive analysis and typology of existing practices. *The Governance Lab.* Available at https://datacollaboratives.org/static/files/existing-practices-report.pdf.

Viljoen S (2020) Data as property? *Phenomenal World.* Available at https://phenomenalworld.org/analysis/data-as-property.

Wetsman N (2020). Personal privacy matters during a pandemic—but less than it might at other times. *The Verge.* Available at https://www.theverge.com/2020/3/12/21177129/personal-privacy-pandemic-ethics-public-health-coronavirus.

Young A (2020) *Responsible Group Data for Children*. UNICEF. Available at www.unicef.org/globalinsight/media/1251/file/UNICEF-Global-Insight-DataGov-group-data-issue-brief-2020.pdf.

Young A, **Verhulst SG** (2020) Data collaboratives. In: Harris P, Bitonti A, Fleisher C, Skorkjær BA (eds) *The Palgrave Encyclopedia of Interest Groups, Lobbying and Public Affairs*. Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-030-13895-0_92-1.

Young A, **Verhulst, SG**, **Safonova N and Zahuranec AJ** (2020) The data assembly: Responsible data re-use framework. *The Governance Lab.* Available at https://thedataassembly.org/files/nyc-data-assembly-report.pdf.

Cite this article: Verhulst S. G (2021). Reimagining data responsibility: 10 new approaches toward a culture of trust in re-using data to address critical public needs. *Data & Policy*, 3: e6. doi:10.1017/dap.2021.4

https://doi.org/10.1017/dap.2021.4 Published online by Cambridge University Press