Information Research - Vol. 31 No. *iConf* (2026)

Framing data inequality as an information justice lens: Access, representation, and control in algorithmic decision-making

Shiwei Jia, Hui Yan, and Jia Tina Du

DOI: https://doi.org/10.47989/ir31iConf64175

Abstract

Introduction. This study investigates how data inequality manifests within algorithmic systems and explores how an information justice framework can be applied to reframe these inequalities.

Method. In-depth interviews were conducted with 36 participants representing three distinct roles — algorithm design-related users, algorithm-for-work users, and algorithm-for-daily-life users — within the Chinese internet context. The qualitative data gathered capture their perceptions and lived experiences of data inequality in algorithmic systems.

Analysis. Thematic analysis was employed to identify and categorize forms of data inequality. These categories were then interpreted through the lens of information justice using the data, information, knowledge, wisdom (DIKW) model to illuminate their moral implications.

Results. The study provides empirical evidence for three forms of data inequality — access inequality, representation inequality, and control inequality. Each of these undermines information justice in distinct ways, particularly with respect to iParticipatory justice and iRecognitional justice, manifesting as exclusion from participation, distortion of identity, or erosion of data sovereignty.

Conclusion. Data inequality represents a systematic violation of information justice. Reframing data inequality as an information justice lens not only advances research on digital inequality but also deepens theoretical discussions at the intersection of data ethics, algorithm studies, and information science.

Introduction

Data, especially big data, has long been recognized as the foundational material for the development of intelligent algorithms, particularly since the advent of the big data era. Similar to the myth of technology neutrality, a common misconception about big data is the belief that it is objective, fair, and non-ideological (Cohen, 2013; Van Dijck, 2014). However, a growing body of research has challenged this assumption, documenting the widespread prevalence of data inequality. For example, Wang et al. (2021) showed how smart city development produced data inequality when geo-communities were disproportionately represented due to the uneven distribution of sensors, resulting in structural disadvantages for marginalized groups. Similarly, Cinnamon (2020) emphasized that the absence or underrepresentation of data is often more common than the popular notion of '*ubiquitous big data*' suggests. Algorithmic systems built on biased or imbalanced datasets often reproduce—and even amplify—existing inequalities, a phenomenon commonly captured by the phrase: '*bias in, bias out*.'

Although some attention has been given to fairness at the data stage, research on algorithmic fairness remains largely concentrated on model design and decision-making, leaving data-level inequities under-theorized and under-documented (Li et al., 2022). In this study, we adopt Cinnamon (2020)'s categorization of *data inequality*, defining it as the unequal distribution of data resources in terms of access to data, representation of the world as data, and control over data flows. Data inequality is deeply embedded in socio-technical infrastructures (Park et., 2020; Pfeffer & Verrest, 2016), shaping who and what becomes visible in datasets, as well as who remains invisible (Heeks & Shekhar, 2019). Consequently, data inequality foregrounds questions of justice in information systems.

Justice is closely associated with positive concepts such as fairness, equality, equity, and inclusion. These values are central to the professional mission of librarianship, which seeks to ensure that all individuals, especially those from marginalized groups, have equitable access to information. (Fallis, 2007). Although many data-related issues are not unique to information justice, the framework of information justice can provide valuable guidance for the ethical evaluation of data practices. (Johnson, 2014). Mathiesen (2015) drew on social justice theory within philosophy to develop a framework for Library and Information Science (LIS) called *information justice*, which she defined as '*the just treatment of persons as seekers, sources, and subjects of information*.' This definition emphasizes three dimensions: *(a) iDistributive justice* — ensuring seekers equitable access to information; *(b) iParticipatory justice* — enabling individuals as sources to contribute perspectives and participate in decision-making about information resources; and *(c) iRecognitional justice* — ensuring individuals as subject fairly and accurately representation in information environment. From this perspective, *iDistributive justice* is primarily concerned with information rather than data, while *iParticipatory* and *iRecognitional justice* directly address the challenges posed by data inequality.

As such, we analyze data inequality through the lens of information justice, exploring its manifestations and implications within algorithmic decision-making systems. To structure this analysis, we employ the Data, Information, Knowledge, Wisdom (DIKW) model from information science (Zeleny, 1987; Ackoff, 1989). DIKW model is a conceptual model that offers a hierarchical transformation from data to wisdom, making it suitable for locating data inequality within the broader information landscape in the age of AI (Peters et al., 2024). Specifically, we address the following research questions: 1) How does data inequality manifest in algorithmic decision-making systems? 2) How can an information justice framework reframe data inequality in these systems?

By framing data inequality through the lens of information justice, this study offers both a theoretical framework and empirical evidence that advance LIS scholarship on equitable access to, representation in, and governance of data.

Research design

Identifying data inequality is challenging because data is often hidden, invisible, or embedded in socio-technical infrastructures. To better capture its manifestations in algorithmic systems, this study adopts a stakeholder lens, moving beyond a sole focus on algorithmic system users. Specifically, we categorized stakeholders into three distinct roles — sources, subjects, and users — each representing different forms of engagement with data:

(a) Algorithm design-related users refer to individuals who are directly involved in the design and development of algorithmic systems, such as data collector, data annotator, data scientist, and algorithm engineer.

(b) Algorithm-for-work users refer to individuals who utilize algorithmic system for non-algorithm design related work tasks rather than for personal decision-making, for example, employing algorithms for advertising distribution. Examples of this role include recruitment specialists, advertising managers, and government officials.

(c) Algorithm-for-daily-life users refer to individuals who utilize algorithmic system only for personal purposes (e.g., leisure), such as content consumers and non-profit content producers, whose experience of using the algorithmic systems is integrated into their daily lives.

Although these three algorithmic roles are analytically distinct, in practice individuals may occupy multiple roles depending on the context. An individual's algorithmic role can also shift across different scenarios. For example, Alex (he/him), an algorithm engineer at Company A, acts as an algorithm design-related user while programming, as an algorithm-for-work user when using Company B's system for non-algorithm design related work tasks, and as an algorithm-for-daily-life user when using YouTube for leisure. Consequently, we developed distinct interview outlines based on the characteristics of the roles. Finally, using the finalized formal outlines, we conducted in-depth interviews with 36 participants (12 participants per role) representing various roles, exploring their perceptions and experiences of data inequality in algorithmic systems. Participants were recruited in China through snowball sampling and online recruitment. Table 1 shows the characteristics of participants.

Roles	Characteristics		Count
Algorithm design-related users	Gender	Man	6
		Woman	6
	Age	18-29	12
		30-44	0
		Over 45	0
	Occupation	Algorithm engineer	8
		Data annotator	1
		Data scientist	3
Algorithm-for-work users	Gender	Man	8
		Woman	4
	Age	18-29	8
		30-44	2
		Over 45	2
	Occupation	Government official	6
		Product manager	5
		Recruitment specialist	1
Algorithm-for-daily-life users	Gender	Man	8
		Woman	4
	Age	18-29	10
		30-44	2
		Over 45	0
	Occupation	College student	8
		Information professionals	3
		Salesperson	1

Table 1. Basic characteristics of participants.

All interviews were recorded and transcribed after obtaining informed consent from each participant. Participants were anonymized throughout the study to ensure confidentiality and privacy. Algorithm design-related users were labelled AD1-AD12, algorithm-for-work users AW1-AW12 and algorithm-for-daily-life users AU1-AU12. Then, we adopted thematic analysis (Braun & Clarke, 2006) to code the data. A coding frame was developed combining deductive and inductive approaches. At the first step, concept-driven codes were deductively derived from the definition of data inequality: (a) data access inequality; (b) data representation inequality; and (c) data control inequality. At the second step, data-driven codes were added inductively as new sub-themes emerged during the close reading of the transcripts. To enhance reliability and validity, two researchers collaborated to code a subset of the data. Any coding inconsistencies that arose during this process were resolved through discussion.

Findings

The coding results confirmed the presence of all three forms of data inequality — access, representation, and control — in our interview data, each of which can be further subdivided into specific categories. These findings provide exploratory evidence of data inequality in the Chinese internet context.

Data access inequality

Data access inequality, which emerges during the stage of data collection, has been a long-standing issue for algorithms-owning organizations despite the data revolution and open data initiatives (Johnson, 2014). Prior studies have conceptualized data access in binary terms, such as having and not having data (Cinnamon, 2020), or the '*Big Data rich*' and the '*Big Data poor*' (Boyd & Crawford, 2012). However, evidence from our interviews reveals that data access also extend to differences in data quality. Therefore, we define data access inequality in this study as a continuous inequality spectrum ranging from complete absence to complete access.

Complete absence of data

The left end of the spectrum represents complete absence, i.e., the *'no data'* state, encompassing two main scenarios.

The first scenario is the absence of data itself, often occurs when lack of digital infrastructures. An example of this is AW3's village, which remained invisible in smart city datasets due to the absence of data-sensing devices until 2020.

> AW3: Our village only got a cement road in 2015, and internet access followed shortly after. Now the whole place has coverage. The county government has also implemented a digital village project to assist with public safety management in rural areas by increasing and updating digital infrastructures. We also developed a county-level governmental big data center that connects all villages.

The second scenario is restricted data inaccessibility, caused by technical barriers, platform censorship policies, etc. For example, AW6: '*Some companies put in anti-scraping mechanisms, and they set the bar super high, making it really hard to crawl their data.*'

Uneven quality of data

The middle of the spectrum represents another form of data access inequality, namely, uneven data quality, which encompasses two main scenarios: incomplete data and inaccurate data.

The former scenario refers to the incompleteness of variable content, a common issue reported by most algorithm design-related users regarding the datasets they utilized. For example, AD4, a data annotator in the ByteDance, explained that recommendation systems often require dozens of variables, such as daily active users, time on page, and other variables from external sources. However, as previously noted in the discussion of complete absence of data, these variables are not always fully available. AD4: '*One challenging thing of data annotation is that it is not always possible to access data for each variable. Variable incompleteness is very common.*'

The second scenario involves various forms of inaccuracies, including those present in cross-domains datasets, and historical data, as reported by participants.

The right end of the spectrum represents complete data access - an ideal state that is rarely attainable in real-world scenarios.

Data representation inequality

Data representation inequality refers to the uneven representation during the transformation of the world into data. It focuses on whether a dataset can represent the characteristics and conditions of its target population objectively and fairly. We identified two forms of data representation inequality: one rooted in technical limitations, and the other in social structures.

Technical data representation inequality

Technical data representation inequality arises when certain groups or characteristics are systematically ignored, overrepresented, or mischaracterized due to sampling bias during data collection. This form of inequality is statistical in nature, as illustrated by AD2. Specifically, hiring college students to distribute and collect questionnaires may limit the dataset's ability to reflect the preferences and evaluations of the broader population for a particular product, resulting in the dataset disproportionately reflecting the views of younger and more educated populations. Although this sampling might be unintentional, it does narrow the range of data representation. Similar issues arise in other domains, such as healthcare. AD1: '*For example, if the training data is mostly collected from men or from women, then the model's predictions could reflect gender bias.*'

Societal data representation inequality

Societal data representation inequality occurs when data collection and processing practices are influenced by social inequality, including three scenarios.

The first is labeling bias or mislabelling due to embedded cognitive bias in the data labeling stage. This scenario is particularly difficult to detect, as data annotators are often unaware of how their own perspectives shape the data they produce.

The second is association bias within datasets. As AD3 observed, stereotypes such as '*men are more suited for medical professions*' were embedded in several training datasets, leading to more serious distortions in representation.

The third is interaction bias, which emerges when users introduce biased content during their interactions with algorithmic systems. Once embedded into datasets, such inputs distort the representation of social groups and reinforce existing imbalances. As AE6 noted, biased prompts to large language models can result in outputs that reproduce or even amplify these inequalities.

Data control inequality

Data control inequality refers to unequal power dynamic between algorithm owners and customers regarding data ownership and control. Specifically, who can access to the data, where and when it is used, for what purposes it is used (Cinnamon, 2020). More than half of the participants reported relevant experiences in their daily-life usage. Two main scenarios emerged:

Unequal power dynamics in data collection

This form of inequality is characterized by opaque and compulsory practices. Participants described how their personal data was unknowing or involuntary gathered by algorithmic systems. 'Agree it or leave it' consent mechanisms legitimized such practices, depriving users of meaningful autonomy and reinforcing power asymmetries between platforms and users.

> AU4: It feels like the app is forcing me. In the user agreement, it always says the app requires access to data and permissions like contacts and recoding. They claim it can improve the recommendation experience, but if I don't agree to the authorization, I can't use this app at all.

Unequal power dynamics in data circulation

This form of inequality is manifested as data surveillance. Participants reported that certain apps (e.g., Taobao, Xiaohongshu) engaged in internal and cross-platform data monitoring. This surveillance occurred through data sharing, trading, or technical integrations, further consolidating control in the hands of algorithm-owning organizations.

Discussion

The findings reveal that the three forms of data inequality can be reinterpreted as systematic challenges to information justice in algorithmic systems. Figure 1 reframes data inequality through an information justice lens across DIKW model. This reframing shifts the discourse from technical issues to moral imperatives, focusing on the moral aspects to consider individuals as both sources and subjects of information.

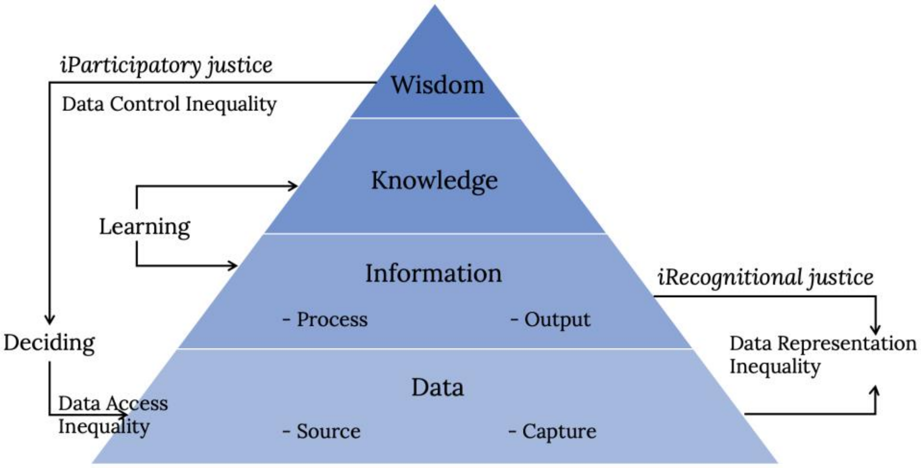

Figure 1. Reframing digital inequality as information justice lens across DIKW model.

When individuals are regarded as sources of information, corresponding to iParticipatory justice, the issues of data access and control emerge as fundamental justice concerns. Specifically, algorithm designers and owners have autonomy over which data are collected, who collects them, how they are processed, and which models are employed. In this process, individuals often become passive data points, subjected to rules of data collection and circulation that they neither control nor consent to. This dynamic effectively deprives individuals of their agency and ownership, resulting in data control inequality, where power over informational resources is concentrated in the hands of organizations that own the algorithmic systems. Beyond control, data access inequality further shapes the informational landscape by restricting who can contribute to or benefit from algorithmic systems. In contexts where data is absent, incomplete, or inaccessible, certain groups are systematically excluded from algorithmic systems, effectively denying them participation in a data-driven society. These forms of inequality represent systematic violations of participatory justice, not only constructing the hierarchy of information, but also transforming what might appear as technical limitations into deeply normative concerns.

When individuals are regarded as subjects of information, corresponding to iRecognitional justice, the central concern shifts to data representation inequality. In the process of translating the world into data, the identities of certain groups might be erased or distorted through biased collection and labelling practices, thereby reproducing the existing the social hierarchy. Within the DIKW model, representation inequality manifests at both the data and information layers, systematically denying individuals just treatment as informational subjects. Marginalized groups are particularly vulnerable, as misrepresentation in algorithmic systems perpetuates symbolic exclusion.

Conclusion

In this study, we investigated data inequality through in-depth interviews with 36 participants representing various roles within the Chinese internet context. The findings provided empirical evidence for three distinct forms of data inequality — access, representation, and control — in this context. By reframing these inequalities through the lens of information justice, this study demonstrated how data inequality undermines both iParticipatory and iRecognitional justice. This reframing extends existing research on digital inequality and enriches theoretical dialogue between data ethics, algorithm studies, and information justice. It further calls for governance approaches that address not only the distribution of digital infrastructures (access) but also the allocation of power and rights (control) and the legitimacy of symbolic representation (representation). A key limitation of this study is that, as a short paper, it necessarily prioritizes specific manifestations of data inequality and qualitative insights, and therefore does not provide normative design principles for algorithmic systems. In addition, the analysis focuses primarily on the data experiences at the organizational and individual levels, leaving unexamined the role of the state in enforcing informational justice through legal frameworks.

Acknowledgements

This research has been made possible through the financial support of the National Social Science Foundation of China under Grant No. 23AZD093.

About the authors

Shiwei Jia is a lecturer in School of Smart Governance, Renmin University of China, and a research fellow in Institute for AI Governance. She received Ph.D. from Renmin University of China, and her research interests include social impact of digital technologies, digital inequality, and algorithmic governance. She can be contacted at jiashiwei422@ruc.edu.cn

Hui Yan is a Professor in School of Information Resource Management, Renmin University of China, also a research fellow in Institute for AI Governance. His research interests include social impact of digital technologies, community informatics and digital inequality. He can be contacted at hyanpku@ruc.edu.cn

Jia Tina Du is a professor in the School of Information and Communication Studies, Charles Sturt University, Australia. Her research interests include interactive information seeking, social impact of digital technologies and data governance. She can be contacted at tdu@csu.edu.cn

References

Ackoff, R. L. (1989). From Data to Wisdom. Journal of Applied Systems Analysis, 16(1), 3–9. https://www.scribd.com/document/518769223/Ackoff-Russel-L-From-Data-to-Wisdom

Boyd, D., & Crawford, K. (2012). Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Information Communication & Society, 15(5), 662–679. https://doi.org/10.1080/1369118x.2012.678878

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cinnamon, J. (2020). Data inequalities and why they matter for development. Information Technology for Development, 26(2), 214-233. https://doi.org/10.1080/02681102.2019.1650244

Cohen, J. E. (2013). What Privacy is For. Harvard Law Review, 126(7), 1904–19033. https://dialnet.unirioja.es/servlet/articulo?codigo=4832440

Fallis, D. (2007). Information ethics for twenty-first century library professionals. Library Hi Tech, 25(1), 23–36. https://doi.org/10.1108/07378830710735830

Heeks, R., & Shekhar, S. (2019). Datafication, development and marginalised urban communities: An applied data justice framework. Information, Communication & Society, 22(7), 992-1011. https://doi.org/10.1080/1369118X.2019.1599039

Johnson, J. A. (2014). From open data to information justice. Ethics and Information Technology, 16(4), 263–274. https://doi.org/10.1007/s10676-014-9351-8

Li, N., Goel, N., & Ash, E. (2022, July). Data-centric factors in algorithmic fairness. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 396-410). https://doi.org/10.1145/3514094.3534147

Mathiesen, K. (2015). Informational justice: A conceptual framework for social justice in library and information services. Library trends, 64(2), 198-225. https://doi.org/10.1353/lib.2015.0044

Park, B., Rao, D. L., & Gudivada, V. N. (2020). Dangers of Bias in Data-Intensive Information Systems. In Advances in intelligent systems and computing (pp. 259–271). https://doi.org/10.1007/978-981-15-4851-2_28

Peters, M. A., Jandrić, P., & Green, B. J. (2024). The DIKW model in the age of artificial intelligence. Postdigital Science and Education. https://doi.org/10.1007/s42438-024-00462-8

Pfeffer, K., & Verrest, H. (2016). Perspectives on the role of Geo-Technologies for Addressing contemporary Urban Issues: Implications for IDS. European Journal of Development Research, 28(2), 154–166. https://doi.org/10.1057/ejdr.2016.4

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. Surveillance & Society, 12(2), 197–208. https://doi.org/10.24908/ss.v12i2.4776

Wang, G., Pan, S., & Xu, S. (2021). Decoupling the unfairness propagation chain in crowd sensing and learning systems for spatio-temporal urban monitoring. Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 200-203. https://doi.org/10.1145/3486611.3486669

Zeleny, M. (1987). Management support systems: Towards integrated knowledge management. Human Systems Management, 7(1), 59–70. https://doi.org/10.3233/hsm-1987-7108