PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Research

Cite this article: Taylor L. 2016 The ethics of big data as a public good: which public? Whose good? *Phil. Trans. R. Soc. A* **374**: 20160126.
http://dx.doi.org/10.1098/rsta.2016.0126

Accepted: 6 September 2016

One contribution of 15 to a theme issue 'The ethical impact of data science'.

Subject Areas:
theory of computing

Keywords:
big data, mobile phones, ethics, development, human rights

Author for correspondence:
Linnet Taylor
e-mail: l.e.m.taylor@uvt.nl

The ethics of big data as a public good: which public? Whose good?

Linnet Taylor

Tilburg Institute for Law, Technology and Society (TILT), Tilburg, Noord-Brabant, the Netherlands

LT, 0000-0001-7856-7611

International development and humanitarian organizations are increasingly calling for digital data to be treated as a public good because of its value in supplementing scarce national statistics and informing interventions, including in emergencies. In response to this claim, a 'responsible data' movement has evolved to discuss guidelines and frameworks that will establish ethical principles for data sharing. However, this movement is not gaining traction with those who hold the highest-value data, particularly mobile network operators who are proving reluctant to make data collected in low- and middle-income countries accessible through intermediaries. This paper evaluates how the argument for 'data as a public good' fits with the corporate reality of big data, exploring existing models for data sharing. I draw on the idea of corporate data as an ecosystem involving often conflicting rights, duties and claims, in comparison to the utilitarian claim that data's humanitarian value makes it imperative to share them. I assess the power dynamics implied by the idea of data as a public good, and how differing incentives lead actors to adopt particular ethical positions with regard to the use of data.

This article is part of the themed issue 'The ethical impact of data science'.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3515205.

THE ROYAL SOCIETY
PUBLISHING

Privacy is your right. So is access to food, water, humanitarian response. The challenge is that we see a lot of regulatory frameworks which don't have the right litmus test.

— Robert Kirkpatrick, Head of UN Global Pulse

The ethics of not acting is the ethics of being afraid of doing something wrong.

— Geoffrey Canright, Head of Data Analytics Group at Telenor Research.[1]

1. Introduction

As data analytics have evolved over the last decade, the discussion about their uses and problems has evolved to include considerations of ethics, privacy and legal boundaries. The notion of 'data as a public good'[2] [1–3] has been widely cited since the high-profile use of data from mobile phones, in particular, in the humanitarian and international development sectors. The value of digital data on low-income areas to these sectors has given rise to calls for 'data philanthropy' [4], where corporations donate data to non-profit actors through development intermediaries, such as the UN. There are some high-profile examples of data philanthropy, for instance, where Twitter has granted access to its data to UN Global Pulse [5]. However, despite their data's value to the development sector, mobile operators are proving reluctant to share call detail records (CDRs) under the kind of general principles that are being discussed by development and humanitarian actors. This paper will ask whether the argument for 'data as a public good' fits with the reality of big data, using two case studies where mobile operators have shared CDRs for development and humanitarian purposes.

There are two possible definitions of a public good that are relevant in this scenario. One is the notion that data should be made available to help international organizations promote social good in the public sphere, i.e. that data should be more public to create more good impacts from it. The second is the argument that data should be formally defined as a public good, because its potential power to fight poverty and disease and to inform emergency response is such that international institutions should have access to it. This is effectively an argument for reordering power relations with regard to digital data. The public good argument with regard to information and knowledge has been defined by information economists Stiglitz [6] and Varian [7]. In their definition, due to the low costs of reproduction, knowledge is a resource that is both non-rivalrous (there is no extra cost incurred when others use it) and non-excludable (it is impossible to keep others from using it).

However, in this definition, there are caveats with relevance to big data: Stiglitz and Varian warn that knowledge can be made functionally excludable where the private sector gains value from controlling it, and that regimes also determine the extent to which it is excludable, for example, in the form of taxes and patents. Purtova [8] brings this debate up to date, identifying digital data deriving from people as a 'system resource' comprising an ecosystem of people, platforms and profiles, and concluding that while it may be possible for knowledge to be a public good, it is not possible to make the same claim for digital data. In fact, the language of the 'personal data ecosystem' is already in use by the World Economic Forum, among others [9], to explain the ways in which the knowledge produced through digital data is inherently commercial, and operates as an interaction between individuals and firms.

This paper will build on Purtova's logic to challenge the rhetoric of data as a public good. I will show that multiple, often contradictory considerations are at work behind any corporate decision to share data *pro bono*, and that claims that data can be a public good underpin a particular politics of data and a particular power dynamic that risks empowering some at the expense of others, and that actually decreases opportunities for data to have positive social impact.

[1] Geoffrey Canright. Head of Data Analytics Group, Telenor, 17 September 2015.

[2] See https://theodi.org/blog/why-is-open-data-a-public-good.

To explore the different considerations in play with data sharing, this paper will compare the experiences of two firms that have shared mobile phone CDRs for research purposes: Telenor (based in Norway) and Orange (based in France). The research involved fieldwork, observation and interviews conducted over the period 2013–2016. It comprises 300 formal and informal interviews conducted at events relevant to the 'responsible data' movement, participation in advisory groups and in public discussions on data ethics, and scans of the media over the same 3-year period. It is also based on visits, observation and interviews conducted during 2015–2016 at the two firms that form the case studies for this paper.

2. Background: 'responsible data' and call detail records

Since the exponential growth in mobile usership in low- and middle-income countries (LMICs)[3] during the 2000s (ITU 2015), CDR data from these countries have become a potentially powerful way of understanding what is being missed by national statistics, which are often lacking due to low resources [10], and for planning and evaluating interventions in the development and humanitarian spheres. Large-scale datasets from international mobile operators are being used in the realm of big data analytics to model the spread of diseases, such as dengue [11] and cholera [12], and to map poverty [13] and mobility [14]. Data from mobile financial transactions combined with mobile surveying are also showing potential as a way to plan savings and banking interventions [15], and records derived from mobile money platforms also have significant potential value for tracking income and spending.

Development and humanitarian research using CDR data began with one-off, targeted collaborations between mobile operators and researchers in public health [16–18], which then sparked a move to create larger scale, more crowdsourced forms of data sharing for research. Spain's Telefonica hosted a datathon in 2013 that opened up CDRs to the technology community,[4] and Orange established a 'Data for Development' challenge[5] that released CDR data from its subsidiaries first in Côte d'Ivoire and later in Senegal to data science research teams worldwide. These two models, targeted collaboration and crowdsourcing research, are explored in the two case studies in this paper.

The demonstrated power of CDRs to supplement national statistics and generate new insights with regard to LMICs has created a community of researchers, funders and intermediaries interested in using them. Central to this process has been the United Nations' call for a 'Data Revolution' [19] and the World Economic Forum's claim that big data should be considered a tool for development and humanitarian action [20]. Despite a growing awareness of the power of the new sources of digital data, however, no generalizable approach to data sharing has emerged. The prevailing vision is one where data are shared through intermediary institutions whose job would be both to analyze data and to channel directly to non-governmental organizations that could use it,[6] such as the UN data science initiative Global Pulse. Proponents of this model have put pressure on mobile operators to share data through them as intermediaries, which the firms have resisted on the basis that they remain responsible for the data they collect and should therefore determine how they are used.

The community interested in creating a public good approach to data sharing has for some time been working to create ethical guidelines to underpin such a framework, recognizing that for data about developing countries to flow more freely, rules and norms are necessary to protect

[3] I use the World Bank's definitions grouping countries, see: http://data.worldbank.org/about/country-classifications, where LMICs have incomes of US$1036–$12 616 *per capita* and high-income countries above that threshold.

[4] For more information, see http://dynamicinsights.telefonica.com/blog/1008/campus-party-2.

[5] See http://www.d4d.orange.com.

[6] Robert Kirkpatrick, director, UN Global Pulse, interviewed 18 August 2014.

their populations. These concerns were epitomized by a statement by Robert Kirkpatrick of UN Global Pulse in 2012:

> Even if you are looking at purely anonymized data on the use of mobile phones, carriers could predict your age to within in some cases plus or minus one year with over 70 percent accuracy. They can predict your gender with between 70 and 80 percent accuracy. One carrier in Indonesia told us they can tell what your religion is by how you use your phone. You can see the population moving around.[7]

It is against this background that various Responsible Data groupings and discussions have emerged. These include the International Data Responsibility Group, convened by Leiden University's Centre for Innovation, the Responsible Data email list which involves influential privacy and data experts worldwide, and the series of Responsible Data events held by universities including Stanford, NYU and the MIT Media Lab.

Figure 1 shows the institutions and connections shaping these discussions, as observed over the period 2014–2016. The network is characterized by two key features: first, it is multipolar, with at least eight key institutions convening discussions (the names of institutions are shown proportionate to the institutions' betweenness centrality, so that the size of the type indicates an institution's relative importance in connecting other institutions in the discussion). Second, the institutions with the highest betweenness centrality are those who stand to gain from data's increased availability to research and policymakers in terms of their funding, status or influence. The diagram can be interpreted as a negotiation between a group of key players over what should constitute practices of responsible data sharing, with each key player largely convening its own discussion community in an effort to define the emerging frameworks and practices. Crossover between communities occurs in the case of mobile operators (Orange, Telenor and Telefonica), academic bodies and civil society organizations. Also remarkable is the extent to which the conversation focuses on mobile data, as seen by the absence (or peripheral nature so that they were invisible to this research) of other producers of data, such as governments, Internet search and social media firms. This is in line with the fact that the conversation is largely about data from LMICs, where mobile operators are the key actors in collecting and potentially sharing data.

The high-level international actors that feature in Figure 1 demonstrate that there is much to play for in this debate, where funding and status potentially attach to becoming a facilitator or intermediary. Alliances between international bodies and academic research groups bring funding and status with them, and countries also stand to benefit: for example, both the Dutch Foreign Ministry and The Hague's city administration fund Leiden's data governance discussions as a way to increase the Netherlands' international profile as promoting humanitarian innovation. When initial calls for data to be made more available did not result in large-scale data sharing, particularly by mobile operators, discussions towards the creation of ethical frameworks took centre stage as a strategy to present a coherent set of community standards that would make data proprietors more comfortable sharing data.

To create such standards, however, numerous technical and practical issues had to be negotiated across widely differing disciplinary worldviews. For example, at a meeting of the International Data Responsibility Group in 2016,[8] a representative of an organization with a substantial body of data scientists presented a decision-making tool for weighing the costs and benefits of using a given dataset. The steps were as follows: first, identify the potential risks and harms of a particular intervention, and evaluate their potential likelihood of occurring. Next, identify a potential beneficiary. Finally, identify and quantify the intervention's positive effects, and weigh them against the harms identified. For social science researchers, this decision-making process was hard to understand, because it involved balancing a general idea of potential harms

[7] Robert Kirkpatrick, interview with *Global Observatory*, 5 November 2012. Accessed online 7 July 2016 at http://theglobalobservatory.org/interviews/377-robert-kirkpatrick-director-of-un-global-pulse-on-the-value-of-big-data.html.

[8] International Data Responsibility Group meeting, 19 February 2016, The Hague, the Netherlands.

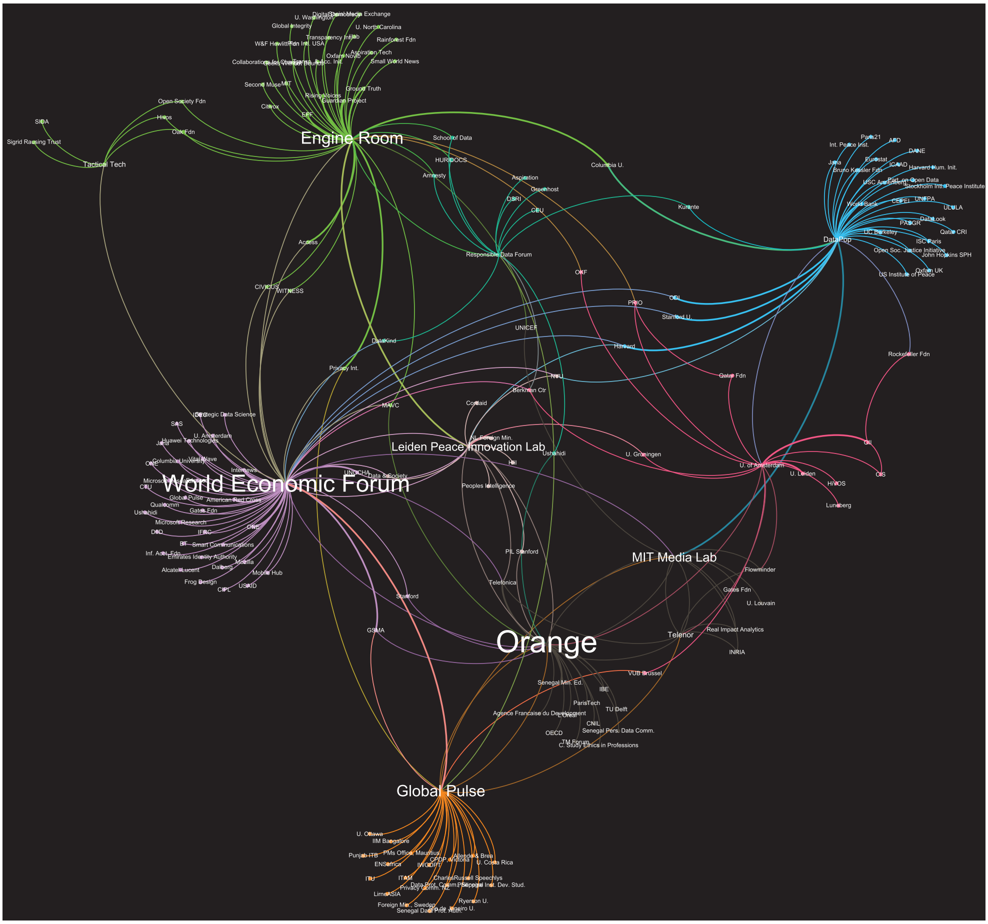

Figure 1. Network of participants in Responsible Data discussions, 2014–2016.

against a specific and contextual idea of benefits, and would always, it seemed, come out in favour of intervening and therefore in favour of the data sharing that would enable intervention.

While such a model may be ethically incomplete, it makes sense from a data scientific perspective whereby data are acquired, then mined, and a use for the data emerges from that process. This contrasts with the established social scientific process of defining a problem then using data to seek an answer. In the data-for-development and humanitarian sectors, these perspectives clash in a context of emergencies, extreme need and strong arguments for data reuse and sharing.

(a) The centrality of call detail record data to the responsible data debate

Although CDR research long predates its use in emergencies and development [14], it was the use of CDRs in Haiti after the 2010 earthquake that sparked widespread interest on the part of the development and humanitarian communities. The independent research organization Flowminder used CDRs to analyze the spread of the ensuing cholera epidemic [16], demonstrating that such data could be of use in tracking disease transmission. This was followed by several high-profile projects in collaboration with mobile operators who were interested in how they could put their data to good (and scientifically productive) use, including data challenges sponsored by the Spanish firm Telefonica in 2013 and Telecom Italia in 2015, and the Orange Data for Development challenges in 2013 and 2015.

In 2015, a media storm arose over the (non)use of CDRs for tracking Ebola [21], demonstrating how widely they had become seen as a potential magic bullet for emergencies and epidemics. It was implied that access was being restricted only by an oversensitivity to privacy considerations which were far outweighed by the potential good of the data [22,23]. In reality, however, the applied research being done using CDR data has been incremental and long term, involving overlapping scientific, humanitarian and core business concerns. These concerns are generated by the ecosystemic nature of private-sector data repositories, where many different groups are involved in sharing data, each with their own rights, duties and claims with regard to the data and its outputs. The following section will explore the considerations at play in developing research collaborations, and in particular the roles privacy and ethics have in shaping those collaborations.

3. Choosing collaborations: the influence of the ecosystem

At the research department of the Norwegian mobile operator Telenor in late 2015, studies were being conducted with outside collaborators using CDRs to detect poverty levels in LMICs, predict users' income and gender, and to measure population in combination with satellite data showing night-time lights. These collaborations built on a knowledge base the company had been creating since 2001 with its first studies of how social networks influence behaviour. The firm had also just released a study conducted in Pakistan that showed CDRs could predict the transmission of dengue [11]. At the same time in Paris, Orange was funding follow-up work from its 'Data for Development' challenge which had generated 56 studies by data scientists from its Senegal CDR dataset. The company was also working on its next approach to sharing data on developing countries: an 'open algorithm' project (OPAL) that would act as a bridge into the data held by the company and allow researchers to ask questions of the data without the firm having to release it to them. OPAL would 'bring the code to the data, rather than the data to the code', according to an internal presentation, and would present a platform not only for data proprietors, including Orange, but also for others, such as Internet search firms and financial intermediaries to develop their own algorithms to allow others to ask questions of their data.

These projects illustrate the different ways in which mobile operators are conceptualizing the value of their datasets, and the way in which those operators are driving the development of new options for sharing it. Sharing CDR data, however, is always a risk. In their raw form, the data are highly privacy-sensitive, and sharing it means legal checks to comply with data protection laws, precautions to make sure competitors do not gain access to privileged information, and input from corporate social responsibility (CSR) specialists to make sure the exercise will not result in bad publicity. Moreover, when CDR data are shared (in de-identified form so that individual users are not trackable), the mobile operator gives consent for the data to be reused, not the firm's customers. This puts an extra onus on the firm to be cautious, because it is taking individuals' privacy into its own hands. On balance, it is safer not to share data. However, CDRs are increasingly shared for research. Why, then, does this occur?

(a) Business advantage

According to the interviewees for this project, CDRs are shared in a situation where the firm perceives an overall benefit from sharing them, in terms of both business advantage and social impact (which also, ideally, confers a business advantage over the longer term). Pål Sundsøy,[9] a Senior Data Scientist at Telenor Research, described the company's research into dengue transmission in Pakistan as being situated precisely at this intersection:

> In the long run it seems important to have good media impact. You choose projects based partly on business value if you want to get them approved ... Dengue is pretty altruistic in terms of our projects, we want to be a positive force in the markets where we operate.

[9] Pål Sundsøy. Senior Data Scientist, Telenor, 16 September 2015.

Researcher Taimur Qureshi[10] of Telenor similarly stated that he saw the process of using data as being 'a positive force in the markets where we operate'. This project clearly did achieve media impact: a week after the study's publication more than 100 positive articles had appeared online around the world. The dengue case was carefully chosen: the researchers had considered the possibility of creating epidemiological models for other diseases using the same data, but there were factors—including the potential political sensitivity of becoming involved in interventions in the lowest-income areas of Pakistan—that mitigated for dengue and against other diseases. Geoff Canright, head of the Data Analytics Group at Telenor Research, explained that the choice of dengue was influenced by the desire not to create tensions with the Pakistani government.

> A privacy breach or bad PR even based on false information is dangerous. We have to take a strict interpretation of the regulations.

Job performance criteria were similarly a motivation for researchers at Telenor to create projects at this intersection of business and altruism: employees are judged on both their internal and external value creation, and scientific publications and advisory work are named important components of good performance.

At Orange, the reasons to facilitate CDR research with their Data for Development challenge and subsequently with the OPAL project were similarly a mixture of altruism and core business. The first Data for Development challenge was at least partly driven by curiosity: a collaboration with Vincent Blondel, Professor of Applied Mathematics at the Université Catholique de Louvain, the challenge was an effort to go beyond the established model of collaborating with outside researchers to research business challenges, such as predicting churn.[11] The organizers believed that researchers would do more innovative and adventurous work if they believed they were contributing to African development, an entirely new topic for data science. In all these cases of data sharing, altruism alone is not sufficient for a business case, but a perception of social value is necessary to overcome the business risks of sharing data outside the company.

(b) Matching data with opportunity

The collaborations conducted by Orange, Telenor and the other operators spoken to for this project all had an element of serendipity. Firms did not generally formulate a policy of sharing CDRs with external researchers unless something happened to make them perceive the value of the data differently. This might be a social connection between a company executive and an external researcher, or, as occurred in the case of Telenor, a contact made at a conference that resulted in an introduction by an academic third party to interested research scientists. Telenor's data from Pakistan, where the dengue study was conducted, were unusual in that the company had captured its market share by going to more remote and rural areas where coverage was sparse, and the data, therefore, represented a fuller picture of human mobility than competitors' data might. This made the data particularly good for epidemiological purposes, and once Telenor's research department was in conversation with researchers from Harvard's School of Public Health, dengue emerged as a good choice for several reasons. One was that a recent epidemic in Pakistan had made it a serious national concern, and another was that it was a disease where information was key. There is currently no cure for dengue, and instead public health authorities need to identify where a wave of cases will occur so that they can concentrate resources on treating its symptoms and minimizing its impact. This created the possibility of a proof-of-concept study that would make it possible to address dengue more efficiently when it next emerged.

Another factor in creating opportunity was the national context, and specifically the existence—or sometimes the lack—of national supplementary data. One Telenor researcher

[10] Taimur Qureshi. Senior Data Scientist, Telenor, 16 September 2015.

[11] Interview with Prof. Vincent Blondel (29 March 2013).

observed that dengue presented a good choice for collaboration because 'public health information was ... collected by the government of Punjab which is pretty good administratively [and] was open to sharing data'. By contrast, Orange's Data for Development challenges revolved around two West African countries—Côte d'Ivoire and Senegal—where lack of resources made the collection of national statistics a challenge, so that supplementing or replacing national statistical products was part of the rationale of the challenge.

(c) Data-sharing models and underlying strategies

Beyond the social aspects of data sharing—perceptions of advantage and one-time opportunities—data sharing also depends on firms being able to create the right technical model for channelling and sharing the data. De Montjoye *et al.* [24] identify four main models currently in use: remote access, limited release, a question-and-answer model, and the use of pre-computed indicators and synthetic data. Of these four models, Orange's Data for Development challenge took the form of a limited release of data: researchers received an initial description of the data, then sent in abstracts on the basis of which they received four datasets. The datasets showed traffic between antennae, movement trajectories for a large group of phone users referenced by antenna location, movement trajectories according to country administrative regions, and communication network details for a group of users. The data were de-identified by Orange's local subsidiary, and antennae locations were blurred 'to protect Orange's commercial interests' [25]. This model for data sharing allowed researchers to map and track mobility as well as communication network activity in terms of space and time, but without knowing the names of the users in question.

Telenor's collaboration with Harvard Public Health was more conservative in that the data were not only de-identified by hashing and encrypting in terms of identifiers, but also aggregated and processed by the local subsidiary [26]. Following the remote access model, the Harvard researchers worked with Telenor's research department to identify the maximum spatial and temporal aggregation level for the data that would still make it possible to answer the research question. The Telenor researchers aggregated the data to base-station level locally at the Pakistan affiliate's offices, making it possible to move the data out of the country. At the firm's Oslo headquarters, they then aggregated the data up another level owing to the concerns about business sensitivity, so that it showed movements of mobile phones not between base stations but between Tehsils, Pakistani administrative units below the province level. The Harvard team were then given access to the aggregated matrices, rather than the dataset itself as with the Orange challenge.

These two models share one central feature: in each case, the local subsidiary is the only one with access to the raw data. This demonstrates the considerable power that subsidiaries retain in terms of endorsing data sharing, evaluating potential impacts and processing the data to avoid problems. In the case of Telenor, it was clear from interviews that the Pakistani subsidiary had had substantial input in the decision-making process based on its operational and political knowledge, and that without the explicit consent and involvement of the subsidiary data cannot be released. This extra layer of corporate and political checks and balances involved in any decision to share CDRs from outside a multinational firm's home country has not been explicitly considered by previous analyses of data-sharing processes [14,21], and is worth taking into account as these actors are often invisible in retrospect.

In the case of Orange, an unexpected problem occurred once the data had been shared, and the research outputs had been made public. The winning paper from the first Data for Development challenge [27] was written by a team based at IBM. This created a business conflict, since to follow up research with action, as Orange had originally hoped, further data and funding would have to be shared with a competing business. This problem led to an imaginative response by Orange: the OPAL project, which aims to provide an open platform that will effectively provide a firewall through which outside researchers can extract answers from the firm's CDRs without accessing the data directly [28].

4. Ethical questions

Data protection laws in the US and EU protect raw CDR data from being shared for reasons of privacy. However, given the potentially greater sensitivities of data about LMICs [14], even the most cautious data-sharing models may create only more nuanced problems of consent and potential harm.

(a) Consent

Geoffrey Canright of Telenor identifies consent as central to the process of sharing data. 'Informed consent is understanding what you are saying yes to, and knowing what data are being used and why. You can take your consent back at any time'. However, in the case of mobile operators sharing data with outside researchers, Nikolai Pfeiffer, Telenor Group Privacy Officer,[12] noted that 'there are two layers: consumer consent and affiliate consent'. The mobile phone user consents via their original contract with the operator to their de-identified data potentially being used for research purposes. Then the affiliate operator must consent to the firm itself using the data. The same process was described by executives at Orange. This raises the potential problem that informed consent on the part of customers is very general, and does not involve revisiting the contract when data are repurposed. In the case of Telenor, the company addresses this by making informed consent on the part of the country affiliate as specific as possible to a single use of data, and checking back if new uses are proposed. In this case, the weakness is that the affiliate takes the place of the individual in the consent process, but strength is that purpose limitation is observed.

(b) Identifiability

The second question that arises from these processes of data sharing is that of pinning down the concept of identifiability. What is personal information in the case of de-identified calling data? This is not legally a problem as long as data are de-identified. Canright explains that working internationally means dealing with multiple layers of regulation:

> We have to invent procedures, the regulators look at Personal Information … Our data is deidentified. If you aggregate location data over weeks, no way is that personal data—unless everyone goes to the same place at the same time as a group.

The group problem, however, is possibly the next frontier for privacy. It is possible to envisage situations where even de-identified data may be sensitive on the basis that people can be tracked as groups and networks [14,29], so that data aggregated for disease prediction can, instead, be used to track groups of political interest, such as separatists, smugglers, undocumented migrants or dissidents.

The dengue project was almost an ideal case from an ethical point of view because the results of the model could be shared with public health authorities in the form of paper maps rather than sharing the data or the model itself. Tracking many other diseases, however, requires disaggregated data. In the case of Ebola [21], it is necessary to know not only where people are travelling, but also what they are doing and what kind of contact they have with each other, something Mcdonald points out is so far beyond the law that it would require government to commandeer raw data from mobile operators.

Taimur Qureshi, senior data scientist in Telenor's research group, explained how sensitive location data can be:

> Even de-identified but nonaggregated data can be sensitive because you can identify someone because of his location. Location patterns also reflect behaviour. [The data shows] if you are up at night, if you are male or female—men have a larger radius of gyration

[12] Nikolai Pfeiffer. Telenor Group Privacy Officer, Telenor, 16 September 2015.

> away from home—your salary, because those who earn more tend to move more widely and often, and to socialise more widely.

This level of sensitivity in even de-identified data makes it difficult for a single framework to emerge for the sharing of CDRs because different issues will arise depending on location, political context and cultural differences. Qureshi's awareness of the difficulty of separating data from people illustrates why mobile operators are justifiably cautious about sharing their datasets: ultimately the potential for misuse depends on the user, not on the processor of the data, and therefore controlling use is the end-game for those interested in ethical data sharing.

5. Conclusion: can call detail record data be framed as a public good?

The new sources of digital data are powerful tools for building knowledge about populations and activities in LMICs. The claim that datasets, such as CDRs, should function as essentially public data, though, rests on the idea that they—like news coverage or the Internet—are so essential that restricting access effectively constitutes censorship, and 'not using data is the moral equivalent of burning books'.[13] Similarly the statement, 'Privacy is your right. So is access to food, water, humanitarian response'[14] draws a clear link between data analytics and the provision of essential goods. These claims draw on a utilitarian model of ethical reasoning where creating 'the right litmus test' (refer footnote 14) for regulation will result in private-sector data becoming more available. Yet despite 6 years of discussion, there has yet to be taken up the idea of data as a public good by regulators or mobile operators.

This paper has argued that to understand why data-as-a-public-good has not gained traction with its target audience, it is necessary to see from the point of view of that audience. Rather than censoring out of sheer risk-aversion, corporations are navigating the challenges of CDRs' potential for both good and bad in nuanced, reflexive ways that result in the evolution of new ethical approaches to data-sharing infrastructures and practices. They also experience data production as an ecosystem where different stakeholders have contractual rights, interests and property claims. On this basis, firms are highly incentivized to keep control of data from LMICs because they have high value as a proprietary resource that contains insights about business processes and customer preferences as well as the potential for professional learning and positive perceptions on the part of customers and governments—and this value is balanced by a high negative potential for the firm if the data are misused.

One underlying reason for not wanting CDRs not to become a public good is that this value is linked to their scarcity. The calls for data to operate as a public good effectively highlight the fact that it does not currently do so, and the debate highlighting CDRs' importance to researchers and policymakers effectively increases the status of firms as gatekeepers of high-value data. Out of the two overarching models for data sharing—one where an intermediary holds and provides access to multiple datasets, versus another where firms control the process of data sharing—the second is likely to be more attractive to data proprietors because it both creates scarcity and focuses the good publicity arising from such projects mainly on the firm.

The ecosystemic characteristics of private-sector data demonstrated in this paper make it harder for data to be shared without an idea of its eventual purpose. Firms experience data-sharing projects as single-use constructs created through negotiations with a constellation of different actors, including country affiliates, country governments and security services, the firm's customers, and different departments within the firm's headquarters, including CSR management and privacy officers. Furthermore, research collaborators' domain knowledge also defines how data must be channelled and packaged to address a particular research problem. This ecosystem

[13] See http://www.vodafone-institut.de/event/not-using-data-is-the-moral-equivalent-of-burning-books/.

[14] Robert Kirkpatrick, UN Global Pulse, interviewed 18 August 2014.

makes data use and sharing highly contingent, contextual and incremental, and highlights the functional reasons why firms are not sharing data more freely.

Orange's OPAL project can be seen as an attempt to sidestep this ecosystem problem by ensuring the dataset and its value remain within the firm's auspices: companies will use the OPAL platform to allow others to ask questions of their proprietary databases without actually releasing the data, and the algorithms can be regulated, controlled and certified by official bodies based on ethical considerations. However, as with the other models, this does not address the question of who has ultimate ethical responsibility for the use of the research outputs. The OPAL approach avoids the utilitarian cost–benefit analysis by placing the onus on public certification bodies to make sure that algorithms and their output are not being used in harmful ways. This may be the system's chief benefit in terms of data ethics: by removing layers of consent and permission it draws attention to the unpredictable outputs of sharing CDRs, and the necessity for a separate layer of project vetting by public bodies.

OPAL would effectively allow proprietors to rent out their datasets, and would shift ethical responsibility for the use of the data to a public body (at least as the system is currently conceived). One potential benefit of OPAL would be that, like the current system of limited sharing, it imposes the principle of purpose limitation on data analysts, forcing researchers to approach the data with an already-defined problem rather than mining the data for all possible correlations. This, in turn, limits the possibility that researchers with little contextual knowledge will be able to mine potentially sensitive data about LMIC populations, thus answering one central ethical problem with regard to the practice of data science for development [14].

The two quotations that began this paper are fundamentally mutually compatible. It should be possible to both claim that LMIC data should be used for the benefit of all, and for corporate data proprietors to be active in providing those data. However, the responsible data debate has not been able to overcome the obstacles to this by producing technical and ethical frameworks within which data can operate as a public good. It has, however, contributed significantly to the understandings of data ethics among a highly diverse disciplinary and professional community by demonstrating that the search for a litmus test, a single set of guidelines or a decision-making template, will ultimately fail because it asks the wrong question. By asking how to make datasets safe for use through the 'right' processes of de-identification or aggregation, one is effectively asking the notion of 'data ethics' to carry the twofold burden of conceptualizing the right to privacy and autonomy across widely varying cultural and political contexts, and of substituting for legal parameters where these are lacking.

The claim that data should be a public good is important to evaluate because it is a statement about power, both power over data and power over the outputs data can produce. The struggle to find an overarching framework for data sharing and use is also the struggle to assert a claim over data on the basis of moral necessity. Creating greater access to corporate datasets would distribute some of the power currently held by firms to intermediaries and researchers, and, by incorporating more decision makers and less purpose limitation in data's value chain, would also shift power away from the individuals who originally generated the data and who are most likely to be affected by possible misuse. Since an overarching framework, as currently conceptualized, would effectively substitute for legal or regulatory measures in LMIC contexts where those measures do not yet exist, corporate resistance to this strategy may have positive effects on the development of the new field of data ethics. This resistance highlights an institutional gap in terms of independent public bodies capable of auditing and authorizing new uses of corporate data. Without independent oversight, it is unrealistic to expect LMIC data to flow freely among guardians who propose to guard themselves.

Competing interests. The author declares that there are no competing interests .

Funding. Funding for this research was provided by an EU Marie Curie fellowship, no. PIEF-GA-2013-624583.

Acknowledgements. The research for this paper was conducted as a Marie Curie postdoctoral fellowship at the University of Amsterdam. I am grateful to Geoff Canright of Telenor and both Nicolas de Cordes and Zbigniew Smoreda of Orange for granting me access to their research groups for interviews and observation, and to Scott Hale and Taha Yasseri for advice on network visualization.

References

1. UN Global Pulse. 2014 Annual Report 2014 (cited 3 May 2016). See http://www.unglobalpulse.org/sites/default/files/Annual Report_2014_FINAL-DIGITAL-FOR EMAIL_0.pdf.
2. Olsen L, Saunders RS, McGinnis JM. 2011 Clinical data as a public good for discovery. In *Patients charting the course: citizen engagement and the learning health system*. Washington, DC: National Academies Press.
3. Ritchie F, Welpton R. 2011 Sharing risks, sharing benefits: data as a public good. In *Work session on statistical data confidentiality 2011; Eurostat, Tarragona, Spain, 26–28 October 2011*. (http://eprints.uwe.ac.uk/22460)
4. Kirkpatrick 2013. Data philanthropy: public and private sector data sharing for global resilience. UN Global Pulse blog. See http://www.unglobalpulse.org/blog/data-philanthropy-public-private-sector-data-sharing-global-resilience.
5. Bellagio Big Data Workshop Participants. 2014 Big data and positive social change in the developing world: a white paper for practitioners and researchers. See https://www.rockefellerfoundation.org/report/big-data-and-positive-social-change-in-the-developing-world/.
6. Stiglitz JE. 1999 Knowledge as a global public good. In *Global public goods: international cooperation in the 21st century* (eds I Kaul, I Grunberg, M Stern), pp. 308–325. Oxford, UK: Oxford University Press.
7. Varian HR. 1999 *Markets for information goods*, vol. 99. Institute for Monetary and Economic Studies, Bank of Japan. (http://people.ischool.berkeley.edu/~hal/Papers/japan/index.html)
8. Purtova N. 2015 The illusion of personal data as no one's property. *Law Innov. Technol.* **7**, 83–111.
9. World Economic Forum. 2014 *Rethinking personal data: trust and context in user-centred data ecosystems*. Geneva, Switzerland: World Economic Forum. (http://www.weforum.org/reports/personal-data-emergence-new-asset-class)
10. Jerven M. 2013 *Poor numbers: how we are misled by African development statistics and what to do about it*. Ithaca, NY: Cornell University Press.
11. Wesolowski A, Qureshi T, Boni MF, Sundsøy PR, Johansson MA, Rasheed SB, Engø-Monsen K, Buckee CO. 2015 Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proc. Natl Acad. Sci. USA* **112**, 11 887–11 892. (doi:10.1073/pnas.1504964112)
12. Bengtsson L, Gaudart J, Lu X, Moore S, Wetter E, Sallah K, Rebaudet S, Piarroux R. 2015 Using mobile phone data to predict the spatial spread of cholera. *Sci. Rep.* **5**, 8923. (doi:10.1038/srep08923)
13. Gutierrez T, Krings G, Blondel VD. 2013 Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. (http://arxiv.org/abs/1309.4496)
14. Taylor L. 2016 No place to hide? The ethics and analytics of tracking mobility using mobile phone data. *Environ. Plan D Soc. Sp.* **34**, 319–336. (doi:10.1177/0263775815608851)
15. Bellagio Big Data Workshop Participants. 2014 *Big data and positive social change in the developing world: a white paper for practitioners and researchers*. See https://www.rockefellerfoundation.org/report/big-data-and-positive-social-change-in-the-developing-world/.
16. Bengtsson L, Lu X, Thorson A, Garfield R, von Schreeb J. 2011 Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. *PLoS Med.* **8**, e1001083. (doi:10.1371/journal.pmed.1001083)
17. Pindolia DK, Garcia AJ, Wesolowski A, Smith DL, Buckee CO, Noor AM, Snow RW, Tatem AJ. 2012 Human movement data for malaria control and elimination strategic planning. *Malar J.* **11**, 205. (doi:10.1186/1475-2875-11-205)
18. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO. 2012 Quantifying the impact of human mobility on malaria. *Science* **338**, 267–270. (doi:10.1126/science.1223467)
19. United Nations. 2014 *A world that counts: mobilising the data revolution for sustainable development*. New York, NY: United Nations.
20. World Economic Forum. 2015 *Data-driven development pathways for progress*. Geneva, Switzerland: World Economic Forum.

21. Mcdonald SM. 2016 Ebola: a big data disaster. Privacy, property, and the law of disaster experimentation. CIS Papers 2016.01. See http://cis-india.org/papers/ebola-a-big-data-disaster.
22. Economist. 2014 Waiting on hold. *The Economist*, 25 October 2014. See http://www.economist.com/news/science-and-technology/21627557-mobile-phone-records-would-help-combat-ebola-epidemic-getting-look.
23. Talbot D. 2014 Cell-phone data might help predict ebola's spread. *MIT Technol. Rev.* See http://www.technologyreview.com/news/530296/cell-phone-data-might-help-predict-ebolas-spread.
24. De Montjoye Y-A *et al.* 2016 Privacy-conscientious use of mobile phone data. MIT Work Paper. See http://openscholar.mit.edu/sites/default/files/bigdataworkshops/files/draft_modelsformobilephonedatasharing.pdf.
25. Blondel VD, Esch M, Chan C, Clerot F, Deville P, Huens E, Morlot F, Smoreda Z, Ziemlicki C. 2012 Data for Development: the D4D challenge on mobile phone data. (http://arxiv.org/abs/1210.0137)
26. Buckee CO. 2015 Mobile phone data for public health: a data-sharing solution to protect individual privacy and national security. Telenor report 2/2016. Oslo, Norway: Telenor Group.
27. Berlingerio M, Calabrese F, Di Lorenzo G, Nair R, Pinelli F, Sbodio ML. 2013 AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data. In *Machine Learning and Knowledge Discovery in Databases: European Conf., ECML PKDD 2013* (eds H Blockeel, K Kersting, S Nijssen, F Železný), *Prague, Czech Republic, 23–27 September 2013*, pp. 663–666. Berlin, Germany: Springer.
28. Orange. 2016 OPAL 2016. See http://www.data4sdgs.org/dc-opal/.
29. Taylor L. 2016 Data subjects or data citizens? Addressing the global regulatory challenge of big data. In *Freedom and property of information: the philosophy of law meets the philosophy of technology*. London, UK: Routledge.