# Contextual Integrity Up and Down the Data Food Chain

*Helen Nissenbaum**

*According to the theory of contextual integrity (CI), privacy norms prescribe information flows with reference to five parameters — sender, recipient, subject, information type, and transmission principle. Because privacy is grasped contextually (e.g., health, education, civic life, etc.), the values of these parameters range over contextually meaningful ontologies — of information types (or topics) and actors (subjects, senders, and recipients), in contextually defined capacities. As an alternative to predominant approaches to privacy, which were ineffective against novel information practices enabled by IT, CI was able both to pinpoint sources of disruption and provide grounds for either accepting or rejecting them. Mounting challenges from a burgeoning array of networked, sensor-enabled devices (IoT) and data-ravenous machine learning systems, similar in form though magnified in scope, call for renewed attention to theory. This Article introduces the metaphor of a data (food) chain to capture the nature of these challenges. With motion up the chain, where data of higher order is inferred from lower-order data, the crucial question is whether privacy norms governing lower-order data are sufficient for the inferred higher-order data. While CI has a response to this question, a greater challenge comes from data primitives, such as digital impulses of mouse clicks, motion detectors, and*

---

* Information Science, Cornell Tech, NYC. Author gratefully acknowledges grants from US NSF CNS-1801501, CNS-1704527, SES1642553, SES1537324, and NSA H98230-18-D-006. This work has benefited from exposure at several events, principally, Tel Aviv University, Conference on Theorizing Privacy, Apple University, EPFL, and CCS 2018. It reflects incisive comments from colleagues and collaborators -- Sebastian Benthall, Anupam Datta, Seda Guerses, Kirsten Martin, Vitaly Shmatikov, Michael Tschantz and, particularly Michael Birnhack and Julie Cohen for urging me forward. Enormous credit goes to the editors of Theoretical Inquiries in Law for a remarkable job on substantive and stylistic matters. Finally, thanks to Katherine Magruder for stellar research assistance at all stages.

Cite as: Helen Nissenbaum, *Contextual Integrity Up and Down the Data Food Chain*, 20 Theoretical Inquiries L. 221 (2019).

*bare GPS coordinates, because they appear to have no meaning. Absent a semantics, they escape CI's privacy norms entirely.*

## Introduction

The theory of contextual integrity offered a conception of privacy that characterized not only how people think about privacy but why it is worth caring about.[1] It explained the existential threats to privacy from information and computational technologies and why these threats evaded entrenched conceptions and regulatory approaches. At the present time, we are experiencing a new wave of privacy threats. These are riding the wave of transformations in data practices enabled by scientific breakthroughs proceeding, some might say, at a revolutionary pace. Even when technological progression is historically cumulative — with seeds of the present evident in the past — it may be felt as sudden and discontinuous when what occurs below the surface of public attention crosses a threshold into practical impact and bursts into public view. We have experienced one such discontinuity in the arena described by a series of terms, including "big data" (breathlessly uttered), data mining, predictive analytics, Internet of things (IoT), mobile applications (Apps), data science, machine learning (ML), and new wave artificial intelligence (AI) based on ML.

My aims for this Article are twofold. One is to describe the fundamentals of the theory of contextual integrity (CI) in relation to other approaches and to information (and now, data) technologies. For some this will serve as an introduction, but even to those familiar with CI the description in its current form helps to elucidate key ideas. The second aim is to describe a facet of big data technologies that accounts for their distinctive challenges to many approaches to privacy (in some cases, their undoing), including CI. The Article concludes with ideas for mitigating the damage to privacy were nothing to be done but proceed on the existing course. It is premature for the announcement of solutions and recommendations whose nature, in my view, will depend on how successful we are in engaging stakeholders, key actors, and public interest advocates in collaborative rather than adversarial paths forward.

At a glance, privacy threats from big data technologies resemble those of previous decades, as they carry the seeds of prior digital state-of-the-art, including information systems, databases, database matching, profiling,

---

1 Helen Nissenbaum, Privacy in Context: Technology, Policy and the Integrity of Social Life 162-64 (2010) [hereinafter Nissenbaum, Privacy in Context]; Helen Nissenbaum, *Respecting Context to Protect Privacy: Why Meaning Matters*, 24 Sci. & Engineering Ethics 831 (2015).

the Net, *etc.* In *Privacy in Context*, I identified the sources of threat to the capacities of digital technologies to disrupt patterns of data flow profoundly, and roughly organized them into three categories: monitoring and surveillance, dissemination and communication, and aggregation and analysis.[2] Although time and the longer view reveal characteristic practices of big data technologies (ML, AI, *etc.*) as close kin of their predecessors, it would be a disingenuous understatement to frame them as *merely* incremental. The confluence of advances in hardware and software and a permissive political economy has resulted in virtually unfettered data flows and capacities to exploit this data in orders of magnitude more powerful than before, resulting in threats to privacy orders of magnitude more acute.[3] As noted above, that incremental technical developments can bring about quantum shifts in socially relevant outcomes, exceeding discrete thresholds and bursting boundaries of once-sound concepts, is not a new idea in the history of technology, and I am not alone in believing we are at such a juncture with privacy.[4]

Further, to reflect relevant changes, I have devised the metaphor of a data chain based on that of a food chain. Though literal accuracy is not asserted, the idea of a hierarchy or directional chain offers a useful explanatory vehicle for distinguishing between the challenges of aggregation and analysis, roughly a decade ago, and the contemporary experience. Data collection and monitoring technologies (*e.g.*, video surveillance, Web tracking) provide a point of entry, an initial position on the data chain; analytics provides the means of traversing up and down it. For many of the privacy challenges that can be mapped onto the data chain, my point of reference is contextual integrity. Whereas CI has overcome pitfalls and blind spots from which other conceptions have suffered, the data chain metaphor exposes sociotechnical practices associated with big data and AI that challenge not only these conceptions but contextual integrity, as well.

---

2 Nissenbaum, Privacy in Context, *supra* note 1.

3 *See, e.g.,* Solon Barocas & Helen Nissenbaum, *Big Data's End Run Around Procedural Privacy Protections*, 57 Comm. ACM 31 (2014).

4 For classical perspectives, see Landon Winner, *Do Artifacts Have Politics?*, 109 Daedalus 121, 122 (1980); Melvin Kranzberg, *Technology and History: "Kranzberg's Laws,"* 27 Tech. & Culture 544, 545 (1986); Bryan Pfaffenberger, *Technological Dramas*, 17 Sci. Tech. & Hum. Values, 282 (1992). Some of my views on the matter were made public elsewhere. *See* Helen Nissenbaum, *From Preemption to Circumvention: If Technology Regulates, Why Do We Need Regulation (And Vice Versa)?*, 26 Berkley Tech. L.J. 1367 (2011). For current scholarship on the matter, see Mireille Hildebrandt, *Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning*, 20 Theoretical Inquiries L. 83 (2019).

This Article comprises two main parts. Part I provides an overview of the theory of contextual integrity in terms of four fundamental theses, which also serve to differentiate the conception of privacy as contextual integrity from those defined in other accounts of privacy. Part II introduces the metaphor of a data food chain as a mechanism for explaining a family of challenges confronting normative accounts of privacy. Although the idea of a data food chain could be applied retroactively to challenges of previous decades, it reveals one of the deepest, most intractable challenges of contemporary big data technologies, namely operations and processes at the layer of non-semantic data primitives. The Article reveals why these challenges are particularly relevant to supporters of privacy-by-design, seeking to enforce privacy constraints within technical systems.

## I. Contextual Integrity: Briefly Motivated and Described

The origin of CI was not as an alternative to important philosophical accounts seeking to define a coherent and distinctive concept and explain privacy's normative force.[5] Rather, it emerged in an attempt to understand what people saw threatened by novel sociotechnical practices wrought by a family of technologies, including computers, digital networks, information systems, databases, communications media, electronic hardware, and software. Thus, it stands (or falls) on its ability to account for phenomena many have labeled *assaults on privacy*.

### A. CI Fundamentals: Four Theses

According to the theory of contextual integrity (CI), privacy, defined as CI, is preserved when information flows generated by an action or practice conform to legitimate contextual informational norms; it is violated when they are breached. To elaborate this definition, this Article characterizes the theory of contextual integrity (CI) in terms of four fundamental theses, each an incremental progression from the one before. These theses convey the substantive assertions of CI, while providing a vehicle for comparing CI to other conceptions, definitions, and theories.

---

5 *See* Ferdinand David Schoeman, Philosophical Dimensions of Privacy: an Anthology (2007); Anita L. Allen, Unpopular Privacy: What Must We Hide? (2011); Judith W. DeCew, In Pursuit of Privacy: Law, Ethics, and the Rise of Technology (1997).

*1. Thesis 1: Privacy is the Appropriate Flow of Personal Information*

This thesis acknowledges the critical importance of information about persons as fuel for a robust social sphere. There is nothing wrong with sharing or gathering information about ourselves and others; there is no presumption in favor of hoarding, holding, or stopping flow. The theory of contextual integrity (CI) does not valorize information containment. Privacy as contextual integrity does not accept the implications of other definitions that identify privacy with no flow, with stoppage, secrecy, and data minimization. It does not agree that when Alice and Bob are talking, Eve always violates their privacy. It does not identify data leakage as a privacy harm, or any collection as privacy violation. CI cares only whether the flow is *appropriate*, not whether it takes place at all.

*Flow* plays such an important role in the articulation of contextual integrity that it deserves its own paragraph. I chose flow to serve as a neutral term to refer to the passage or transmission of information or data from party (or parties) to party (or parties). Each of the alternative terms I considered, such as share, collect, disseminate, distribute, transmit, receive, or communicate, was richer in meaning than flow, and their augmented meanings bundle assumptions that theory requires be made explicit.

Thesis 1 gives CI an important strategic advantage over definitions of privacy as secrecy, which, in my view, imperil privacy's moral standing. Although privacy as secrecy, or stoppage of flow, provides clarity to the concept, its legitimate scope is extremely narrow. How many times have we heard people announce that privacy must be balanced against — insert your favorite alternative — security, efficiency, convenience, usability, functionality, commercial profit, public health, *etc.*, or the generic version, "trade privacy for utility!"? What these people almost always mean is that information must flow in order to support security, convenience, and so on, thus implicitly adopting a definition of privacy as stoppage of flows. If privacy *were* stoppage, or secrecy, it would stand to reason that more often than not it would need to be traded off against other useful, socially valuable exchanges. But, as contextual integrity, privacy allows for information flows that are appropriate, including flows needed to promote utility — *i.e.*, security, convenience, and the like.

*2. Thesis 2: Appropriate Flows Conform with Contextual Informational Norms ("Privacy Norms")*

Thesis 1 leaves open the question of what it means for flows to be appropriate. To answer, Thesis 2 introduces the construct of contextual informational norms that express or characterize information flows. Building on the work of social theorists and philosophers, this construct presumes a conception of social life

not as an undifferentiated whole, but as constituted by distinct social contexts.[6] Although these works posit differing logics and labels — spheres, realms, fields, institutions, domains — nevertheless, they share a view of society as constituted by diverse domains. CI takes on board some of the common insights without committing to any specific one of these theories. With the general term *context* CI connects with neighboring theories,[7] while also drawing on intuitive understandings of distinct social domains and even quite common societal arrangements, such as law and policy in the contemporary United States, which acknowledges these differing domains with differing bodies of law for distinct spheres of engagement and respective institutions, including, for example, commercial, constitutional, family, financial, workplace, and health. As noted, although contextual integrity is committed to differentiated social space, it is tied neither to any one theoretical account nor to any one paradigmatic society.[8]

In assuming that society comprises multiple social spheres, CI does not commit to a particular set or arrangement of spheres and allows for different societies (historical eras, cultures, *etc.*) to comprise different ones. It does, however, conceive of a particular structural arrangement, including several key contextual constituents. These include roles or capacities in which people act; paradigmatic activities and practices; and respective ontologies. They may also include paradigmatic institutions and venues. Physicians, physical therapists, and patients; physical examinations, blood tests, and treatment; and symptoms, insurance forms, illnesses, diagnoses, and medications; hospitals, ambulances, and physicians' offices are among the constitutive elements of the healthcare context, just as teachers, students, reading, writing, studying, schools, and universities partially constitute education.

In the past,[9] I had not placed sufficient emphasis on functions, purposes (goals, ends), and values around which contexts are oriented. I would like to make up for this prior lapse by saying that these — let's call them teleological — factors define the very essence of a respective social context. Even though,

---

6 *See, e.g.,* Pierre Bourdieu, Distinction: A Social Critique of the Judgment of Taste (Richard Nice trans., 1984); John R. Searle, The Construction of Social Reality (1995); Michael Walzer, Spheres of Justice: A Defense of Pluralism and Equality (1984); Nissenbaum, Privacy in Context, *supra* note 1, at 129-58.

7 For an elaboration of this discussion, see Nissenbaum, Privacy in Context, *supra* note 1, at 132-37.

8 Admittedly, many of my illustrations are drawn from contemporary life in the United States, revealing my somewhat limited experience. Supporting observations drawn from a wider range of societies and historical periods would enormously strengthen the empirical basis of CI.

9 *See e.g.*, Nissenbaum, Privacy in Context, *supra* note 1.

or perhaps because, they are the most important aspects of social contexts, they frequently are the most contentious, debated, and controversial aspects. We may agree that among the defining aims of healthcare are alleviating pain, preventing contagion, or curing disease, but we may disagree over whether prevention is more important than cure, prolonging individual life more important than average population health, and so forth. Some have argued that healthcare values include equity, the provision of care (or organs for transplants) according to need, irrespective of ability to pay; others disagree. Some hold that physicians should respect whatever paths patients choose; others insist that physicians have a right and duty to steer. Although, from the beginning, teleology has been a steady part of CI, experience has shown that it is frequently overlooked. Social contexts are what they are because of respective contextual aims, purposes, and values.

One caution is to avoid thinking of contexts in spatial terms, although, admittedly, standard usage allows for both spatial and non-spatial meanings. Respective roles, activities, purposes, information types do not exist *in a* context; rather, these factors *constitute* a context. Although certain places are generally associated with respective contexts — hospitals, schools, department stores, churches, governors' mansions — they do not, by themselves, define a context.

We have mentioned roles, practices, goals, and values but not yet contextual informational norms, a fundamental building block of contextual integrity. Though we now turn to them, space constraints require that we do so with great brevity and insufficient detail.[10] Consider the term *norm* to have a meaning close to the term *rule*, and social or societal norms to be norms that govern individuals insofar as they are members of societies. The reason for preferring *norm* to *rule* is the former's flexibility. While rules tend to be explicit and emanate from authoritative sources, norms may be explicit or implicit, may emanate from a variety of sources, may or may not be enshrined in law, may be commanded or merely emergent, may vary over time and across cultures, may be strict or approximate, may be universally or merely locally known, and so forth. I apply the term *contextual norm* to norms that describe, prescribe, proscribe, and establish expectations for characteristic contextual behaviors and practices.

As an empirically or historically discoverable sociological matter, contextual norms can be finely or coarsely grained and finely or coarsely tuned. In certain situations, such as a court of law, behaviors are finely governed by norms (and rules), whereas in others, such as a social event, they may be

---

10 Readers wishing more details are advised to consult, see Nissenbaum, Privacy in Context, *supra* note 1, at 129-58.

quite nonspecific. In general, because the existence of norms shapes people's expectations, behaviors or practices that contravene them are commonly met with surprise, shame, anger, or even punishment.

Among contextual norms are those that govern information flows. Thesis 2 asserts that information flows are judged appropriate insofar as they conform to, or at least do not contravene, these norms. The presence of norms may explain, for example, why one feels angry or disappointed when a good friend reveals to others the details of one's troubled marriage, why we do not ask coworkers (even friends) about their salaries, or why we assume that our votes in a democratic election are not known by government officials. Nonetheless, constituents expect the voting patterns of their political representatives to be available to them, clients expect lawyers to share their educational credentials, the Internal Revenue Service requires citizens to reveal earnings, and building owners expect tenants to provide details of their financial standing. *Notice: Although the last four involve information sharing, none of them involve tradeoffs of privacy because the flows are appropriate.*

Thesis 2 differentiates CI from procedural approaches to privacy, such as "notice and choice," ubiquitous in the contemporary digital landscape of websites, social platforms, mobile systems, and apps. They are procedural because no matter what the substance of the practice, as long as subjects are notified and are allowed either to refuse or consent, privacy has been duly respected. I have argued that the original *Code of Fair Information Practice Principles*[11] is largely procedural, exhorting information collectors to follow prescribed steps in their bid for "fair" practices.[12] Although Thesis 2 does not rule out procedural constraints, it stands by the idea that substantive, normative dos and don'ts define appropriate flows.

A word on terminology: Going forward, I use the shorter term *privacy norm* interchangeably with *contextual informational norm.*

### *3. Thesis 3: Five Parameters Define Privacy (Contextual Informational) Norms: Subject, Sender, Recipient, Information Type, and Transmission Principle*

Thesis 3 asserts that fully articulated contextual informational norms prescribe flows in terms of (actors) who sent the information, who received it, about whom it is, what types of information are involved, and constraints imposed on them (transmission principles). To ascertain whether an action or a practice

---

11 U.S. Department of Health Education & Welfare, Records, Computers, and the Rights of Citizens: Report of the Secretary's Advisory Committee on Automated Personal Data Systems (1973).

12 The principle governing database security might be an exception.

respects privacy, values for all five parameters must be specified in order to map resulting flows onto governing norms. The comparative template that CI provides has, arguably, served as one of its most successful contributions, for it effectively reveals changes to which other approaches are blind. Thus, defenders of a video-cam may reject complaints because video subjects are "in public" and can be seen by anyone passing by. CI reveals, at the very least, that a transmission principle has changed — reciprocity — for no longer is it possible for subjects to see those who see them. Similarly, the "no change" defense of public records (including court records) rendered digitally and placed online ("public is public") can be challenged by carefully comparing the five parameters before (paper records in courthouses) and after (digital records, available online).[13]

Several points are worth noting, both to elaborate on Thesis 3 and to reveal the line it draws between CI and other approaches. First, values for actor and information-type parameters are identified in terms of respective contextual ontologies. Subjects, senders, and recipients are described in their contextual roles, that is, are acting in capacities drawn from contextual ontologies, whether physician, teacher, elected politician, priest, customer, police officer, investor, friend, or congregant. Information likewise is conceptualized according to contextual ontologies, whether symptoms, medications, grades, voting records, demographics, salary, social security number, or church donations.

Second, evidence supports the fidelity of contextual informational norms to privacy expectations. One source is U.S. regulation where a close analysis of privacy rules for financial and health information, following passage of the Gram-Leach-Bliley Act (GLBA) and Health Insurance Portability and Accountability Act (HIPAA), respectively, revealed the presence of CI's five parameters. This suggests that where precision is needed relevant parameters are not overlooked.[14] A second source is direct empirical scientific study of people's expressed privacy expectations, including results from two large

---

13 *See* Amanda Conley et al., *Sustaining Privacy and Open Justice in the Transition to Online Court Records: A Multidisciplinary Inquiry*, 71 Md. L. Rev. 772 (2012); Robert Gellman, *Public Records — Access, Privacy, and Public Policy: A Discussion Paper*, 12 Gov't Info. Q. 391 (1995).

14 To illustrate, one of the HIPAA rules asserts, "A covered entity can disclose a patient's psychotherapy notes to the patient only with the prior approval of the patient's psychiatrist." The Health Insurance Portability and Accountability Act of 1996 (HIPAA).
P.L. No. 104-191, 110 Stat. 1938 (1996). For more detail, see Nissenbaum, Privacy in Context, *supra* note 1, at 96; Adam Barth et al., *Privacy and Contextual Integrity: Framework and Applications*, *in* 2006 IEEE Symposium on Security and Privacy 184 (2006).

factorial vignette studies which not only demonstrated that people have a refined appreciation of all five parameters, but reinforced the proposition that descriptions of information practices that do not cover all the parameters are ambiguous and may be interpreted in different ways by different people.[15]

Third, many have found the transmission principle (TP) parameter to be puzzling because, on its face, it is less familiar to accounts of privacy than actor-capacities and information types. A closer look reveals that TPs do perform a decisive function in other accounts, and definitely in law and policy, but are not explicitly recognized as such. Recognizing transmission principles as a distinct dimension of privacy norms has given CI a richer set of variables with which to characterize data flows in privacy-relevant ways.

Transmission principles are quickly demystified once one considers common instances, such as consent. No-one following privacy is unfamiliar with subject-consent, but conceived as a TP within a norm, its action is to condition the flow of information from sender to recipient on the consent of the information subject. Beyond consent, the possibilities for constraints that may serve as TPs are endless. Although some are more salient than others, such as, "with notice," "entitled," "required by law," "coerced," "reciprocal," "in confidence," "buy," and "sell," TPs are structurally generative. Take consent. The most commonly assumed is subject's consent, but it need not be so; consent may be required from parents or other legal guardians and may be either necessary or sufficient, or both. Authorization is similarly generative as to both the authorizing parties and the conditions that need to be met. The 4th Amendment of the US Constitution is a case in point, requiring "a warrant," meaning authorization from a judge provided on condition of showing "probable cause."

Fourth, Thesis 3 challenges privacy truisms that generally hold sway in public debates, research, and approaches to public opinion polls: a) challenging the dichotomy of public/private data and positing, instead, the promotion of a multiplicity of information types reflecting contextual ontologies; b) challenging accounts that hang privacy status on only one factor, *e.g.*, subject control, or whether the information is "sensitive," instead maintaining a simultaneous focus on all parameters (In the Cambridge Analytica scandal, for example, it was not only that subject consent was bypassed; outrage focused

---

15 *See* Kirsten E. Martin & Helen Nissenbaum, *Privacy Interests in Public Records: An Empirical Investigation*, 31 Harv. J.L. & Tech. 111 (2017) [hereinafter Martin & Nissenbaum, *Privacy Interests*]; Kirsten E. Martin & Helen Nissenbaum, *Measuring Privacy: An Empirical Test Using Context to Expose Confounding Variables*, 18 Columbia Sci. & Tech. L. Rev. 176 (2016) [hereinafter Martin & Nissenbaum, *Measuring Privacy*].

on Cambridge Analytica as an unacceptable recipient of the data that were gathered by Facebook.); c) challenging the chokehold of subject control as the sole arbiter of privacy, instead highlighting alternative TPs.

Fifth, and lastly, CI requires that values for *all five* parameters be specified. As argued above, failing to do so results in an incomplete and ambiguous account of information flows. However, because working with five independent parameters is demanding, certain circumstances may justify a reduction. For example, in a healthcare context, it is usually safe to assume that the data subject is the patient, in education, the subject is a student, *etc*. Or, in developing an app for specific needs where one can count on a closed set of values for one or more of the parameters, one may fix others; for example, if the app is for use among friends, one might justifiably settle on subject consent (or subject control) as the universal TP. Or, in an access control system built for use in a hospital the number of norm permutations that need to be considered can be controlled, since the context and circumstances of use impose natural limits on the variety of actors and information types for which it must account. Finally, within carefully defined settings, it might even make sense to bundle information into two categories, enabling different constraints for each. Thus, information about mental disorders may be judged more sensitive than most other medical conditions and fewer known parties may have access, under more restrictive conditions. The circumstances in which such reductions are justified nevertheless call for vigilance, particularly to stay true to the assumptions about context of study and use.

4. *Thesis 4: The Ethical Legitimacy of Privacy Norms is Evaluated in Terms of: A) Interests of Affected Parties, B) Ethical and Political Values, and C) Contextual Functions, Purposes, and Values.*

The CI narrative proceeds as follows: Social life is thick with flows of personal information, which may and sometimes may not conform with entrenched contextual informational norms. A practice violates a privacy norm if resulting flows fail to map onto expected values for the parameters. Consider, hypothetically, if the U.S. Census Bureau were to share raw data with the Immigration and Naturalization Service, or if a priest were to divulge a congregant's confession to a third party. In such cases, CI would hold that a *prima facie* violation of contextual integrity has occurred, but would not necessarily cease evaluation at that point. Always resisting practices that violate entrenched norms would make CI an exceedingly conservative theory and would be exceedingly problematic for two reasons. One is that CI is designed to respond to the challenges of rapidly evolving sociotechnical practices and many novel flows may be highly beneficial. The other is that

intransigence in the face of change sets CI as a descriptive theory, with no capacity to evaluate — either entrenched norms or contravening principles.

The approach outlined in *Privacy in Context* remains conservative, but instead of flatly rejecting novel flows, it presumptively favors entrenched norms, simultaneously offering a process for adjudicating between the two. To begin, evaluators follow the approach laid out in CI for pinpointing changes, followed by a three-layered comparative analysis to locate respective strengths and weaknesses. Based on these findings, the evaluator recommends in favor of either the status quo or the challenger on grounds of moral quality.

According to Thesis 4, the first two layers are a) interests and b) ethical and political values, both unsurprising to anyone following the privacy debates. Public deliberations surrounding law and policy pay great heed to asking familiar questions. For a given information practice, who wins and loses? Whose preferences and interests are served; whose are not? What are the costs, what are the benefits? While economic arguments often downplay the interests of data subjects, privacy advocates have highlighted harms to subjects from data exposure, such as identity theft and embarrassment. In the literature, researchers have pointed to longer-term threats and subtler harms, such as helplessness, shrinking self-determination, losses of power, boundary control, and the reduced ability to modulate relationships.[16] Beyond the interests of affected parties, there are many who have focused on the ethical and political values at stake, going back to the U.S. drafters of the FIPPS, who sought to level the playing field for the owners and controllers of information systems and the subjects of these systems. To critics, it was not merely that decision makers learning facts about you could harm your prospects (*i.e.*, your interests), but they could be acting against your interests unfairly and unjustly. Privacy advocates defended the connections between privacy and ethical and political ends about which there was broad agreement — freedom of speech and thought, political freedom, and autonomy.[17]

Layer c) of Thesis 4 introduces a distinctively contextual consideration. Influenced by Priscilla Regan's groundbreaking work on the social value of privacy,[18] CI requires that information practices be evaluated in terms of contextual functions, purposes, and values. Let us return to the healthcare context where, for centuries, medical professionals have sworn to the secrecy

---

16 *See* Schoeman, *supra* note 5; Julie E. Cohen, *Examined Lives: Informational Privacy and the Subject as Object*, 52 Stan. L. Rev. 1373, 1438 (1999).

17 *See* Neil M. Richards, Intellectual Privacy: Rethinking Civil Liberties in the Digital Age (2017); Cohen, *supra* note 16.

18 Priscilla M. Regan, Legislating Privacy: Technology, Social Values, and Public Policy 92-108 (2009).

of their patients' health conditions. Confronting the risks of information technologies and pressures to share information, privacy advocates have warned against inadequate protections, citing harms of discrimination, shame and embarrassment, reduced employment opportunity, and so forth. These are all good reasons to ensure that healthcare professionals honor the obligations of entrenched privacy norms to keep patient information out of the grasp of, say, advertisers, prospective employers, curious onlookers, or the press.

As convincing as these warnings are, a dispassionate analyst would proceed to weigh the interests of patients against others whose interests are affected by flows of health information, for example, prospective employers seeking to hire candidates with the best health prospects, marketers seeking promising customers for a new drug, or even users of a dating site wanting the most robust partners. For many privacy debates, this is the battleground; sleeves are rolled up and fervent arguments weigh in favor of one or the other side. For CI, however, missing from these debates is a critical consideration, namely contextual purposes and values. Beyond stakeholder interests, analysts must seek constraints on information flow that promote the goals and values of the healthcare context. This argument proceeds as follows: if patients are fearful that medical information will flow to the wrong parties, they may lie to their physicians, hop from doctor to doctor, or not seek medical advice at all. So doing, they place not only their health at risk, but the health of others, thereby undermining contextual purposes.

Contextual purposes and values sometimes may disfavor the data subject's interests, even in the healthcare context. Having advanced in our understanding of environmental health hazards and communicable diseases, CI may support overriding individual patient interests or preferences in favor of onward sharing of information with others, for example, public health officials. This would allow information to be aggregated and analyzed, hazards (toxic chemicals or restaurants with poor hygiene practices) to be pinpointed, and the further spread of disease contained. Similar lines of reasoning support adherence to privacy norms for other contexts, *e.g.*, student privacy, or voter privacy, in support of contextual ends, not only to secure respective interests.

In the years since the first comprehensive account of CI was published in *Privacy in Context*, the theory has confronted important questions. Relevant to Thesis 4 is a question about recommended steps for evaluating the moral legitimacy of entrenched norms. According to the book's narrative, when established, normative practices confront competing, novel, disruptive sociotechnical practices, the analysis that ensues compares the two in terms of the three-layer criteria. In presenting CI to various audiences, I have come to appreciate the innumerable, unprecedented practices and associated data flows that defy any easy comparison with preexisting practice that may

serve as reasonable precursors or counterpoints. It turns out that although the advantage of having a clear counterpoint is being able to choose the better performer, in fact, conducting an evaluation, as described in Thesis 4, does not require locating a reasonable counterpoint as a first step. An analysis that first carefully maps out data flows in terms of the five parameters and then proceeds to consider interests, societal values, and contextual ends and values may provide sufficient insight to reveal the moral legitimacy or moral hazard of a given practice.

### B. CI General Highlights

Before concluding Part One, I offer a few concluding remarks. First, again as regards why CI begins a comparative evaluation with a presumption favoring entrenched norms, this conservative stance relates to the choice of appropriateness, not correctness or excellence, as the initial entry point. Appropriateness suggests a balance, already a societal compromise. It deserves utmost protection not because it favors the interests of individual subjects above all others, or the opposite, but because it already represents a settled accommodation of diverse interests as well as societal and contextual ends. I have favored entrenched norms, presuming that they reflect a settled accommodation. Critical theorists may chide me; at best, this is a naïve ideal, a fiction. The starker reality is that practices are entrenched because they are favored by society's powerful, the entitled. If this is so, my hope is that the clear-eyed pursuit of CI's evaluative analysis is as likely as any to reveal the tyranny of convention at the same time as it reveals the delicate balance of multiple conflicting interests and plurality of values that complex constraints on flow seek to realize.

The cascade of four theses allows for stops along the way. One may, for example, subscribe to Thesis 1, but not the rest; Theses 1 and 2, but neither 3 nor 4. And so on. Worth noting are the challenges to Thesis 3, asserting that the set of five parameters is incorrect or incomplete. I'd like to highlight one version of this challenge, posed by Anupam Datta, an important collaborator whose work formalizing CI significantly sharpened it. Datta has urged the inclusion of a use parameter, noting the frequency with which use is cited as a factor in the U.S. legal domain, importantly, in HIPAA and GLBA privacy rules.[19] I am increasingly persuaded that adding a use parameter might, after all, be a necessary antidote to a policy environment in which data holdings

---

19 Deepak Garg, Limin Jia, & Anupam Datta, *Policy Auditing Over Incomplete Logs*, *in* CCS'11 Proceedings of the 18th ACM Conference on Computer and Communications Security 151 (2011).

may be obtained through unimpeded company takeovers even when data flow between the companies in question may be heavily conditioned, or even prohibited.

CI opens the door to privacy regulation and design that is, at once, more nuanced as well as more precise. Resistance to regulation frequently is framed as fear of overly blunt regulation — all or nothing, yes-flow or no-flow. Because five parameters provide five dimensions of variation, regulation can be precisely tailored to need. Regulators can consider restricting recipients, or articulating particular fields of information, or can adjust transmission principles (as discussed above) with significant sensitivity. Similar strategies work for system design as well. Turning to CI, technology developers with ambitions to promote privacy-by-design could ensure that the languages they choose to express the rules governing data flow have sufficient expressive power to embed multiple variables. Formal language experts are beginning to offer such languages and associated logics.[20]

## II. The Data Food Chain

I offer the data food chain not as a precise theoretical construct but as a helpful metaphor to expose dimensions of sociotechnical practices that disturbed privacy scholars in the 1980s and whose progeny, almost forty years later, magnified by the unprecedented advances of big data technologies, pose ever more vexing challenges. Decades earlier, no one was taking lightly the warnings about "Big Brother" and the surveillance state or the excessive brute powers that massive data repositories gave to those in possession of them.[21] But, in the 1980s, a new alarm was triggered by the practice then called database matching, which allowed data holders to perform computations across multiple databases. By matching a database of welfare recipients with a database of federal employees, for example, it would be possible to identify people committing welfare fraud by locating those on the federal payroll who were also claiming welfare. Celebrated by some, reviled by others,[22] database matching scared the U.S. Congress enough that they passed The Computer

---

20 *See Id.*; Yan Shvartzshnaider et al., *RECIPE: Applying Open Domain Question Answering to Privacy Policies*, *in* Proceedings of the Workshop on Machine Reading for Question Answering (MRQA) 71 (2018).

21 See, for one, Arthur Miller, The Assault on Privacy: Computers, Data Banks, and Dossiers (1971).

22 Richard P. Kusserow, *The Government Needs Computer Matching to Root Out Waste and Fraud*, 27 Comm. ACM 542 (1984); John Shattuck, *Computer Matching is a Serious Threat to Individual Rights*, 27 Comm. ACM 538 (1984).

Matching and Privacy Protection Act in 1988.[23] The intuition behind matching was captured in a vignette that Ruth Gavison relayed: a party guest is able to deduce that the person who warmly greets the priest is a murderer because that person reveals he was the priest's first confessor and the priest earlier had revealed that his first confessor owned to a murder.[24]

By standards of the present day, both these cases seem quaint, but they provide an initial framing for the construct of a *data food chain* (or *information food chain*[25]). By analogy with a food chain — a hierarchy, or an arrangement of organisms in the order of predation, placing higher on the chain those that depend on those lower as a food source — a data food chain is a hierarchy in which data of a higher order is a function of data of a lower order.[26] Matching the two federal databases revealed someone to be a fraudster as a function of being both a welfare recipient and a federal employee, and a murderer is exposed as a function of being a confessor and being a first confessor. In each case, the data of interest is above, in the hierarchy, the data in hand from which it is derived.

Computer matching, Sherlock Holmes, Gavison's priest and confessor, and experimentalists in controlled scientific settings may utilize a range of reasoning principles to infer legitimately from observed phenomena to new knowledge — deductive, inductive, empirical, and semantic. At some level, big data technologies enable the same, that is, reasoning from observed phenomena to new knowledge, only now occurring at unprecedented scale and analytical complexity.[27] Its powers allow those who possess it to employ advanced mathematical and statistical techniques over voluminous assemblages of data from diverse sources to learn more about individuals than is readily observed. For privacy scholars and other social critics, the remarkable potential to draw knowledge and power far beyond recorded data in hand posed threats to privacy that they have attempted to describe according to their preferred ethical, political and legal concepts, but unlike the practices of computer

---

23 *See* Regan, *supra* note 18, at 92-108.

24 Ruth Gavison, *Privacy and the Limits of Law*, 89 Yale L. Rev. 421, 430-31 (1980).

25 Although there are some who insist there is a distinction to be drawn between data and information, I am not convinced the various attempts at so doing are helpful. I'm even more skeptical of those who add knowledge to that continuum because, in my view, knowledge describes a person's state of mind.

26 Food Chain, Merriam-Webster Dictionary, https://www.merriam-webster.com/dictionary/food%20chain (last visited Aug. 23, 2018).

27 For an early discussion, see Viktor Mayer-Schonberger & Kenneth Cukier, Big Data: A Revolution that Will Transform the Way We Live, Work, and Think (2014).

matching and Sherlock Holmes, mapping dimensions of big data practices onto these societal concepts required intensive scrutiny of the technologies (and sciences behind them) and a level of understanding that strained non-experts (including myself).[28]

### A. Up and Down the Data Chain

One uncomfortable privacy truism, which CI countered, had this form: if fellow shoppers are able to see the contents of your shopping cart, or members of the public able to access public records about you, surely you have no privacy interests against anyone, data brokers, for example, from sweeping together such information. For me and other frustrated privacy researchers, public surveillance, that is, the trawling of so-called public facts or facts from public places, was the canary in the coal mine.

While entrenched accounts were resigned to this truism, it ran counter to an intuition that met these practices with indignation. Insisting that privacy required consideration of five key parameters, CI offered a systematic way to distinguish among cases that previously had been seen as the same; in so doing, not breaking from intuition but providing theoretical fortification to it. CI's claim is that analysts were ignoring parameters whose values varied across the different cases. In other words, it mattered whether it was a data broker who observed you rather than a fellow shopper or the cash register operator, that it was a fellow protester noting your presence at a political rally rather than a law enforcement surveillance camera recording it. CI provided a framework for asserting that ten different people observing the contents of your shopping cart on ten different occasions is *not* the same, from a privacy perspective, as one party recording and storing it; and a Google Street View car capturing the image of your home is different from a passerby observing or even taking a photograph of it.

These are cases that illustrate what goes wrong when different parameters are elided and actions and practices pass privacy scrutiny that under keener examination should not. Not acknowledging that privacy expectations shift, for example, when the data recipients change, even when the information types remain the same, illustrates one source of this fallacy, or when information

---

28 I stand with those who hold that societal, ethical, political, and legal concepts that may have been more-or-less adequate for these precursors are grossly ineffectual for the information trawling across populations that enables analytical derivation of further information. Without pretending to understand, let alone explain these analytical methods, they stood out to me, even in the early 2000s, as a class of challenge to privacy that any successful theory would need to address.

type stays steady but flows under different transmission principles, another. And so on. Without belaboring the point, I will not cycle through the many permutations that CI can generate, thereby revealing the many ways digital technologies (and associate practices) may threaten contextual integrity by disrupting information flows to which other conceptions of privacy are blind. This Article, in contrast, focuses attention mostly on one parameter alone, namely data type (topic, content). Employing the metaphor of a data food chain, the Article reveals how big data technologies expose a fissure in the regulatory landscape that allows troubling practices to escape control. Whereas, to date, CI has been able to contain some of the practices that have vexed other approaches to privacy, data analytics based on machine learning creates a class of challenges equally vexing to CI. Understanding and responding to these challenges is germane to the success of GDPR's ambitious requirement of "data protection by design and by default."[29]

To set the stage, it will be useful to assemble a handful of cases — many familiar — to illustrate some of the challenges before us in mapping privacy onto the data food chain. As noted, data science and technology, including predictive analytics, KDD, ML and AI, has enabled powerful inferential transitions from data in hand to data of interest. Quantities of available data have grown by orders of magnitude with sources such as the Internet and Web, and the rise of born-digital transactions and recordkeeping in virtually all spheres of life, including healthcare, human resources, courts, and education adding to those with a longer history of digitization such as banking, insurance, and commerce.[30] Although it has been decades since small businesses (even farmers) transitioned core functions to computers, the adoption of web and cloud-based services has meant that the data they generate flows into the data pipeline (presumably under transformations that offer protections), flooding data pools, just as credit card data, itemized communications records, and online shopping did in the recent past. A prevailing political economy that is lax — or one might say, friendly — in its regulation of the information industries has allowed the consolidation of data into massive centers, ultimately funneled into the hands of relatively few proprietors.

---

29 Council Regulation 2016/679 2016 O.J. (L 119) 1 (EU) [hereinafter GDPR].

30 In 2012, Anjul Bhambhri, IBM Vice President of Big Data Products, estimated that the world produces 2.5 quintillion bytes of data every day. This statistic is frequently reiterated in more recent big data reports and infographics. *See* Anjul Bhambri, *Looking for Data Scientists from Within — Start with Marketing*, Dataversity (July 25, 2012), http://www.dataversity.net/looking-for-data-scientists-from-within-start-with-marketing/;Ibm, *10 Key Marketing Trends for 2017*, IBM https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN (last visited July 30, 2018).

Consider some of the paradigmatic ways that analysts may transition up and down the data chain. One paradigm is the discovery of regularities (clusters, correlations) in an aggregate dataset with diverse data fields, in turn enabling discovery of data about subjects who are not represented in the initial set. Another is learning more about those who are in a given dataset than was originally provided in the data fields. There is also the paradigmatic move of combining data in-hand with data or knowledge that is external to the set but readily available, in an effort to learn about data subjects beyond what one already knew. The fictional detective, Sherlock Holmes, infers that a person of interest was the murderer on the basis of a speck of mud on his shoes or the dogs' not barking (not, however, by deduction, as he so often asserted); Netflix predicts the likely interest a person may have in a movie based on past movie ratings; Facebook is able to infer that a person is gay based on his or her situation in a social network;[31] a border agent is able to infer that people in question are U.S. citizens when learning they were born in the United States; healthcare providers may infer those at high risk of heart disease based on physiological measurements and lifestyle choices.[32]

Also among them is one of the most publicized cases, the "Target-pregnancy case," in which data analytics was performed over purchase data in order to identify pregnant customers.[33] Its shock value came partially from its flouting conventional wisdom in relating pregnancy to attributes seemingly unconnected to it, and partially from the sobering news that ordinary department stores had the power to uncover personal and intimate information. Similarly head-scratching was the oft-cited correlation between creditworthiness and the purchase of premium birdseed,[34] which, similarly to the Target case, demonstrated shocking connections between data types from disparate ontologies, miles apart. In a similar vein are numerous demonstrations of the past decades that have shown how identity can be derived from supposedly anonymized datasets either by

---

31 Josh Halliday, *Facebook Users Unwittingly Revealing Intimate Secrets, Study Finds*, The Guardian (Mar. 11, 2013), https://www.theguardian.com/technology/2013/mar/11/facebook-users-reveal-intimate-secrets; Heather Kelly, *Facebook 'Likes' Can Reveal Your Secrets, Study Finds*, CNN (Mar. 11, 2013), https://edition.cnn.com/2013/03/11/tech/social-media/facebook-likes-study/index.html.

32 *See* Carter Jernigan & Behram F.T. Mistree, *Gaydar: Facebook Friendships Expose Sexual Orientation*, First Monday (Oct. 5, 2009), http://www.firstmonday.dk/ojs/index.php/fm/article/view/2611/2302.

33 *See* Charles Duhigg, *How Companies Learn Your Secrets,* N.Y. Times (Feb. 16, 2012), https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

34 *See* Charles Duhigg, *What Does Your Credit-Card Company Know About You*, N.Y. Times (May 12, 2009), https://www.nytimes.com/2009/05/17/magazine/17credit-t.html.

triangulating fields within a single database or cleverly matching datasets in hand with data readily available beyond them.[35]

These cases illustrate movement up the data chain as data analysts reason from some data to some other data, warranted by a developing class of inferential principles that includes logical, semantic, empirical, statistical, probabilistic, and so on. In terms of the data chain metaphor, these inferences are transitions upwards as lower-order data is synthesized according to logically, scientifically or mathematically warranted pathways to produce higher-order data. Or, said another way, higher-order data is derived from, inferred from, or constructed from lower-order data. This statement avoids circularity because *lower* and *higher*, used as I have, indicate a relation between data fields (or attributes) and are not fixed labels for specific types of data.

I want also to draw attention to cases of ascending the data chain that on their face are different but, in my view, importantly similar. I have in mind such cases as consumer profiles created by data brokers for purposes of marketing, derived — they often boast — from billions of data points, and assigned labels, such as "Bible lifestyle," "Affluent Baby Boomer," "Rural everlasting," and "Urban Scramble," purportedly offering a heightened ability to sort large populations into categories that are more or less likely to respond to, or deserve, targeted offers, including advertising.[36] These types of cases will be familiar to those who follow discussions and literatures of the privacy world (academics, professionals, regulators, industry, media, *etc.*) and know that marketing is not the only domain in which categories are created in service of expressed needs, goals, or purposes. Thus, the Transportation Security Administration (commonly, TSA) creates no-fly lists based on numerous data inputs; auto insurance companies ascribe numeric ratings for more or less desirable customers based on attributes, presumably including demographics and driving history; hiring decision support systems both define and identify

---

35 *See, e.g.,* Arvind Narayanan & Vitaly Shmatikov, *Myths and Fallacies of "Personally Identifiable Information"*, 53 Comm. ACM 24 (2010); Latanya Sweeney, *Discrimination in Online Ad Delivery*, 56 Comm. ACM 44 (2013); Michael Barbaro & Tom Zeller Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. Times (Aug. 9, 2006), https://www.nytimes.com/2006/08/09/technology/09aol.html; Alessandro Acquisti & Ralph Gross, *Predicting Social Security Numbers from Public Data*, 106 Proc. Nat'l Acad. Sci. USA 10975 (2009).

36 *See, e.g.,* Craig Timberg, *Brokers Use Billions of Data Points to Profile Americans,* Wash. Post (May 27, 2014), https://www.washingtonpost.com/business/technology/brokers-use-billions-of-data-points-to-profile-americans/2014/05/27/b4207b96-e5b2-11e3-a86b-362fd5443d19_story.html.

employable candidates;[37] banks decide on mortgages on the basis of a Fair Isaacs Co. or more commonly, FICO score, or other functionally equivalent scores.

This second class of cases also involves synthesizing higher-order from lower-order data, but with a difference that may be easier to highlight with a concrete comparison between, say, heart disease and rural everlasting. Starting with heart disease, let us assume that reliable studies have shown a persistent relationship with factors such as blood pressure, cholesterol level, weight, and lifestyle choices, such as exercise, smoking, and drinking alcohol. From these, physicians may infer that a given patient either suffers from heart disease or has a high likelihood of suffering from it in the future. In this case, we have a previously labeled physical attribute (possibly discoverable by other means) that is inferable from a cluster of attributes associated with it. *Rural everlasting,* by contrast, is a profile category that data brokers have constructed, presumably applying clustering techniques across a range of attributes to discover that older, less educated people of low net worth living in rural areas correlated with characteristics of interest, such as purchasing power (wealth, salary, *etc.*) and relevant consumption patterns. There is no preexisting concept or natural class to which the newly minted label *rural everlasting* refers; instead, this invented label for older, less educated people of low net worth living in rural areas is thought to capture a category that will be useful for or relevant to marketing.

Both heart disease and *rural everlasting* involve data analytics performed over large datasets in support of inferential moves up the data chain. They are different because the former labels the higher-order concept with a familiar natural language term, while the latter is an artificial construct of the data broker. Artificially constructed concepts can serve valuable functional purposes. In the medical context, for example, syndromes, such as Reye's Syndrome, Adams-Stokes Syndrome, Irritable Bowel Syndrome, to name a few, identifying a common clustering of symptoms with or without known causes, have proliferated.[38] Over time, artificial constructs that take on a life of their own, enter the mainstream, and accrue independent meaning may come to resemble natural counterparts. As Barocas and Selbst have pointed out in their important article, however, even when such terms become normalized,

---

37 *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact,* 104 Calif. L. Rev. 671 (2016).

38 *See, e.g.,* Nitesh V. Chawla & Darcy A. Davis, *Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework*, 28 J. Gen. Internal Med. 660 (2013).

a critical stance on how their meaning has been construed and how these meanings serve some interests over others is vital.[39]

Differences between naturally emerging concepts and purposefully constructed concepts are less clear from a privacy perspective. The tenets of CI, in particular, would discourage drawing a bright line between artificially created constructs, such as rural everlasting and Irritable Bowel Syndrome, and natural concepts, such as heart disease and wealthy, because both gather individuals into meaningful classes for purposeful, targeted, differential attention, be it decision, action, reward, or punishment. For those identified as having or likely to have heart disease, this differential attention might involve supplementary preventative treatment, higher insurance premiums, or reduced employment prospects; for those identified as *rural everlasting*, it might mean gains or losses of certain ads, a reduction in special offers and discounts, or differential pricing.

One final point worth noting, before turning attention to implications for privacy, is the support provided by analytics (and machine learning, AI) for differential treatment even without interpolating meaningful concepts — natural or artificial. Analytics based on ML can support the clustering of populations based on learned vulnerabilities to targeted, disparate treatment in the interest of data processors (or their customers) without needing the middlemen, as it were, of any concept. Wading into this issue would take us into speculative territory that is beyond the scope of this Article (and my expertise) but is certain to be of increasing importance as the implications settle in of the tools of machine learning and AI applied over vast data holdings by parties whose interests diverge from our own.

Turning now to implications for privacy, the theory of contextual integrity predicts that unease, if not protest, is likely to follow at the heels of applications of big data methods (analytics, ML, *etc.*) that enable the discovery or derivation of data previously unknown. After all, transitions up the data chain may result in the flow of data to parties — in CI terms, *recipients* — contrary to entrenched informational norms. It is little comfort to be told that the lower-order data either was "public," meaning captured in public, available in a public database, not marked as protected or sensitive, or already legitimately in the hands of the data processors in question. In the "gaydar" case, for example, individuals who accept that Facebook *knows* (in a manner of speaking) whatever data they have chosen to post as well as their network of friends may still, at a minimum, be surprised and even be appalled that it *knows* their sexual orientation. They may accept that Target holds information about their purchase history but be shocked and uneasy that it *knows* they are pregnant. Why is it, then, that

---

39 Barocas & Selbst, *supra* note 37, at 698.

although common sense, as it were, tells us that these practices are wrong, a decisive case against them has remained elusive?

It has been impossible to break the chokehold over privacy regulation of a prevailing logic that allows data predators to assemble what they need in order to derive valuable higher-order data content. Key premises include, first, a presumed entitlement to the lower-order data obtained in the prevalent model of an implicit *quid pro quo* — customer, consumer, or user data in exchange for services. Such entitlements are unilaterally asserted by service providers via privacy policies, which remain the dominant vehicle for defining the terms of first-party acquisitions of data, despite overwhelming evidence that they fail hopelessly to serve the purported underlying rationale of subject control over information. Second, a political economy that refuses to modulate these assertions of entitlement underwrites the treatment of data holdings as property and allows an unchecked marketplace in human data with little regard for content or provenance.[40] This means that if Party A holds data assets that Party B wants, for example, revealing to B that X suffers from heart disease, there are many legally sanctioned routes for obtaining it. Among them are buying, selling, or reciprocal sharing of data with the option of corporate takeover, so widely practiced that a clause allowing this appears in all but a few privacy policies.[41] Not only has this allowed information and media service incumbents to swell their data assets, but it has also created fertile grounds for the flourishing of a huge, diverse data broker ecosystem.[42]

A third premise is that if a company manages, let us say legitimately, to assemble data, over which it performs sophisticated algorithmic learning to ascend the data chain and infer (derive, discover) higher-order, presumably more incisive, valuable, revealing, surprising, complex, *etc.* data, the company has not compromised privacy. The critic who dares to challenge this premise will suffer the disapprobation, wrath, scorn, and ridicule of defenders citing the First Amendment, the right to think and reason, to perform research, and to "use and enjoy" the data, as any party is entitled to do with what it owns. This

---

40 In a few sectors there are explicit rules, for example HIPAA privacy rules, which constrain actions of "covered entities." The Health Insurance Portability and Accountability Act of 1996 (HIPAA), Pub .L. No. 104-191, 110 Stat. 1936 (codified as amended in scattered sections of 18, 26, 29, and 42 U.S.C.).

41 *See, e.g.,* Charles Arthur, *Google Facing Legal Threat from Six European Countries over Privacy*, The Guardian (Apr. 2, 2013), https://www.theguardian.com/technology/2013/apr/02/google-privacy-policy-legal-threat-europe.

42 *See, e.g.,* Marshall Allen, *Health Insurers Are Vacuuming Up Details About You — And It Could Raise Your Rates,* ProPublica (July 17, 2018), https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates.

applies if the company has siphoned data from public repositories (common practice for data brokers) where it is believed to be there for the taking and subject to no systematic privacy restrictions, despite evidence to the contrary that strong and systematic privacy expectations attach to information flows even held in public records, such as, for example, whether a person voted in the last election.[43] Such expectations stoke common sense disapprobation.

Accounting for the discrepancies between common sense and common practice in terms of our data food chain, the view that common practice is justifiable follows from a belief that privacy claims travel with data up the chain; in other words, the privacy constraints that apply to lower-order data stay constant under transformations that occur as they ascend the chain. According to this belief, if B, a medical insurance company, acquires the data holdings of A, a fitness-tracking social media company, it may justly apply the privacy constraints on information about X, including weight, lifestyle habits, cholesterol levels, step count, calorie intake, *etc.*, to its predictions about X's heart condition.[44] The prevailing logic erects few regulatory hurdles in the way of powerful industry actors controlling vast assemblages and, for the most part, is impotent against the assertion of lower-order privacy constraints traveling up the chain, attaching to higher-order, inferred (learned, derived) information. Common sense resists; if privacy expectations hold constant it is in the other direction. Accordingly, instead of privacy norms about toiletries and vitamins attaching to pregnancy, those applying to pregnancy should travel *down* to toiletries and vitamins. Prevailing logic would deem this perverse, but let us pause here and consider how applying contextual thinking may inform this stalemate.

In CI terms, what matters is that information of a given type flows in accordance with legitimate privacy (contextual informational) norms, unambiguously specified in terms of five parameters. If privacy norms allow flows of information about a user's sexual orientation from the user to, say, Facebook only if the user explicitly posts this information, the fact that Facebook may learn it by other means, whether from a data broker or from the upshot of sophisticated analytics, does not sweep aside the *prima facie*, normative prohibition. Similarly, body weight and lifestyle choices may flow appropriately to mobile health devices and backend social platforms; medical tests administered in clinical settings may reveal cholesterol levels and heart disease. There are different expectations of terms of flow that, justifiably, attach to the two scenarios — in the latter, an expectation that it is governed by strict

---

43 *See* Martin & Nissenbaum, *Privacy Interests*, *supra* note 15.

44 *Cf.*, Orin Kerr, *The Mosaic Theory of the Fourth Amendment*, 111 Mich. L. Rev. 311 (2012).

rules of medical confidentiality. The fact that heart disease is derivable from lifestyle choices and body weight does not necessarily transform the norms of flow governing heart disease accordingly. In other words, justifying data aggregation and derivations up the data chain on the basis of economic and technical principles, with no regard for the constraints of CI's parameters, is to confuse what is possible with what is morally defensible.[45]

According to CI, in contrast to prevailing logic (described above), a case can be made against extending privacy norms from lower-order to higher-order data. Going farther, a sound argument may even support increasing restrictions on flows of data lower on the chain in light of its revelatory significance. Those who consider this proposition absurd should bear in mind that realignments involving greater stringency, at times, have garnered broad support. The United States Social Security Number (SSN) is a case in point.[46] Although, decades before, concerns had been voiced about the potential for the SSN to become a *de facto* universal identifier, the growing call to restrict its flow was finally heeded as the role of the SSN in enabling identity-based fraud was exposed.[47] A second, more recent case is metadata, a term of convenience referring to a class of data previously considered fair game for law enforcement, such as pen register and other data about communications besides what traditionally has been called content. Following revelations in 2013 by Edward Snowden of NSA surveillance practices, there was widespread demand for additional constraints on the flow of metadata from communications providers to government agencies.[48] Outside the realm of personal information, there are useful if unlikely analogies.

It is well-known that fertilizer bombs have been responsible for many of the most heinous terrorist attacks around the world, from the Oklahoma City bombing in 1994 to the Madrid commuter train bombing in 2004.[49] One reason

---

45 *See* David Hume, A Treatise of Human Nature (David Fate Norton & Mary J. Norton eds., 2000).

46 *See* U.S. Department of Health Education & Welfare, *supra* note 11.

47 Privacy Act of 1974, Pub. L. No. 93-579, § 2(a)(1), 88 Stat. 1896 (codified as amended at 5 U.S.C. § 552a (2012)).

48 *See* Paula H. Kift & Helen Nissenbaum, *Metadata in Context - An Ontological and Normative Analysis of the NSA's Bulk Telephony Metadata Collection Program*, 13 ISJLP 334 (2017).

49 *See, e.g.,* Fox Butterfield, *Terror in Oklahoma: The Bomb; Ideas Abound, but Blocking Oklahoma-Type Bombs Is Seen as Unlikely,* N.Y. Times (May 3, 1995), https://www.nytimes.com/1995/05/03/us/terror-oklahoma-bomb-ideas-abound-but-blocking-oklahoma-type-bombs-seen-unlikely.html; Lauren Johnston, *Fertilizer Used in Terror Bombs*, CBS News (Apr. 14, 2004), https://www.cbsnews.com/news/fertilizer-used-in-terror-bombs/.

for this is that key ingredients, diesel fuel or nitromethane, a cleaning solvent, and Tovex, combined with ammonium nitrate fertilizers, are relatively cheap, plentiful, and had been readily available in the United States. To give a sense of scale, the amount of fertilizer purchased for the Oklahoma City bomb, roughly equivalent to the amount needed for a 12.5-acre cornfield, would not have raised suspicion. Thus, even though we have a lethal brew derived from innocuous components, I doubt anyone would defend the making of a bomb on grounds that its components, individually, are benign, and in fact, already the Bureau of Alcohol, Tobacco, Firearms, and Explosives regulates explosive mixtures that include ammonium nitrate.[50] Without belaboring the point, it is quite obvious that the creation and distribution of a fertilizer bomb calls for radically different constraints from those we would impose on its components, because the former serves reprehensible purposes with dreadful outcomes. Although this is an extreme example, it illustrates the point that the moral standing of a construct does not derive from its constituent parts; instead, there may be many alternative factors, in this case including intent and dire consequences of its use.

I chose this case for its dramatic value, but there was another reason as well. Not only is there a marked disconnect between respective normative judgments of the components and the construct, but the case also illustrates a push downward from normative adjudication of the construct to the components. Thus, in 2011, the Department of Homeland Security issued proposed regulation of the sale and transfer of ammonium nitrate fertilizers, which had previously been freely available for purchase in the U.S. Unlike some counterparts in Europe and Asia, it did not include an outright ban on sales but includes a registration requirement for anyone purchasing above a set amount.[51]

Returning to the data food chain and contextual integrity, the point I wish to establish is neither that privacy norms travel up with inference, nor down from inference, but that a normative evaluation of the subsequent data flows and uses is independent of the logical (mathematical, statistical, algorithmic) principles enabling movement up the chain. As previously discussed, evaluating data flows, in the first place, requires establishing departures from expected flows in terms of shifts in one or more of the CI parameters followed by

---

50 *See* Associated Press, *U.S. pushing new rules to curb fertilizer bombs*, CBS News (Aug. 2, 2011), https://www.cbsnews.com/news/us-pushing-new-rules-to-curb-fertilizer-bombs/.

51 *Id.* For the Department of Homeland Security's proposed program regulating ammonium nitrate sales, see U.S Dep't of Homeland Sec., Ammonium Nitrate Security Program (ANSP) https://www.dhs.gov/ammonium-nitrate-security-program (last visited Aug. 22, 2018).

comparative assessment in terms of interests, values, and contextual purposes. The challenges to this process in the contemporary data-AI environment should not be understated, because novel information types and newfangled actors are emerging at a dizzying pace, sometimes even transforming the contours of contexts themselves. If one considers the mobile health arena alone, it involves several new layers of data and service intermediaries: not merely a handful of new recipients, not only new types of data, but also data in new forms, such as continuous capture of heart-rate; step counters that are able to record features of a person's gait and incipient musculoskeletal problems, or infer the nature of his or her activities.[52] This means that it is not always straightforward and may even be impossible to locate the relevant entrenched norms against which to compare the novel flows. Although this leaves CI without a quick assessment option, it fares no worse than other accounts of privacy, or societal values generally, in which social norms play a role.

Defending on ethical ground practices that give rise to unprecedented data flows requires CI's three-layered analysis: harms and benefits to affected parties, impacts on ethical and political (*i.e.*, societal) values, and impacts on the attainment of contextual ends and values. In the familiar case of targeted advertising, the derivation of user profiles from online behaviors and other data serves the interests of advertising networks while arguably undermining the wellbeing of data subjects. The practice, further, has been shown to unfairly discriminate on grounds of gender[53] and race.[54] To these historically significant categories, I would add synthetic categories, such as rural everlasting, undeservedly targeted for differential treatment. Finally, growing empirical evidence suggests that targeted advertising, particularly price discrimination, lessens consumers' trust in online commerce, arguably undermining the values and goals of the commercial marketplace.[55]

---

52 *See* Natasha Singer, *How Big Tech is Going After your Health Care*, N.Y. Times (Dec. 26, 2017), https://www.nytimes.com/2017/12/26/technology/big-tech-health-care.html; Daisuke Wakabayashi, *Freed form the iPhone, the Apple Watch Finds a Medical Purpose*, N.Y. Times (Dec. 26, 2017), https://www.nytimes.com/2017/12/26/technology/apple-watch-medical-purpose.html.

53 *See* Amit Datta, Michael Carl Tschantz & Anupam Datta, *Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination*, Proc. Privacy Enhancing Tech. 92-112 (2015), https://arxiv.org/abs/1408.6491.

54 Sweeney, *supra* note 35.

55 *See* Kirsten Martin, *The Penalty for Privacy Violations: How Privacy Violations Impact Trust Online,* 82 J. Bus. Res. 103 (2018); Zeynep Tufecki, *We're Building a Dystopia Just to Make People Click on Ads*, TED (Sept. 2017), https://www.ted.com/talks/zeynep_tufekci_we_re_building_a_dystopia_just_to_make_people_click_on_ads.

The practical upshot of analysis and evaluation is to suggest adjustment and calibration between novel practices (and associated data flows) and informational norms governing flows, so that deleterious consequences may be minimized. CI offers far more nuanced alternatives than simply "ban access!", "ban flow," or "give subjects control!" Dissemination of data may be channeled to a judicious selection of recipients, or constrained by innovative transmission principles, including well-modulated authorization strategies, or auditing, as was proposed for the distribution of ammonium nitrate. In "Tragedy of the Data Commons," Jane Yakowitz implicitly searches for more nuanced alternatives when she advocates for legal, even criminal, sanctions against researchers who re-identify anonymized datasets, even when they could.[56] Surprising to some readers of the Article, generally critical of privacy-based constraints on data flows, its proposal usefully, if inadvertently, lends support to the thesis that the ability to extract information does not justify its extraction, in practice. One could argue about Yakowitz's paradigmatic cases that trust of researchers is paramount to the success of the research enterprise and practices that diminish this trust may not only harm data subjects but also undermine the enterprise itself, a thoroughly contextual argument.

Summarizing the key element of this section, I have observed that actions and practices that flout expectations and pose existential threats to privacy as a legitimate claim and right have become the new normal because they are opaque to examination, unchallenged by a dominant political economy, and not grasped for what they are by a regulatory framework that fails to distinguish technical from ethical merit. This is not to say that big data technologies (ML, AI, *etc.*) are an easy target; I must acknowledge the harsh challenges they pose to the original heuristics of contextual integrity.[57] Radical changes in the nature of data and surrounding technologies outpace the capacity to proceed according to the discrete steps of the CI heuristic — first to locate divergences in entrenched norms measured in terms of the five parameters and then to evaluate them. However, despite the need to relax and revise the mechanics of the heuristic, I remain convinced that a) mapping the data flows in terms of the five parameters and b) evaluating these flows, to the best of

---

56 Jane Yakowitz, *Tragedy of the Data Commons*, 25 Harv. J.L. & Tech. 1 (2011).

57 The capacity to draw unexpected inferences is the central concern of an ongoing research project with Sebastian Benthall, Anupam Datta, and Michael Tschantz. We are developing a construct of "origin," to supplement CI for a world when "information type" might no longer function reliably to produce adequately protective constraints on flow. This project is titled "Origin Privacy: Protecting Privacy in the Big-Data Era" and is funded by the Defense Advanced Research Projects Agency (DARPA) Brandeis Program.

our knowledge, in terms of interests, values, and contextual ends — messy, uncertain, and burdensome as it may be — remains an ideal worth striving for.

The Facebook-Cambridge Analytica incident sounded a sharp warning bell, but its impact will dissipate if we fail to grasp the full range of its lessons. Yes, it exposed Facebook's reckless opportunism and Cambridge Analytica's venal ambition, and yes, information was shared without explicit consent. Yet even this case, which more than almost any other caused public trauma and exacerbated broad distrust of information and platform industries, does not stray very far from the mandates of the three premises of the prevailing logic discussed above. Contextual integrity, by contrast, forces us to confront the reasons for people's shock and disgust to measure the distance between actual and expected practice. It mattered that the betrayal was not for commercial gain alone but, in this instance, for the purpose of manipulating voters. Finally, it underscores the importance of respecting privacy not to protect the interests and claims of individual data subjects, alone; the actions of Facebook and Cambridge Analytica should be understood, in addition, as compromising contextual ends and values, in this case the aims and functioning of democratic governance, no less.

### B. Data Primitives

In this concluding section on the data food chain, I take a final plunge down to what I call "data primitives" and the distinctive challenge they pose to contextual integrity. Although it is not newly sprung in the wake of big data technologies (AI, *etc.*), the burgeoning of smart, connected devices and appliances, that is, "things" in the so-called Internet of Things (IoT), has pushed this phenomenon to center-stage. Like "big data," "IoT" functions more as a marketing concept than a term with scientific coherence; here, however, we finesse some of the term's messiness and find it a useful shorthand to cover a broad array of disparate networked devices and systems, from thermostats to "smart" luggage and Weber Grills, self-tracking, and mHealth wearables that register and record activity and a host of bodily functions.[58] It can refer to conventional household appliances but also to novel in-home systems such as Amazon's Echo or networked sensors that can detect a range of environmental factors. I see no reason not to include mobile or "smart" phones and the myriad of mobile apps they host and mediate, including specialized physical devices from baby-cams to smart forks, and others that draw on telemetry data generated by the mobile phone itself (*e.g.*, gyroscope, light, *etc.*).

---

58 *See* Philip N. Howard, Pax Technica: How the Internet of Things May Set Us Free or Lock Us Up (2015).

Less important than defending the exact borders of IoT is establishing the types of data to which it draws attention. IoT systems and devices render into digital, recordable form ("datafy")[59] and send down the chute into the data pool the ephemera of everyday life, dimensions of experience and expression, affect, sound, text, image, video, type and strength of network connections, mouse clicks, location, mobile telemetry and other metadata, biometric markers, and even brainwaves. Before we have text, a photo, a place, a shoe order, or a social network, we have mouse clicks registered as digital (electric) pulses, environmental phenomena (temperature, airborne chemicals, *etc.*) and biological features rendered as sensor signals, as mathematical templates, and metrics, flowing via digital networks to software platforms. We have electrical signals passing from transmitters to transceivers, activated pixels producing digital images, and geospatial coordinates communicated from satellite to GPS-enabled devices. These event imprints, the base-layer of the informational universe, are what I am calling, *data primitives.*

In a previous article, I offered the following illustration.[60] Imagine you are walking on the beach and you see an impression in the damp sand along the water's edge. The impression, one could say, is data captured in the sand, as the electrical pulse is data captured by a computer's operating system. As you approach, you see that it's a footprint, likely the footprint of a man, with a long middle toe. Someone with tracking skills might see more; that it's the footprint of someone running or someone with a limp. He might even recognize it as the size 10 footprint of a recently escaped prisoner. And so on. The primitive, in this case, is the impression in the sand. It moves higher in the chain as it acquires meaning and significance, often through context, *e.g.*, escaped prisoner on-the-loose. Examples from the digital world abound: a pattern of pixels is an image of a suspect on video surveillance footage; a word or phrase is a web search term or an element in a letter, email, or novel; a timed series of location coordinates recorded by a GPS device is the route driven from home to work.

Why are data primitives problematic for the theory of contextual integrity? As we've seen, CI's fundamental commitment is to appropriate flow based on substantive, informational norms. The social norms we live by — *real*

---

59 *See* Katherine J. Strandburg, *Monitoring, Datafication, and Consent: Legal Approaches to Privacy in the Big Data Context, in* Privacy, Big Data, and the Public Good 5, 5-43 (Julia Lane et al. eds., 2014).

60 I developed this example for my paper on privacy, collection, and use regulation. *See* Helen Nissenbaum, Deregulating Collection: Must Privacy Give Way to Use Regulation? 9 (May 1, 2017) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3092282.

norms versus theoretical or mathematical constructs — are messy along many different dimensions, but their capacity to regulate behavior, mold institutions, and shape the built environment (including technology) suggests that they are understood by people; they have meaning and we grasp it. The cultural or historical stories about sources of the informational norms we live by are likely to be as varied as any we might tell about the sources of social and cultural norms and their evolution — custom, religion, law, ethics, practical necessity, even fashion and whimsy. The legitimacy of norms is another matter; we may trust the wisdom of the ages, or we may decide an evaluation of convention is warranted.[61] CI is interested in assessing whether entrenched informational norms stand up to ethical scrutiny and that, too, requires taking stock of the *meanings* of the parametric values and their significance for interests, values, and contextual purposes. To be more concrete, it matters that *Recipient* is *Doctor* and not simply *Person A* because *Doctor* is a capacity that has rich significance, including function, training, obligation, *etc.* Similarly, it matters whether *Information* is *Address*, or *Heroin Overdose*, or *Muslim*, just as it matters that police acquired evidence *With a Warrant* and not willy-nilly. The challenge posed by data primitives is that they present to us under descriptions that do not engage our privacy norms. It is only when we understand their significance, or meaning, that we can map them onto norms.

Concrete thinking may be helpful. A digital pulse is registered in an operating system. It requires an interpretative act, albeit a modest one, to register it as a mouse click and further to situate it within the browser, or at a given website, and beyond this, whether I have ordered a pair of shoes or searched WebMD for syphilis.[62] Each transition is a hop up the data chain, not necessarily ending with shoe order but potentially even beyond this, as profilers imbue the purchase with additional meaning, situating it within other data, inferring a predictive pattern. The initial pulse, that is, the data subject clicking the mouse, is unlikely to trigger a privacy norm; but the shoe order, or acceptance of terms of service, placement of a marker in a document, expression of interest in an ad or search term or medical condition, placement of the subject in rural everlasting, or assignment of an insurance risk score, *etc.*, gives the click significance and, in all likelihood, engages a norm.

If one agrees that CI's strength is its framing of privacy expectations in terms of a rich semantics, this strength could be viewed as a liability when challenged with the flows (capture and transmission) of data primitives. CI insists that only when senders, recipients, and subjects are conceived as

---

61 *See* Nissenbaum, Privacy in Context, *supra* note 1, at 162-64.

62 No surprises here: these connections are precisely what systems architects and software engineers have set out to achieve.

actors in contextual capacities (roles), and only when the attributes (data or information types) are specified according to contextual ontologies, the norms prescribing and proscribing flows are meaningful to people. Only in these terms can we grasp the significance of data flows for society and forge regulation that is conducive to the interests of affected parties and to contextual (hence social) integrity. It seems to follow that being unable to capture data flows in semantically rich terms, CI is at a loss as to whether they respect or violate privacy.

Before explaining why CI needs to address this challenge, it is worth taking a brief sideways glance at other theoretical and regulatory approaches from which, earlier, I had set CI apart. If these approaches fare better under the challenge from data primitives, it makes more sense to revert to them than to prolong the struggle. Thesis 1, for example, differentiates CI from definitions of privacy as the withholding of information — as secrecy. These definitions, by contrast, do not need to distinguish between data primitives and semantically rich data, for all that matters is suppressing the release of data emanating from a given, identifiable individual. Likewise, Theses 2 and 3 differentiate CI from procedural definitions, such as privacy as compliance with FIPPs or as control over information about oneself, which also do not distinguish between data primitives and semantically meaningful data. Although, on the face of it, these alternatives sidestep the challenge, a closer look indicates they may, in fact, be kicking it down the road.

Take secrecy or control. A major challenge is scope, because no matter how committed one is to these definitions, none of them are defended as an absolute right over all data and in all circumstances. Indeed, much of the literature and public deliberation on the topic of privacy is devoted to establishing the legitimacy of privacy claims in the face of other, competing claims. It is impossible, in my view, to advocate on behalf of particular privacy claims, or types of claims, without explaining why certain flows or uses of data deserve protection while others do not and, in so doing, accounting for their impact or significance in meaningful terms. Accordingly, when advocates argue that protection is warranted for, say, sensitive information — as often they do — they are bringing into play assumptions about the topic or substantive nature of data.

Without belaboring the point, even if one's formal definition of a right to privacy does not require semantics for data, defending the *ethical* force of privacy claims will require an account of data flows in terms of whatever moral currency matters, whether economic harm, embarrassment, injustice, autonomy, freedom, *etc*. In turn, this will involve imbuing the data in question with meaning and significance that people (and institutions) grasp — in other words, a semantics — so that we, the data subjects, and they, societal

regulators, can become aware of what is at stake and adjudicate accordingly. Taking the position, according to versions of privacy-as-control, that subject consent, or choice, makes flows of data right, no matter its place on the data chain, makes a mockery of choice. Choosing is not mere picking but requires that the subject understand that to which he or she is consenting, which is lacking in our interactions with data primitives, defined so precisely because they are absent of meaning.

Returning to the question why data primitives are a pressing concern for CI (and, as I've demonstrated, other central privacy theories), the answer lies in the ascending importance of privacy by design (PBD).[63] To this end, discussions with computer scientists and engineers have sought to engage CI in an effort to pursue PBD.[64] In one case, a perplexed manufacturer of "smart" systems for controlling lights and temperature in large buildings insisted privacy could not be designed into the system because the data in question was merely that produced by temperature, light, and motion sensors. Even though this data was responsive to people in a room, or in the building, it did not seem relevant to any conventional notions of privacy. Yes, moving up the data chain from these primitives, the potential exists to infer where people are, individually and collectively, how many in a room, whether animal or human, and so forth. More can be inferred if we pair these primitive sensory inputs with other information, such as people who work in the building, office numbers, *etc*. Although these might appear to be "tame" concerns, they have been noted by both proponents and critics of "smart grid" systems, which may reveal to networked thermostat providers and utility companies quite a lot of information about goings-on in a home.

Variations of these questions arise for conscientious developers[65] of IoT devices and systems, which choose PBD rather than leave privacy as an afterthought. The question they face is what principles exist for limiting ("minimizing") the flow (collection, storage, onward sharing) of primitive data when they know that the potential exists for deriving data higher in the data chain that is likely to trigger contextual norms. Even more perplexing is what responsibilities they may have when, at the time of design, whether such higher-order data is derivable and what that might be is neither known nor predictable? Another real world case involves machine learning over smartphone telemetry

---

63 GDPR, *supra* note 29.

64 Sebastian Benthall, Seda Gürses & Helen Nissenbaum, *Contextual Integrity through the Lens of Computer Science*, 2 Found. & Trends® Privacy & Security 1 (2017).

65 Mary Flanagan & Helen Nissenbaum, Values at Play in Digital Games 11-13 (2014).

data to predict whether users suffer from the personality disorder, "impulsivity," defined as "behavior without adequate thought, the tendency to act with less forethought than do most individuals of equal ability and knowledge, or a predisposition toward rapid, unplanned reactions to internal or external stimuli without regard to the negative consequences of these reactions."[66] Here, too, the challenge is what guidance contextual integrity is able to provide for such data, *e.g.*, readings from gyroscopes and accelerometers, rates of reaction and reaction time responding to texts and other pings, light, motion, *etc*., when we know that this data, though primitive, may yield semantically rich information when interpreted, such as how late you stay awake, and when aggregated and processed yields may reveal a personality disorder.

Without offering the depth of analysis it deserves, I will end this section with a brief mention of location privacy, which is mired in confusion, in my view, due to the poorly grasped disconnect between data primitives and meaningful attributes. Progress has been made in the realms of regulation and design, recognizing location as a distinctive attribute requiring explicit protection both because it is vulnerable to the broad class of IoT technologies and because it offers profound insight into a person's life, even though there is a growing sense of when it is and is not appropriate for a given party (*e.g.*, a mobile app) to obtain a user's location: for example, yes to Google Maps, no to a flashlight app.[67] Contrary to the approach CI would recommend, current policy allows apps to seek consent. The trouble is that most human users have an understanding of location that is quite different from its representation as a data primitive, a system or device. The operators of an app, system, or media platform that respects a user's refusal to share location data by, say, not recording GPS coordinates may pursue other, analytic means to learn that he or she is, say, in a mall, a hospital, on a road, at home, or at work.[68] It turns out that people are more attuned to appropriate flows of location data when the data are presented in terms of meaningful places (mall, hospital, *etc.*) or signify higher-order inferred data (is shopping) than as data primitives (*e.g.*, GPS coordinates), despite their greater precision.[69]

---

66 Int'l Soc'y for Research on Impulsivity, https://www.impulsivity.org (last visited Aug. 3, 2018).

67 *See, e.g.,* Robert McMillan, *The Hidden Privacy Threat of…Flashlight Apps?*, Wired (Oct. 20, 2014), https://www.wired.com/2014/10/iphone-apps/.

68 *See, e.g.,* Angela Fritz, *A Security Researcher Discovered AccuWeather App Tracked, Shared your Location — even if you 'Opt Out',* Wash. Post (Aug. 24, 2017), https://www.washingtonpost.com/news/capital-weather-gang/wp/2017/08/23/security-researcher-discovered-accuweather-app-tracks-and-shares-your-location-even-if-you-opt-out/.

69 *See* Martin & Nissenbaum, *Measuring Privacy*, *supra* note 15, at 210.

## Conclusion

CI faces a conundrum in how to approach the incommensurability of privacy norms, expressed in semantically rich, contextually meaningful terms with rules governing flows of lower-order data —particularly data primitives within and across networked systems and devices.[70] One option is to do nothing and stick with literal CI norms, until the machinations of big and computational analysis (*e.g.*, AI/ML) emerge into view in forms that higher-order, privacy-relevant concepts can grasp. There was a time when I believed that privacy could be well enough protected atomically, that is, by astutely monitoring and characterizing data flows in terms of contextual parameters and divergences from norms and expectations, case by case. Although there certainly will be occasions when this approach successfully can identify sources of privacy breach, as a general approach, it is overwhelmed by the scale of the shadow data universe. But passively waiting for clear breaches to surface is untenable.

Given a technological landscape that includes vast data holdings, data analytics, AI, machine learning, IoT, mobile devices, and other computational capacities, there is a dire need for systemic principles that will expose the material risks of the current data policy anarchy. Decisions are being made, critical actions taken on the basis of practices that defy higher-order description. These may be unjustly harmful or simply careless of individuals' interests and dangerously risky to contextual integrity with consequences for the fabric of society, including democratic governance, healthcare, and education. Although most of us are not able to anticipate all the ways in which lower-order data betray higher-order attributes, there are two points to note: First, there are stakeholders who have a pretty good idea what higher-order data is derivable from what lower-order data because they are in the business of doing precisely this, that is, going after valuable higher-order data without arousing embarrassing discovery and costly resistance; and second, ignoring clear signs that the risks and benefits are not evenly spread of sophisticated methods for inferring high-order data from data lower on the chain is a mistake. Waiting decades before imposing constraints on the flow of SSNs and failing to acknowledge the substantive value of communications metadata placed many people and important political values in harm's way.

---

70 Arguably, CI does not face this conundrum alone, but faces it together with any other approach to privacy that modulates its requirements according to variation in the meaning or significance of the data in question, such as those that label medical or financial data, for example, as sensitive and worthy of a higher standard of protection.

That was then. At present, surprising moves up the data food chain are coming at us too fast and furious to afford a decades-long learning curve for each of them. The intent of this paper has been to point to the sources of these uncomfortable surprises and to shed light on the latest round of entanglements (not the only ones) of privacy with technology. I do not have a clear set of recommendations to offer, in part, because for real world impact, there is still too much to assay — common practices, technical vanguard, industry and political affinities and stakeholders. Instead, insights from this inquiry into contextual integrity up and down the data food chain point to and reinforce high-level principles that to be effective, need to be elaborated, justified, and situated in real world contexts where knowledge of the levers of policy and technology is critical.

One important principle is that ethical norms do not stay constant under transformations up and down the data chain. It is as plausible for privacy expectations to travel down to lower-order elements as it is to travel up to higher-order constructs or inferred properties. This may apply even to data primitives as we have seen in the cases of SSNs and metadata, discussed above.

Just as important is transparency ("no unpleasant surprises"): Practices that yield higher-order properties ("home") from lower-order properties (GPS traces) should be made available for inspection, particularly in cases where informational norms governing the former are different from those governing the latter. What it means to "make available to inspection" remains an open but critically important issue to address.

Closely related is accountability, in both senses: one, beyond transparency of process, requires that account be given in meaningful, actionable terms of outcomes.[71] The other is to hold data processors to account for steps up the data chain that defy norms and expectations for violations of privacy. Further, they should be held to account for thoughtless derivations that single out individuals or pose risks to societal or contextual values (*e.g.*, voter targeting and democracy).

Lastly is the principle that the burden of proof should fall on data processors to ensure that the outcomes of the computations (*e.g.*, machine learning) meet not only self-serving ends but also produce no harm and, further, meet individual and societal expectations and produce no harm, including what I have called actuarial harms to the unlucky false positives and false negatives.[72] Intent to harm is an insufficient standard.

---

71 *See* Hildebrandt, *supra* note 4.

72 Nissenbaum, *supra* note 60, at 25.