Social Science Computer Review
2025, Vol. 0(0) 1–15
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08944393251365277
journals.sagepub.com/home/ssc
Sage

Ghosts in the Data: The Contested Politics of Absence in Data Infrastructures

Will Orr[1]

Abstract

Absences are inescapable in data. Data collection always focuses on some elements while occluding others. Yet, how absences are considered and recorded within data infrastructures markedly transforms the inferences that can be made. Tracing a genealogy from early databases to contemporary AI datasets, this paper explores how data infrastructures have grappled with the inherent incompleteness of data. Specifically, I uncover a tension between a desire for certainty and acknowledging partiality at the foundation of data science that continues to pervade contemporary AI datasets. Drawing on archival studies and sociological perspectives, I argue that data science must embrace uncertainty by recognizing the "ghosts in the data"—the uncounted, the unrepresented, and the silenced—and how their absence shapes the outcomes of automated systems.

Keywords

Absence, Archives, Artificial Intelligence, Databases, Datasets, SQL

Introduction

In the final weeks of 2020, as the COVID-19 pandemic was well underway, the Los Angeles County Department of Public Health reported that a total of 71 unhoused people had died from COVID-19 in California (Fowle & Gray, 2022). Predicted initially as a "time bomb" within the unhoused community, the coronavirus pandemic contributed to significantly fewer recorded deaths amongst this cohort than expected (Holland, 2020). This surprisingly low casualty rate led the LA Times to wonder whether the local unhoused community had "dodged a COVID-19 catastrophe" (Holland, 2020). Public officials and academics alike grappled with this unexpected outcome, proposing various explanations ranging from infrequent use of indoor spaces and natural social distancing practices to potential biological factors, such as increased vitamin D levels (Holland, 2020).

[1]University of Southern California, USA

Corresponding Author:
Will Orr, Annenberg School of Communication, University of Southern California, 3502 Watt Way, Los Angeles, CA 90089, USA.
Email: orrw@usc.edu

However, an alternative explanation is also possible: there is insufficient data regarding the prevalence of COVID-19 within unhoused communities or deaths relating to the illness. The cause of death of homeless people is rarely recorded. In the absence of evidence of assault or drug abuse, deceased unhoused people do not receive autopsies; their cause of death remains unrecorded and unknown (McFarling, 2021). Moreover, in each month of 2020, there were 40–60% more recorded deaths of unhoused people in Los Angeles compared to previous years (Fowle & Gray, 2022). Of the 1000 deaths between January and June 2020, only 36 were attributed to COVID-19 (McFarling, 2021). Consequently, it is possible that the COVID-19 "time bomb" did go off within unhoused communities, but its effects went largely unnoticed and unaddressed (McFarling, 2021). This outlook extends upon the invisibility often experienced by unhoused communities, which are often overlooked by the public, and their needs neglected by government officials (Banerjee & Bhattacharya, 2021).

The treatment of absent data in these two perspectives highlights a central tension that has shaped the development of data science. The LA Times account of this story seemingly takes the data at face value and treats them as a veritable representation of reality. Here, the absence of data did not mean that information or perspectives were missing but rather that these accounts did not exist. The alternative perspective treats the absence of data as a source of explanatory uncertainty, highlighting the possibility of alternative and unaccounted-for explanations. The tension between these explanations pervades the history of data science. I therefore ask: How has data science reckoned with the inherent incompleteness of data? What claims of truth can be made from a medium that will always be partial?

This paper traces a genealogy of absence within the field of data science. Beginning with an examination of archival literature, it uncovers the power dynamics embedded in the creation of legitimized knowledge and the political practices of revealing, concealing, and silencing within historical narratives. Particularly, I draw on Avery Gordon's work (1997) to underscore how sociology can benefit from engaging with traces or "ghosts" of social phenomena that are not readily visible but nonetheless shape social realities. The paper then examines three pivotal forms of data infrastructure that have wrestled with the challenges of absent and missing data. Firstly, by exploring early iterations of databases, I highlight the prevailing assumption that data are complete and finite. Next, I demonstrate how these assumptions were destabilized with the relational database model, specifically the invention and eventual dominance of the database management system, Structured Query Language (SQL). By designing a visual representation for the absence of information, SQL enabled the possibility of recording data as flawed, partial, or unknown. Finally, I examine how contemporary AI datasets foreclose the possibility of absences by imposing technological and institutionalized data standards while claiming comprehensive scale and representativeness. Through this historical and infrastructural analysis, I underscore the contingent nature of truth in data analytics, demonstrating how it hinges on underlying assumptions about what is and, crucially, what is not present. Taking a sociological lens to AI entails moving beyond what is "visible" to consider and value the "ghosts" that may be missing or erased from automated decision-making systems but nonetheless constitute their existence.

Archives, Absences, and Infrastructures

To frame the construction of absences in data, we can look towards the rich critical literature that interrogates archives and record-keeping as sites of power. Sociologist Avery Gordon (1997) underscores this preoccupation with visible, empirical evidence within social sciences. The production of social scientific knowledge is predicated on archival utterances as evidence of particular accounts and perspectives. Physical evidence, such as photographs, state records, and recorded transcripts, is taken as veritable facts to bolster the perceived authoritativeness of "social sciences" (Gordon, 1997, p. 21). Archives thus enact "the law of what can be said" (Foucault, 1972, p. 129), delineating the boundaries of what is knowable and commanding what constitutes an authoritative construction of the past and present (Derrida, 1995, p. 9; Trouillot, 2011, p. 52).

Yet, privileging the visible within the archive also creates silences and absences of what is not recorded. As Gordon (1997, p. 15) explains, *"Visibility is a complex system of permission and prohibition, of presence and absence, punctuated alternately by apparitions and hysterical blindness*". No source can capture every detail, nor can any archive contain every source or perspective. Moreover, the very process of creating cohesive narratives necessitates valorizing specific sources and perspectives while silencing others (Trouillot, 2011). The archive operates by being "invisibly exclusionary" (Bowker, 2008, p. 24), selectively remembering a particular set of facts, discoveries, or observations while presenting itself as "the set of all possible statements" (Bowker, 2008, p. 24), and a "totalizing assemblage" (Derrida, 1995, p. 50). Narratives from the archive are thus constructed as a "particular bundle of silences" where presence and absence are "active, dialectical counterparts" (Trouillot, 2011, pp. 27–48).

Importantly, archival absences are not neutral or natural (Trouillot, 2011, p. 48). Rather, "silencing" is an active and intentional process (Trouillot, 2011). Occlusions in archives are not oversights but are integral to their geopolitical formations and epistemic frameworks (Stoler, 2016, p. 10). Upholding institutional standards that require tangible archival sources to produce social scientific knowledge silences the perspectives of those who cannot be readily represented. For instance, the archives of slave trading companies through West Africa, containing records such as ledgers of goods traded, consumed food items, and lists of enslaved bodies alive and dead, fail to capture the violence and dispossession of those they represent (Hartman, 2007, p. 17). As Hartman (2007, p. 17) illustrates, "Commodities, cargo, and things don't lend themselves to representation, at least not easily." By defining the social through what is "seen," social sciences are blinded to the fraught histories of loss, dispossession, and disappearance, such as slavery, racism, and capitalism (Gordon, 1997; Hartman, 2007). Perspectives absent from the archive remain unheard and unrepresented (Gordon, 1997).

Despite their invisibility, the absent, erased, and unseen still haunt our everyday social realities as a "seething presence" upon taken-for-granted social realities (Gordon, 1997, p. 8). "Ghosts" emerge as signifiers or suggestions of what might be missing, traces of what is supposed to be invisible or absent but are nonetheless central to social formation. For example, systemic racism manifesting in persistent inequalities and injustices reminds us of the "lingering inheritance of U.S. racial slavery" (Gordon, 1997, p. 27). Gordon (1997) thus calls for a new way of understanding the social world that takes ghosts seriously as "empirical evidence" of hauntings that are otherwise unseen (p. 8). This rich literature grappling with the partiality of archives thus underscores the importance of making sense of the absences that complicate dominant historical narratives, to examine where and how they emerge in the production of knowledge, and to whose benefit.

Indeed, sociology, too, has concerned itself with studying the negative space of social life, encompassing both things that do not exist due to conscious disengagement or deletion and those that are omitted through passive neglect (Scott, 2018). In her exploration of subaltern epistemologies erased through the dominance of Western rationality, Santos (2015) proposes a transgressive "sociology of absence" that aims to take seriously and make present the "disqualified parts of homogeneous totalities" (p. 174). In the realm of data, Artist Mimi Onuoha (2018) offers a material example of this form of thinking in The Library of Missing Datasets. Phrases such as "Undocumented immigrants currently incarcerated and/or underpaid" and "How much Spotify pays each of its artists per play of song" are compiled as an incomplete list of datasets that should exist but do not due to "quiet complications" and embedded power dynamics inherent in data collection (Onuoha, 2018). By resisting positivist reduction of reality to the extant and analyzable, a sociology of absence can identify the scope of “subtraction and contraction” (Santos, 2015, p. 174), consider counterfactuals to hegemonic experience, and confront the power dynamics that render certain phenomena invisible.

To complicate this further, scholars have also detailed the political potential of absence as an intentional agentic activity. Indeed, attempts to incorporate marginal perspectives into existing data systems may enact “discursive violence” that reinforces, rather than subverts, dominant power structures by normalizing and diffusing radical potential for social change (Hoffmann, 2021, p. 2). In this way, “informed refusal” from regimes of technological inclusion may present a form of resistance that affirms the agency and self-determination of marginalized people in the face of technological dominance (Gangadharan, 2021, p. 25). For Indigenous communities, public domain and open-access information commons can perpetuate colonial legacies in digital spaces, where outsiders profit from knowledge extraction without reciprocal benefits for the community (Christen, 2012). In response, indigenous communities in Latin America and Australia have chosen to exclude themselves from mainstream technological systems, such as data management and mapping systems, instead adopting self-governance frameworks that center their own knowledge systems and cultural relations (Christen, 2012). As such, absence can be an intentional and active form of democratic participation that confronts dominant narratives beyond visible expression (Ananny, 2020).

Science and technology studies also point towards the power embedded within infrastructures that shape how information is remembered. As Dourish (2022, p. 6) draws attention to, the specific “materialities of information… constrain, enable, limit, and shape the ways in which those representations can be created, transmitted, stored, manipulated, and put to use.” The methodologies, systems, and processes used to collect, store, and organize information, including infrastructures, classification schemes, and the criteria for selection and preservation, profoundly shape how the past is transmitted into the future (Bowker, 2008). For instance, Bowker (2008) highlights how archives operate through the categorization and taxonomy of information that distills the essence of actual events into a system of classifications. Through this lens, archives do not preserve facts *per se* but rather the frameworks that allow for these facts to be reconstructed from disaggregated classifications (Bowker, 2008, p. 18). This process facilitates a selective memory of the past, enabling the forgetting of specific details through consolidation into a classification system. Yet, there will always be edge cases that do not fit neatly within the categories provided, thus erasing important information through classification processes (Bowker & Star, 1999; D’Ignazio & Klein, 2020). As such, the “archivization produces as much as it records the event” (Derrida, 1995, p. 17), reifying boundaries between constructed categories and erasing experiences that do not conform to these rigid definitions.

Finally, STS scholars have also underscored how science and technology are not inherently self-evident. While technologies may seem “stable” or “closed,” interrogating the controversies present within technology design underscores alternative interpretations about the potential functionalities and promises of artifacts (Davis, 2020). In their germinal analysis of the evolution of bicycle design, Pinch and Bijker (1984) highlight how different interpretations of a technology’s purpose can lead to conflicting ideas about its structure and design. Technological artifacts are open to multiple meanings and uses, shaped by the social contexts and groups involved in their development (Davis, 2020). Other scholars emphasize how broader social tensions, power dynamics, and cultural norms shape technology design and deployment (e.g., Winner, 1993). These considerations are motivated by collective hopes, fears, and expectations regarding the possibilities and implications of emerging technologies (Jasanoff & Kim, 2015). Tracing controversies, therefore, is a task of excavating the assumptions and broader imaginaries underpinning the development and adoption of technologies.

In the following pages, I will examine historical technical documents to reveal a central controversy regarding the construction of absences within data infrastructures. I will show how this controversy reflects broader divisions regarding the purpose and utility of data-driven technologies.

The Origins of Databases

Like many technologies, databases originated for specific business and military purposes. Haigh (2006) traces the desire for databases back to the concept of Management Information Systems (MIS) in the late 1950s, several years before the term "database" was used. As the volume of files grew in corporations, executives hoped for a central repository of information that could provide instant access to pertinent records about their organizations. Referred to as "data hubs," "data pools," and "data banks," the ideal of MIS lays in their ability to smoothly aggregate interrelated company information of disparate data types into a single "ocean of knowledge" (Haigh, 2006, p. 34). As Milton D. Stone (1960, p. 17) explained at the American Management Association conference on novel computer technologies for corporate administration, MIS represent "a veritable 'bucket of facts'" into which all "information seeking ladles of various sizes and shapes are thrust in different locations" (cited in Haigh, 2006). Here, Stone (1960) illustrates the slippage between "data" and "facts" within databases. By removing the contextual information of data, such as the structure of its origin (e.g., administrative form, sales report, and letter of recommendation), data are presented as a pure and inert substance that can be stored, refined, and extracted as needed.

In this way, databases rely on "the loss of the thing itself," both physically, as an entity becomes understood entirely through numerical representation, and epistemologically, as the elimination of contextual information is necessary for data to be comparable and insights generated (McPherson, 2018, p. 79). Importantly, however, metadata are retained, such as schemas, data types, and dependency rules, that explain how the database is structured and provide a framework for data retrieval and manipulation. Yet, treating these data *as* fact risks reifying them as a veritable abstraction, stripping away their contextual significance and historical embeddedness (Gitelman, 2013). As Rosenberg (2013, p. 18) reminds us, "the semantic function of data is *specifically* rhetorical."

In practice, however, the hopes of MIS were undermined by the technological capabilities of their times. MIS were constrained by small computational memories, tape drives, and inflexible file structures. Users were required to scroll through tapes to find a particular record, which was laborious and inefficient. The invention of hard disk drives in the early 1960s allowed information to be accessed randomly, eliminating the sequential dependence of previous memory devices. In 1963, IBM released the Integrated Data Store (IDS) with the goal of creating a flexible, integrated system that could be used across various business contexts to provide rapid results to queries (Haigh, 2006). IDS became one of the most effective and influential database management systems, prominently used in manufacturing, coordination, and the Apollo space program. Instead of storing records sequentially, IDS represented data through a network, allowing the complex relationships between records and groups to be defined. However, as records were linked in a network structure, they could not be easily queried: users were required to navigate from one record to another to access a specific record. This navigational structure imposed incredible dependence on the system. For instance, removing one record could render paths broken, cause information to be lost, or make it inaccessible.

Navigational databases are grounded in the assumption that all data observed within a database are true (Razniewski & Nutt, 2014). Everything that cannot be inferred from the data as true is considered false. There is no middle ground (Reiter, 1978). For instance, in Figure 1, it is assumed that the supplier Ajax does not sell screws, rather than the price of screws being unknown entirely.

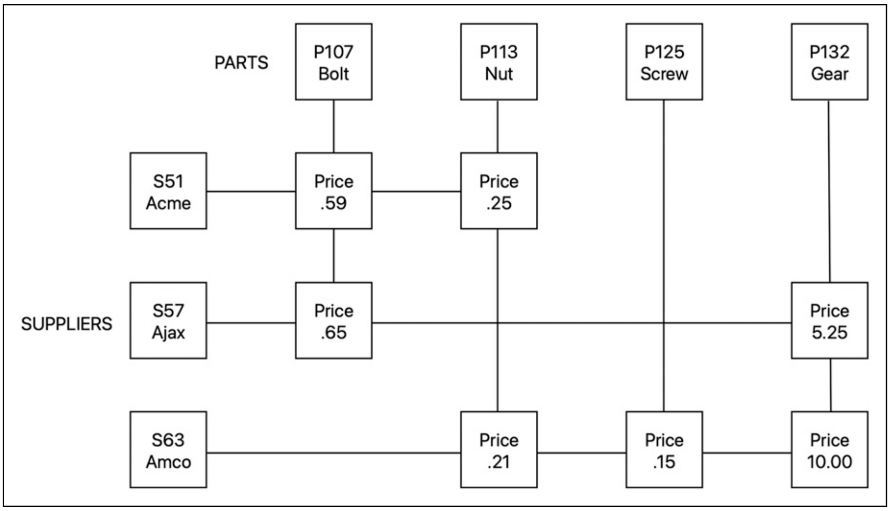

Figure 1. A "Navigational" Database (reproduced from Chamberlin et al., 1981, p. 632).

This logical formulation, known as the closed-world assumption, claims that data represent a verifiable model of reality. If databases represent "a veritable 'bucket of facts'" (Stone, 1960), it is the presence of data that reifies observations as truth. Like archives, digital databases are perceived as a "totalizing assemblage" (Derrida, 1995, p. 50) of all possible and pertinent observations.

Indeed, data analytics traditionally relies on the closed-world assumption to construct certainty about the inferences made from data. In the case of COVID-19 in the Los Angeles unhoused community, the absence of positive coronavirus diagnoses in unhoused people indicated the lack of virus transmission and their overall survival (Holland, 2020). The possibilities of truth about the experiences of homelessness are bound by the parameters of the data themselves: the types and categories of data, and the contexts in which they were recorded (Orr, 2024). This paradigm upholds the assumption that historical truths can be gleaned directly from data through unmediated observation. By relying on digital records to "speak for themselves" without acknowledgments of absent information and perspectives, a database is also a "bundle of silences" (Trouillot, 2011, p. 27) that ossifies particular perspectives as truth—those that are visible and readily represented.

SQL and the Construction of Absence

The relational database model, introduced in 1970 by Edgar F. Codd, revolutionized data management by organizing data into discrete tables, each representing a single coherent concept, and interconnected through shared attributes within rows and columns. The tabular framework enabled data to be efficiently queried, retrieved, and manipulated—a vast improvement over manually traversing navigational databases. Yet, this technological shift also revealed a fundamental tension within data science: how to grapple with the inherent absence and incompleteness of data. This tension within data science became particularly evident through the development of SQL, the *de facto* standard for interacting with relational databases, and the subsequent controversy surrounding how it handles and presents missing information.

Importantly, Codd's (1970) relational model relies on a further flattening of the world as "decontextualized atomized information" (Hayles, 2012, p. 170). The hierarchical structure of information present in earlier iterations of database systems is levelled and replaced with correlative mechanisms that privilege relationships between data elements. A database "represents the world as a list of items, and it refuses to order this list," thus resisting the traditional linear and causal constructions of narrative forms (Manovich, 2002, p. 225). The development of relational databases reflects a broader cultural paradigm that breaks down the world into discrete, modular units—a principle evident not only in computational design but also in the physical buildings in which these technologies were developed (McPherson, 2018, p. 78).

Existing literature on relational databases underscores how databases fragment knowledge with significant implications for capturing the complexities of the real world (Hayles, 2007; McPherson, 2018). Relational databases impose a rigid structure by mandating that relationships between data elements, such as tables, columns, and rows, are predefined and difficult to change without jeopardizing existing data (Dourish, 2022). This rigid structure struggles to accommodate "indeterminate data" (Hayles, 2007, p. 1606) or that which cannot be "cut up and cut down to fit the granular structures of databases" (Liu, 2008, p. 240).

Yet, tracing the controversy surrounding the creation of SQL software for accessing and managing databases reveals an alternative story that destabilizes the closed-world assumption and the foundation of truth claims. This is most evident in the methods through which SQL databases acknowledge and manage the absence of data. IBM practitioners Donald Chamberlain and Raymond Boyce (1974) developed SQL as the commercial application of Codd's (1970) theoretical model. A key feature of SQL was the inclusion of NULL markers—placeholders to designate unaccounted information (Chamberlin & Boyce, 1974; Chamberlin, 2012). NULLs are used in SQL to indicate situations where data are missing, incomplete, or not applicable to a given record. These can include cases where information is unavailable at the time of collection, deliberately omitted by the data provider, or simply does not exist. For the sake of computation, NULLs are "qualitatively different from, and not comparable to, a normal value" (Chamberlin, 1998, p. 30). They are not zero, nor are they an empty string. Rather, they are an unrecorded presence, a placeholder for what is missing or unknown, an indication that something is absent or undefined.

The inclusion of NULLs allowed absences and silences to be explicitly recorded, confronting practitioners with the inherent incompleteness and partiality of data. This violated the closed-world assumption of earlier databases, which treated all unstated information as false (Gottlob & Zicari, 1988). In contrast, SQL's open-world approach acknowledged that knowledge exists beyond the database, potentially unrecorded and unaccounted for. As Chamberlin (1998) argued, the issue of NULLs "forces us to confront, head-on, the fact that databases are sometimes used to model the real world" (p. 28). Unlike traditional archives that obscure absences, NULLs operate by being *visibly exclusionary*, manifesting as ghosts in the data that draw attention to hauntings of the "unstructured and unknown" (Liu, 2008, p. 240) that are systematically excluded or resist capture within datafied form. In this way, NULLs are not just a technical artifact but also a tool for sociological analysis. By revealing the invisible categorical boundaries and omitted accounts that haunt data production, NULLs sensitize researchers to the power dynamics and cultural norms that shape our understanding of data.

For example, NULLs have historically been used to denote individuals who do not conform to the gender binary within database systems (Gaboury, 2018). As such, NULLs draw attention to the dominant heteronormative power dynamics that permeate database systems and their erasure of non-normative experiences. As Butler (1999) articulates, the imposition of social norms regarding natural gender expression creates "the spectres of discontinuity and incoherence" for those who do not conform to these norms, thereby rendering their experiences "false, unreal, and unintelligible" by hegemonic systems of identification (p. 23, xxiii). Incorporating NULLs within data systems thus makes visible these spectres of non-normative experiences that are erased and excluded. Consequently, NULLs offer a means of disrupting the "matrix of intelligibility" (Butler, 1999, p.

24) that underpins hegemonic data practices. Ghosts in the data, made visible and tangible as NULL markers, offer a form of "queer resistance" against data regimes of compulsory identification and categorization for algorithmic profiling and capital extraction (Gaboury, 2018).

By allowing for the absence of data within databases, NULLs also eroded the basis of certainty upon which data analysis was grounded. NULLs increase the computational complexity of every data transaction, requiring a three-value logic system (true, false, or unknown) rather than the traditional two-value logic of navigational database systems (true or false) (Codd, 1970). In doing so, SQL reconfigured the meaning of "truth" as derived from data. Under the open-world assumption, the failure to infer something from data does not mean that it is false, but rather, that it is unknown or yet to be determined. In the COVID-19 example, at least 71 unhoused people died from COVID-19 in Los Angeles in 2020 (Fowle & Gray, 2022). We do not know how many additional documented unhoused people have succumbed to the virus. Nor do we know if any victims were experiencing homelessness, yet undocumented. The presence of NULL markers within databases encourages users to dwell on absences, to consider the perspectives that may be missing (Gordon, 1997), and allows for speculative reimaginings of that which resists representation (Hartman, 2007).

These epistemological implications of NULL markers made their inclusion within SQL contentious. Chamberlin (1998, p. 28) notes that NULLs are a "religious topic" amongst practitioners. While acknowledging their "unpalatable" nature, Chamberlin (1998, p. 30) argued that they are the most suitable approach to capture missing data. Others have disagreed. Upon the release of SQL, Codd (1990) reflected that NULLs should be disambiguated between data that are absent but applicable and those that are inapplicable. Furthermore, Date (2005) argues that NULL markers erode the foundation of authority upon which the relational model of data is built. For Date (2005), the open-world assumption is unnecessarily conservative and misleading. By not considering the likely values of NULLs and excluding records from searches for which information is unknown, NULL markers infuse all results with uncertainty.

> "If you have any nulls in your database, you're getting wrong answers to some of your queries... *all* results become suspect. *You can never trust the answers you get from a database with nulls*" (Date, 2005, p. 55).

For Date (2005), the open-world foundation of databases erodes the very mission of data analysis—to draw trustworthy conclusions from data. This is because NULL markers "contaminate" mathematical operations, as any equation containing NULL markers will result in NULL output (Hayles, 2007). They can easily cascade throughout databases, "rendering everything they touch indeterminate" (Hayles, 2007, p. 1606). Arguing for closed-world approaches to data science, Date (2005, p. 56) maintains that if NULLs are present, "the entire edifice crumbles, and all bets are off."

The controversy regarding NULLs in SQL highlights a central tension within the histories of data analysis: the ability and willingness to capture and communicate uncertainty (Stigler, 1986). In other words, should data science be in the business of certainty or reality? Critics like Date (2005) maintain that data should be supported by probabilistic certainty. However, STS scholars acknowledge that all data are inherently partial and incomplete, embodying some perspectives and occluding others (e.g., Bowker, 2008; Hayles, 2007). NULL markers represent an attempt to recognize and account for the contingency of real-world data.

AI and the Illusion of Completeness

With the emergence of automated decision systems, absences in data take on new significance. While NULL markers in SQL offered a way to acknowledge and represent missing information, contemporary AI systems have largely reverted to closed-world logics. Absent data are seen not as gaps to be acknowledged but as obstacles to be overcome, often at the cost of perpetuating biases and obscuring complexities.

Institutional standards, design norms, and technological affordances now actively work to eradicate the possibility of absence and reinstate a closed-world logic upon datasets. Prior to their use, datasets must be evaluated for their quality. Yet, data quality metrics prioritize "completeness," defined not as capturing all possible data points or perspectives but rather as having values recorded for every observation in every domain (Osbourne, 2013). This standard of completeness has been enshrined within the International Standards Organisation's definitions of data quality (International Standards Organization, 2022). Moreover, many AI techniques simply cannot function with missing values. Practitioners erase these absences by imputing data based on probabilities of other occurrences within the data or merely by removing the records entirely (Broussard, 2018; Osbourne, 2013). These processes institutionalize completeness, as the absence of absence, erasing the possibility of contingency and partial perspectives in AI datasets.

This pursuit of completeness is further amplified by the paradigm of "Big Data," which often assumes that larger datasets inherently lead to more accurate and truthful representations of reality (boyd & Crawford, 2012). Today, training datasets containing terabytes of scraped content present themselves as inherently complete, comprehensive representations of real-world phenomena (Scheuerman et al., 2021). For instance, ImageNet, a popular computer vision dataset, claims to "map out the entire world of objects" by labeling images against a stable universe of nouns that claims to capture the breadth of categories, concepts, and phenomena that humans experience (Crawford & Paglen, 2019). This view attempts to blur the boundary between open and closed-world logics. As information is increasingly recorded, captured, and datafied, fewer pieces of information remain outside the bounds of data. As such, Razniewski and Nutt (2014) argue that completeness is the default state of contemporary datasets. In other words, absences and uncertainty in datasets can be addressed by simply collecting more data (Orr & Crawford, 2024a). The imperative for large-scale datasets in AI research reflects the culture of technological hypervisibility, where "we are led to believe not only that everything can be seen, but also that everything is available and accessible for our consumption" (Gordon, 1997, p. 16). Large-scale datasets attempt to erode the distinction between presence and absence by claiming to encapsulate all experiences and perspectives.

Yet, far from being "the entire internet," even the largest of AI datasets are partial, privileging certain perspectives over others (Baack, 2024). Critical literature highlights how ImageNet contains harmful and erroneous content that perpetuates sexist, racist, and ableist stereotypes of marginalized cohorts (Crawford & Paglen, 2019). Moreover, scraped text datasets, such as Common Crawl, overrepresent the perspectives most prevalent on the web: those of younger English speakers from Global North contexts (Luccioni & Viviano, 2021). Increasing the scale of datasets without accounting for existing disparities is insufficient for ensuring the inclusion of otherwise silenced perspectives (Birhane et al., 2023). In fact, it will only amplify dominant perspectives, leading to a rise in negative stereotypes and hate speech within AI models (Birhane et al., 2023). In this way, the pursuit of scale not only entrenches absences in datasets but also renders them opaque behind claims of universality and completeness.

The problem of absence is further compounded by the datasets used to evaluate AI performance, which also claim to be of totalizing scope. Generative AI benchmarks have attempted to encapsulate the breadth of human knowledge, such as the Massive Multitask Language Understanding (MMLU) dataset, which evaluates Large Language Models (LLMs) across 57 domains from mathematics to US history, medicine, and ethics (Hendrycks et al., 2021). However, datasets that assert completeness occlude the limitations and inherent constraints of AI systems (Raji et al., 2021). Such benchmarks prioritize specific tasks and metrics, thus failing to capture the breadth of real-world complexity (Orr & Kang, 2024).

In practice, dataset creators are often aware of the partial perspectives of their datasets; however, rectifying these issues may be prohibitively expensive or deemed unsuitable by bureaucratic processes or institutional priorities (Orr & Crawford, 2024b). For example, Google's T5 LLM was trained on a dataset that removed any mention of sex, inadvertently also erasing content about LGBTQI+ communities (Dodge et al., 2021). The flawed filtering process was considered acceptable because the primary objective of the dataset was to train an LLM to perform well on conventional benchmarks, and according to the creator, "because most benchmark datasets don't talk about sex, it probably doesn't hurt the model" (Orr & Crawford, 2024a, p. 4965). Here, absences are reinforced through evaluation frameworks, as models are not assessed on the breadth of human experience but rather a subset of standardized instances (Raji et al., 2021). In this way, absences in AI datasets are not "neutral or natural" (Trouillot, 2011, p. 48). Rather, they are explicitly socially constructed and work to silence traditionally marginalized and underrepresented cohorts.

This erasure of absence has profound implications for AI outputs. AI systems reproduce historical relationships and patterns observed within training datasets (Chun, 2021). As such, absences and silenced perspectives in datasets do not simply mean that those observations never occurred, but also that these relations will never occur. Absent information and missing values in training data foreclose the possibilities of alternative explanations (Amoore, 2020; Broussard, 2018), thereby shaping the very possibilities of ML outputs. As Amoore (2020) explains, the vast multiplicity of training data and the potentialities they embody are collapsed into a singular optimized output. This final and finite decision is presented with the authority of the wealth of training data used to construct it. Yet, as Broussard (2018, p. 116) explains, "Not everything that counts is counted." Datasets can never record every observation, perspective, or relevant characteristic (Bowker, 2008). This inherent partiality can lead to misunderstandings and erroneous outcomes when they are relied upon as the foundation of truth claims (Broussard, 2018).

Returning to the COVID-19 example, if LA County developed a classifier to predict whether someone was at risk of contracting coronavirus, given the available data, LA's unhoused population would likely be classified as low risk. This outcome would reinforce this conclusion as true, leading to less funding for pandemic-related health services and testing in these communities. In other words, algorithmic prediction would perpetuate the cycle of neglect and underinvestment that created gaps in data. The inability of a classifier to capture systemic absences highlights the dangers of assuming completeness of data in automated systems.

Indeed, errors, oversights, and breakdowns are all too common in automated systems, which often entrench existing social inequities by relying on incomplete datasets. For instance, automated systems used to predict child maltreatment utilize data about families who use public aid programs, such as food stamps and affordable housing, but occlude information about parents accessing private services, like substance abuse or mental health services (Eubanks, 2018, p. 82). These systems, therefore, may inflate risk scores and unnecessarily scrutinize families with long histories of using public services while potentially missing or justifying the maltreatment of children underrepresented in the system (Eubanks, 2018). In facial recognition, the overrepresentation of pale and male faces within training datasets also led to software not working correctly for people with dark skin tones, worst of all for black women (Buolamwini & Gebru, 2018).

Yet, addressing absences in data is not simply a question of collecting more data. In facial recognition technologies, backfilling datasets to be more representative means collecting data about historically marginalized populations to train technologies that might ultimately be used for surveillance and policing against them (D'Ignazio & Klein, 2020). Likewise, languages not well represented in contemporary training datasets, such as Indigenous languages, may be deliberately excluded by community members to prevent their cultural heritage from being co-opted and profited by tech corporations (Widder & Kneese, 2025). Reckoning with the gaps in AI datasets must go beyond simply collecting more data, but rather grapple with the power structures that have created them.

The resurgence of closed-world logic in contemporary datasets reflects a prioritization of certainty rather than confronting the complexities of the social world. While NULL markers in SQL offer an acknowledgment of absence and uncertainty in data, the pursuit of scale and optimization in automated systems has largely foreclosed this possibility. As such, the ghosts in the data remain unacknowledged and continue to haunt the outputs of automated systems as cycles of error, structural inequalities, and inherent biases.

Towards a Sociology of AI

This history of absence within data infrastructures reveals a persistent tension between the aspiration for certainty from quantitative outputs and the acknowledgment of the inherent incompleteness of data. The inclusion of NULLs within SQL introduced absence as a visible feature that could be counted and tracked. Their presence prompted a fundamental tension regarding the very project of data science, confronting practitioners with the limitations of relying on databases as representations of reality. Yet, the controversy surrounding NULLs in SQL and their eventual omission in contemporary AI datasets highlights the enduring allure of the closed-world assumption and the sociotechnical imaginaries that underpin it. Evaluating the quality of datasets by the absence of visibly absent data produces singular optimizable outputs, at the risk of perpetuating dominant power structures and structural inequalities that haunt the data as unrecorded or unrecordable phenomena.

This case also illustrates the power embedded within infrastructures for recording and remembering. As Trouillot (2011) emphasizes, "the census taker is always a *censor*… he [sic] who counts heads always silences facts and voices" (p. 51). However, this history also draws attention to not only the power of those who create records but also the power of those who create recording infrastructures. Data standards, frameworks, and management software all shape the kinds of information that can be represented and remembered as data, and, thus, what can be known in the future. It is, therefore, imperative to critically examine the affordances and limitations of recording infrastructures that shape the possibilities of knowledge.

Gordon's (1997) analysis of the "unseen" provides a useful framework for uncovering the sociological impacts of absence in data science. Absent data can reinforce dominant epistemologies and narratives, thereby constraining the possibilities of knowledge from data systems. Excavating and dwelling on such absences provides opportunities for alternative explanations and speculative reimaginings that foreground perspectives that resist representation as data. Returning to Los Angeles' unhoused community's experiences with COVID-19, foregrounding the absences of recorded data may necessitate complementing existing data with qualitative sources and firsthand accounts that acknowledge the complexity and variegated impacts of marginalization.

Echoing Gordon's (1997) call to take "ghosts" seriously beyond the "empirical safety net" of what is recorded (p. 45), data science must progress beyond viewing data as a "bucket of facts" (Haigh, 2006) to underscore the potential to both reveal and conceal. A sociology of AI should, therefore, critically interrogate what is not there to uncover the "ghosts in the data" that haunt AI systems. This is not just an examination of the often-hidden labor that shapes data systems (Gray & Suri, 2019) but the underlying social forces and power dynamics that render certain perspectives absent. The sociological toolkit, specifically, lends itself well to this task: the sociological lens is primed for uncovering power dynamics, including those embedded within datasets, data infrastructures, and data collection pipelines. Likewise, sociology's mission to amplify marginalized voices positions it well to highlight those obfuscated or erased from hegemonic readings of data. Moreover, uncertainty and partiality are central to sociology, allowing it to foreground the co-

constitution between data and reality. As Santos (2015) suggests, uncovering the unseen, the unrecorded, and the silenced demands the sociological imagination (p. 181).

Tracing the history of data infrastructures as tools for remembering, I argue for a sociology of AI that acknowledges and embraces partiality and missing perspectives in data. Dataset documentation efforts (e.g., Gebru et al., 2018) are crucial to communicate to users the design decisions and partial perspectives embedded within datasets. However, uncovering absent perspectives is insufficient to achieve data justice and repair. Grappling with the ghosts that haunt data systems necessitates examining the power dynamics that render certain perspectives invisible or erased. It demands not only amplifying complicated histories and marginalized perspectives lost through datafication but also respecting the agency of individuals who prefer to remain unseen.

In practice, this may involve collaborative partnerships between AI designers and community organizations of those affected by automated decisions to critically reflect on the hidden narratives or alternative explanations embedded in data systems. Participatory methods for collecting and curating marginalized voices are central to this process, which can be integrated into data systems if desired by community members. Moreover, instead of upholding the illusion that even the largest datasets are "complete," I advocate for reimagining and actively designing data infrastructures that acknowledge and visualize missing data and absent perspectives. By doing so, AI research can move towards recognizing and valuing multiple forms of evidence, including qualitative sources and firsthand accounts from marginalized communities inadequately captured by traditional data collection pipelines. Engaging with the absences that haunt data infrastructures is necessary for developing AI systems that recognize the complexity of human experience and contribute to a more equitable society.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Will Orr https://orcid.org/0000-0002-6720-5794

References

Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others*. Duke University Press.

Ananny, M. (2020). Presence of absence: Exploring the democratic significance of silence. In H. Landemore, R. Reich, & L. Bernholz (Eds.), *Digital technology and democratic theory* (pp. 141–166). University of Chicago Press.

Baack, S. (2024). *Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI.* Mozilla Foundation. https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/

Banerjee, D., & Bhattacharya, P. (2021). The hidden vulnerability of homelessness in the COVID-19 pandemic: Perspectives from India. *International Journal of Social Psychiatry*, *67*(1), 3–6. https://doi.org/10.1177/0020764020922890

Birhane, A., Prabhu, V., Han, S., & Boddeti, V. N. (2023). On hate scaling laws for data-swamps. arXiv. https://doi.org/10.48550/arXiv.2306.13141

Bowker, G. C. (2008). *Memory practices in the sciences*. MIT Press.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press.

boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. MIT Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html.

Butler, J. (1999). *Gender trouble: Feminism and the subversion of identity*. Routledge.

Chamberlin, D. D. (1998). *A complete guide to DB2 universal database*. Morgan Kaufmann Publishers.

Chamberlin, D. D. (2012). Early history of SQL. *IEEE Annals of the History of Computing*, *34*(4), 78–82. https://doi.org/10.1109/MAHC.2012.61

Chamberlin, D. D., Astrahan, M. M., Blasgen, M. W., Gray, J. N., King, W. F., Lindsay, B. G., Lorie, R., Mehl, J. W., Price, T. G., Putzolu, F., Selinger, P. G., Schkolnick, M., Slutz, D. R., Traiger, I. L., Wade, B. W., & Yost, R. A. (1981). A history and evaluation of system R. *Communications of the ACM*, *24*(10), 632–646. https://doi.org/10.1145/358769.358784

Chamberlin, D. D., & Boyce, R. F. (1974). Sequel: A structured English query language. *Proceedings of the 1976 ACM SIGFIDET (now SIGMOD) workshop on Data description, access and control - FIDET '76'*, 249–264. https://doi.org/10.1145/800296.811515

Christen, K. (2012). Does information really want to be free? Indigenous knowledge systems and the question of openness. *International Journal of Communication*, *6*, 2870–2893. https://ijoc.org/index.php/ijoc/article/view/1618.

Chun, W. H. K. (2021). *Discriminating data: Correlation, neighborhoods, and the new politics of recognition*. The MIT Press.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, *13*(6), 377–387. https://doi.org/10.1145/362384.362685

Codd, E. F. (1990). *The relational model for database management: Version 2*. Addison-Wesley.

Crawford, K., & Paglen, T. (2019). Excavating AI: The politics of training sets for machine learning. https://excavating.ai

Date, C. J. (2005). *Database in depth: Relational theory for practitioners*. O'Reilly.

Davis, J. L. (2020). *How artifacts afford: The power and politics of everyday things*. MIT Press.

Derrida, J. (1995). Archive fever: A Freudian impression (E. Prenowitz, trans). *Diacritics*, *25*(2), 9–63. https://doi.org/10.2307/465144

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286–1305. https://doi.org/10.18653/v1/2021.emnlp-main.98

Dourish, P. (2022). *The stuff of bits: An essay on the materialities of information*. MIT Press.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Foucault, M. (1972). *The archaeology of knowledge*. Harper Torchbooks.

Fowle, M., & Gray, F. (2022). *COVID-19 homeless deaths*. Homeless Deaths Count. https://homelessdeathscount.org/data/covid-19/

Gaboury, J. (2018). Becoming NULL: Queer relations in the excluded middle. *Women and Performance: A Journal of Feminist Theory*, *28*(2), 143–158. https://doi.org/10.1080/0740770X.2018.1473986

Gangadharan, S. P. (2021). Digital exclusion: A politics of refusal. In L. Bernholz, H. Landemore, & R. Reich (Eds.), *Digital technology and democratic theory* (pp. 113–140). University of Chicago Press. https://doi.org/10.7208/chicago/9780226748603.001.0001

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., & Crawford, K. (2018). Datasheets for datasets. arXiv. https://arxiv.org/abs/1803.09010

Gitelman, L. (Ed.), (2013). *"Raw data" is an oxymoron*. MIT Press. https://doi.org/10.7551/mitpress/9302.001.0001

Gordon, A. (1997). *Ghostly matters: Haunting and the sociological imagination*. University of Minnesota Press.

Gottlob, G., & Zicari, R. (1988). Closed world databases opened through null values. *Proceedings of the 14th International Conference on Very Large Data Bases*, 50–61. https://www.vldb.org/conf/1988/P050.PDF

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop silicon valley from building a new global underclass*. Houghton Mifflin Harcourt.

Haigh, T. (2006). "A veritable bucket of facts" origins of the data base management system. *ACM Sigmod Record*, *35*(2), 33–49. https://doi.org/10.1145/1147376.1147382

Hartman, S. V. (2007). *Lose your mother: A journey along the Atlantic slave route*. Farrar, Straus and Giroux.

Hayles, N. K. (2007). Narrative and database: Natural symbionts. *PMLA/Publications of the Modern Language Association of America*, *122*(5), 1603–1608. https://doi.org/10.1632/s0030812900168580

Hayles, N. K. (2012). *How we think: Digital media and contemporary technogenesis*. University of Chicago Press.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. arXiv. https://doi.org/10.48550/arXiv.2009.03300

Hoffmann, A. L. (2021). Terms of inclusion: Data, discourse, violence. *New Media & Society*, *23*(12), 3539–3556. https://doi.org/10.1177/1461444820958725

Holland, G. (2020). *Have L.A.'s homeless people dodged a COVID-19 catastrophe?* Los Angeles Times. https://www.latimes.com/california/story/2020-08-21/why-has-covid-spared-l-a-homeless-people

International Standards Organization. (2022). *ISO 8000-2:2022(en), Data quality—Part 2: Vocabulary.* https://www.iso.org/obp/ui/#iso:std:iso:8000:-2:ed-5:v1:en:term:3.1.2.

Jasanoff, S., & Kim, S. H. (2015). *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power.* University of Chicago Press.

Liu, A. (2008). *Local transcendence: Essays on postmodern historicism and the database*. University of Chicago Press.

Luccioni, A., & Viviano, J. (2021). What's in the box? An analysis of undesirable content in the common crawl corpus. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (volume 2: Short papers)*, 182–189. https://doi.org/10.18653/v1/2021.acl-short.24

Manovich, L. (2002). *The language of new media*. MIT Press.

McFarling, U. L. (2021). *The uncounted: People who are homeless are invisible victims of Covid-19*. Stat News. https://www.statnews.com/2021/03/11/the-uncounted-people-who-are-homeless-are-invisible-victims-of-covid-19/

McPherson, T. (2018). *Feminist in a software lab: Difference + design*. Harvard University Press.

Onuoha, M. (2018). *On missing datasets*. GitHub. https://github.com/MimiOnuoha/missing-datasets/blob/master/README.md

Orr, W. (2024). Counting on stability: The social construction of the Los Angeles homeless count. *International Journal of Communication*, *18*(2024), 808–815. https://ijoc.org/index.php/ijoc/article/view/21845/4469.

Orr, W., & Crawford, K. (2024a). The social construction of datasets: On the practices, processes, and challenges of dataset creation for machine learning. *New Media & Society*, *26*(9), 4925–5572. https://doi.org/10.1177/14614448241251797

Orr, W., & Crawford, K. (2024b). Building better datasets: Seven recommendations for responsible design from dataset creators. *Journal of Data-Centric Machine Learning Research*, *1*(1), 1–21. https://doi.org/10.48550/arXiv.2409.00252

Orr, W., & Kang, E. B. (2024). AI as a sport: On the competitive epistemologies of benchmarking. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1875–1884. https://doi.org/10.1145/3630106.3659012

Osborne, J. W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage.

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, *14*(3), 399–441. https://doi.org/10.1177/030631284014003004

Raji, I. D., Denton, E., Bender, E. M., Hanna, A., & Paullada, A. (2021). AI and the everything in the whole wide world benchmark. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, *1*, 1–20. https://doi.org/10.48550/arXiv.2111.15366

Razniewski, S., & Nutt, W. (2014). Databases under the partial closed-world assumption: A survey. In Proceedings of the 26th GI-Workshop on Foundations of Databases, October, 2014, Bozen, Italy.

Reiter, R. (1978). On closed world data bases. In H. Gallaire & J. Minker (Eds.), *Logic and data bases* (pp. 55–76). Springer US.

Rosenberg, D. (2013). Data before the fact. In L. Gitelman (Ed.), *"Raw data" is an oxymoron* (pp. 13–40). MIT Press. https://doi.org/10.7551/mitpress/9302.001.0001

Santos, B. de S. (2015). *Epistemologies of the south: Justice against epistemicide*. Routledge.

Scheuerman, M. K., Denton, E., & Hanna, A. (2021). Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. Proceedings of the ACM on Human-Computer Interaction, *5*(CSCW2), 1–37. https://doi.org/10.1145/3476058

Scott, S. (2018). A sociology of nothing: Understanding the unmarked. *Sociology*, *52*(1), 3–19. https://doi.org/10.1177/0038038517690681

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Belknap Press of Harvard University Press.

Stoler, A. L. (2016). *Duress: Imperial durabilities in our times*. Duke University Press.

Stone, M.D. (1960). Data Processing and the Management Information System: A Realistic Evaluation of Data Processing's Role in the Modern Business Enterprise. In American Management Association ed. *Data Processing Today: A Progress Report – New Concepts, Techniques and Applications – AMA Management Report Number 46*. American Management Association, Finance Division, New York, 1960, 14-22.

Trouillot, M.-R. (2011). *Silencing the past: Power and the production of history*. Beacon Press.

Widder, David Gray, & Kneese, Tamara (2025). Salvage Anthropology and Low-Resource NLP: What Computer Science Should Learn from the Social Sciences. *Interactions*, *32*(2), 46–49. DOI:10.1145/3714996

Winner, L. (1993). Upon opening the Black box and finding it empty: Social constructivism and the philosophy of technology. *Science, Technology, & Human Values*, *18*(3), 362–378. https://doi.org/10.1177/016224399301800306

Author Biography

Will Orr is a PhD candidate at the University of Southern California's Annenberg School of Communication. His work examines the sociotechnical dynamics of creating and evaluating Artificial Intelligence systems, exploring the cultures, practices, and epistemologies that shape these technologies. .