Big Data for Good: Unlocking Privately-Held Data to the Benefit of the Many

Alberto ALEMANNO*

When something bad happens these days – be it a terrorist attack, the break of a pandemic or a natural disaster – the relevant, yet new, question is: who has the data?

Today, having data means being able to save lives. Thanks to our enhanced societal ability to collect, process and use data, we are able to know exactly where – in the aftermath of an episode – the most vulnerable people are, where the epicentre of the disaster is or how a pandemic, such as Zika, is spreading. Most of this information can typically be inferred from where people check their phones following a disaster or which words they use while communicating about it on social media.[1]

Unfortunately, it is not public authorities who hold this real-time data, but private entities, such as mobile network operators, and – with even greater detail – tech firms such as Google, through its globally-dominant search engine, and, in particular, social media platforms, such as Facebook and Twitter.

Thanks to their many features – including geo-localisation and eye-tracking – and our acritical generosity in giving away so much personal information, these companies know more about us than our partners and closest friends.[2] The growing quantity of data harvested by companies through search engines, social networking sites, photo-sharing sites, messengers, and apps more generally are the result of the interaction of commercial interests, interface designs, algorithmic processes and users' indication of preferences, actions or attitudes.[3]

* Jean Monnet Professor of EU Law, HEC Paris. The author would like to thank the members of the high-level expert group on data for good, that has been established by the Organisation for Economic Co-operation and Development (OECD) and Mastercard, for the insightful conversations undertaken since Autumn 2017. Special thanks to Alessandro Spina, Cliff Wirajendi and anonymous reviewers for their comments on previous versions. The usual disclaimer applies.

[1] Today personal data is widely collected by a panoply of largely invisible parties and generally used without the knowledge or consent of exploited "data subjects". From the tracking cookies which track our movements on the Web to the flows of data generated by FitBits and other devices, personal data about both our on- and offline activities is harvested, bundled, and monetised daily.

[2] You might have heard the story of how a supermarket, Target, worked out that a teenage girl was pregnant before her father – and most probably she herself – did. Target's consumer tracking system identified 25 products that when purchased together indicate a woman is likely to be expecting a baby. The value of this information was that Target could send coupons to the pregnant woman at an expensive and habit-forming period of her life. That is targeted marketing. To know more see C Duhigg, "How Companies Learn Your Secrets" *New York Times*, 16 February 2012 and A Alemanno, *Lobbying for Change: Find Your Voice to Create a Better Society* (London, Iconbooks 2017) pp 57–59.

[3] A Richterich *The Big Data Agenda: Data Ethics and Critical Data Studies* (London, University of Westminster Press 2018).

European Journal of Risk Regulation, 9 (2018), pp. 183–191
doi:10.1017/err.2018.34

By supplementing scant public statistics and informing interventions, including in emergencies, big data functions as a crucial "sense-making resource in the digital era".[4] In particular, by disclosing these data and mashing them up with various data sets, tech companies may enable public authorities to improve situational awareness and response, thus prioritising interventions (eg go where people with asthma are, or identify where people from areas hard-hit by a disease are moving to), thus saving money and lives.[5] Think about the kinds of "safety-check" data from social media applications,[6] or mobility patterns stemming from ubiquitous geo-localisation services. No doubt this information can supplement scarce public statistics and inform interventions, in particular in emergencies.

But there's more.

While real-time data mostly benefit disaster-zones management – when situations change quickly and life-or-death decisions are typically taken under time-pressure – they may also help a speedy and efficient handling of non-emergency situations. In other words, while humanitarian use of private data epitomises their inherent, yet locked, potential in improving society, their positive societal impact is significantly greater. Indeed, also in less salient areas such as urban planning and health, commercial data present a strong life-improving potential. Thus, for instance, by identifying common predictors for liver failures, you can save lives. By assessing certain trigger messages to teen suicide, public authorities can better assess what preventive approach might (or might not) work. Or by monitoring people's use of public and private transport through geo-tagging, you can generate super-rich data capable of determining how a city operates, as well as where and how – for instance – traffic could be improved. These illustrations suggest that, whilst many point out the risks of data in abstract terms, it is the purpose pursued by data-driven and data-dependent technologies that should draw our attention.

Companies' data can tell us not only whether a given policy intervention works or doesn't work, but also how it could be fixed. Thus, policymakers could learn whether citizens consume less of an unhealthy product as a result of the implementation of a given policy, be it a soda tax, a health warning or a sale restriction of that product. In other words, access to social media datasets can improve not only the design of new public policies, but also their real-world impact. Therefore data is and must urgently become the hallmark of risk regulation. However, we should also be aware that the data we intend to unlock might not cover the so-called "data invisibles", ie those people who, generally due to their socio-economics, are not counted or tracked within the formal or digital economy.[7] These individuals are disproportionately migrants, women, children,

[4] M Andrejevic, "Big Data, Big Questions: e Big Data Divide" (2014) 8 International Journal of Communication 1673, at 1675.

[5] For a detailed analysis, see M Hilbert, "Big Data for Development: A Review of Promises and Challenges Big Data for Development: A Review of Promises and Challenges" (2016) 34(1) Development Policy Review 135–174.

[6] Facebook has declared that it will share aggregated data about people checking-in as safe, using location data of people who have agreed to this.

[7] R Shuman and F Mita Paramita, "Why your view of the world is riddled with holes" *World Economic Forum*, 2016 available at < www.weforum.org/agenda/2016/01/data-invisibles-ignore-at-our-peril/ > .

https://doi.org/10.1017/err.2018.34 Published online by Cambridge University Press

rural and slum dwellers, frequently marginalised within their own societies, and as such they are not a data point.[8]

The realities of private data-sharing

However, besides a few isolated and self-proclaimed "data philanthropy" initiatives and other corporate data-sharing collaborations,[9] data-rich companies have historically shown resistance to sharing this data for the public good. Therefore, despite its undeniable life-changing potential, private data remains the prerogative of a few big corporations who jealously guard it.

While open data is becoming more common in government, academic and institutional datasets,[10] this kind of data availability has not yet been taken up by corporations, who struggle to embed these values in their business operations. Thus, for instance, accessibility to private data is not even granted to external actors in, for instance, mandatory data audits. More generally, while companies understand the commercial potential of their data, they lack a comparable awareness about the public utility of those data. Social purpose is not how data is wired into their business models and corporate culture.

The ensuing phenomenon is giving rise to an alarming data asymmetry in society, which the emerging initiatives of voluntary corporate data sharing do not seem adequate to address and overcome. Such a data asymmetry feeds into consumers' lack of agency, and further strengthens the market dominance of a very few tech corporations, such as Google and its parent company Alphabet Inc, Facebook and subsidiary platforms such as Instagram, Whatsapp, and Oculus VR, as well as also increasingly popular apps such as Snapchat.

To justify their restrictive stance, companies typically invoke the need to preserve their competitiveness in the market and to fully protect the privacy of personal information. Moreover, further regulatory, reputational, fiduciary, allocation of risk and other obstacles are typically invoked as preventing companies from sharing their data. As a result, as sharply summed up in the literature:

> "Despite the growing acknowledgement of the benefits, we are far from having a viable and sustainable model for private sector data sharing. This is due to a number of challenges – most of which revolve around personal data privacy,[11] and corporate market competitiveness".[12]

[8] India, for instance, has tens of millions of undocumented immigrants, with 10 million from Bangladesh alone. See supra, note 7.

[9] For an initial collection and taxonomy of corporate data sharing, see GovLab, "Data Collaboratives: Creating public value by exchanging data" at < datacollaboratives org/ > .

[10] As documented by the work of the Open Government Partnership, resistance is also notable in these domains.

[11] While companies highlight the many privacy and security risks stemming from disclosing personally or demographically identifiable information for the social good, they have shown less concern and care when sharing the very same information for commercial purposes. See, eg, the investigative work conducted by C Cadwalladr, "The great British Brexit robbery: how our democracy was hijacked" *The Guardian*, 7 May 2017.

[12] A Pawelke and A Rima Tatevossian, "Data Philanthropy. Where are we Now?" *United Nations Global Pulse Blog* (2013), available at < www.unglobalpulse.org/data-philanthropy-where-are-we-now > .

https://doi.org/10.1017/err.2018.34 Published online by Cambridge University Press

While there do exist several legal and other legitimate obstacles to the release of these data, even in emergency situations, those are far from being insurmountable.[13] There are indeed methods of balancing both privacy and competitive risks with data-sharing for public good. These include aggregating data or sharing insights from datasets rather than the raw data. Yet this – as with any data-sharing exercise – entails significant transaction costs (eg preparing the data, de-risking them, etc) on the supply side.

The responsible data movement

It is against this backdrop that a growing number of organisations – be they international development, humanitarian as well as other civil society organisations and public authorities – have called (and continue to call) for private digital data to be shared for the public good. They want them to be treated as "public goods" because of their inherent value in informing interventions, and not only in emergencies.

The term 'data commons' was popularised by the *United Nations Global Pulse* initiative – a flagship project promoting the use of big data for sustainable development and humanitarian action.[14] In Davos, during the 2011 World Economic Forum, Kirkpatrick – speaking on behalf of *UN Global Pulse* – complained that "[...] while there is more and more of this [big] data produced every day, it is only available to the private sector, and it is only being used to boost revenues".

In response to this claim, a "responsible data" movement has emerged to lay down guidelines and frameworks that will establish a set of ethical and legal principles for data sharing.[15] Yet, in the absence of a common multi-stakeholder platform for data governance, this movement lacks institutionalisation, and as a result appears largely fragmented. Several initiatives, like the Signal Code[16] and the International Red Cross Handbook on Data Protection in Humanitarian Action,[17] are developing. Often based on these frameworks, actors from various sectors – such as social media companies and civil society actors – exchange information to create new public value through so-called "data collaboratives",[18] typically in the form of public-private partnership. While their experimental nature must be praised, these initiatives are – at best – limited to specific sectors, such as health, disaster response, education, poverty alleviation, and – at

[13] More challenging instead is the issue of data bias, as epitomised by the so-called of "data invisibles", ie individuals, generally from vulnerable communities, who are unrepresented in private or public datasets. In other words, the generalisability stemming from extrapolating general observations from such data should be questioned.

[14] R Kirkpatrick, "Data Philanthropy. Public & Private Sector Data Sharing for Global Resilience" *United Nations Global Pulse Blog*, available at <www.unglobalpulse.org/blog/data-philanthropy-public-private- sector-data-sharing-global-resilience>.

[15] Initially a "Data for Good" movement has been encouraging using data in meaningful ways to solve humanitarian issues around poverty, health, human rights, education and the environment. It is now is in the process of being mainstreamed by the OECD.

[16] This seeks to apply human rights principles to data during times of emergencies. See <signalcode org/>.

[17] A Handbook was published as part of the Brussels Privacy Hub and ICRC's Data Protection in Humanitarian Action project. It is aimed at the staff of humanitarian organisations involved in processing personal data as part of humanitarian operations, particularly those in charge of advising on and applying data protection standards.

[18] UN Global Pulse, GSMA, "State of Mobile Data for Social Good," Report Preview, February 2017 and the work led by NYU GovLab under this label that has collected more than 100 examples of public value extraction from privately-held data collected. To know more, see <datacollaboratives org/>.

worst – may hide a commercially-driven attempt to gain a research insight, access to expertise, entering a new market or merely gaining positive visibility, often vis-à-vis the companies' employees.[19] As such they remain generally under-used as solutions for large, complex public problems. As a result, no generalisable approach to data sharing has yet emerged.[20]

How to then move away from one-off, sectoral projects, toward a scalable, broader, genuine approach to private data-sharing? In other words, how to best "institutionalise" data-sharing for public good within the private sector and in collaboration with public sector and philanthropic actors?

This is the mission pursued by the OECD, which has teamed up with the MasterCard Center for Inclusive Growth to identify and formalise data-sharing methods.[21] After several social media corporations and organisations – such as Facebook, the International Committee of the Red Cross, the United Nations Global Pulse and NYU GovLab – explored various frameworks to private-sector collaborations, the OECD and MasterCard have been analysing dozens of data-sharing initiatives within corporations so as to bring them to the next level. In particular, the OECD has been working on the codification of a set of principles and assessment frameworks aimed at the development of a methodology for voluntary, private-sector data sharing. While this effort must be praised insofar as it goes beyond the sector-by-sector approach, it falls short of developing an accountable governance process capable of applying these principles. Data use and sharing are set to remain highly contingent, contextual and incremental and – more critically – to be entirely subject to data controllers themselves.

The cost of not sharing private data

Moreover, the OECD approach focuses on the sharing of insights from corporate data, rather than the transfer of raw data to third-party researchers or organizations. While sharing of insights has emerged as a commonly-used marketing model, one may wonder under what circumstances companies will actually agree to do so, given the high costs involved. More critically, by making data sharing conditional upon a utilitarian cost-benefit analysis ("societal benefits should clearly outweigh other risks"),[22] it is unlikely that the proposed OECD approach alone will instil an institutional culture of sharing in the business community.

Shouldn't we ask more often "what's the risk of not sharing?"

[19] The literature focusing on data-sharing initiatives focusing on Call Details Records (CDR) demonstrates that data sharing occurs where the firm perceives an overall benefit from sharing them, in terms of both business advantage and social impact (which may turn into a business advantage in the longer term). See, eg L Taylor, "The Ethics of Big Data as a public good: which public? Whose good?" (2016) Philosophical Transactions of The Royal Society 374.

[20] Noteworthy, however, is the European Commission's Guidance on sharing private sector data in the European data economy (SWD2018 125 final), which aims to provide a toolbox for companies on the legal, business and technical aspects of data sharing, in particular with respect to machine-generated data, notably for public interest purposes.

[21] The OECD and Mastercard convened a high-level expert group in data for good in Autumn 2017, of which I am a member.

[22] Draft methodology for a Set of Principles/Methodologies for Private Sector Data, OECD, 2018 (on file with the author).

https://doi.org/10.1017/err.2018.34 Published online by Cambridge University Press

Yet, after more than a decade of the rhetoric of ‘data as public good’ floating in public and corporate discourse, the idea of systematically assessing the balance between risks and rewards of sharing that data remain the exception – not the norm – in corporate operations. Only few companies – and not necessarily the large ones – have data management and data governance structures capable of identifying and fostering greater utility for their data or meta-data.

As a result, despite their potentially life-saving nature, these collaborations are entirely left to the goodwill of the private actors involved. In other words, the guardians guard the guardians.

How to unlock private data for good?

The urgent question today, therefore, is how to move from this emerging, sector-to-sector, voluntary approach to a more universal, sustainable and accountable data sharing model (or models). To do so, one has to examine whether, and how, the argument for “data as a public good” fits with the corporate reality of big data.

It has been suggested that to unlock the potential of commercial data for the public good, one has to move away from the concept of data as something to be owned or controlled. Conferring ownership – or, as we do in Europe, giving control to third parties over personal information[23] – seems indeed inherently to inhibit societal beneficial use. It is the language of ownership that makes companies blind to the many opportunities where the data they collect and hold create public value.

Yet, even if we were to posit that the data belong to the users and not the company controlling their data, how do we deal with requests for access to the compilation or aggregation of the data of many thousands or millions of users by public authorities claiming they need these data in order to save lives (or attain other legitimate public policy goals)?

A more sensible, emerging approach is that of stewardship, which is intended to convey a fiduciary level of responsibility toward the data. Under such an approach, companies would suddenly realise how their data can benefit them beyond their businesses’ bottom lines, thus embracing a culture of data sharing that is currently missing. Yet even the stewardship approach does not seem to come to terms to the question of why private sector-data should suddenly be released, regardless of an immediate return.

What then can be done to gain access to and use data collected by third parties? Property law is just one of the legal regimes that control rights and responsibilities in relation to personal data. Many others exist, such as tort law, contracts, as well as regulatory law, including competition law rules.[24] This suggests that the companies’ stubborn refuse to share personal data may lead some of these legal regimes to be triggered soon. Thus, for instance, other companies – be they new entrants on the market,

[23] Under EU law, this third-party control of data is limited to the rights of the “data subject”, which are related to the fundamental right to data protection/privacy.

[24] This is insofar as big data can be seen as a source of market power and, consequently, as a possible field of abuse for undertakings in dominant positions. To know more, see A Thierer, “The Perils of Classifying Social Media Platforms as Public Utilities” (2012)–(2013) 21 CommLaw Conspectus 249; “Essential Data” (2014–2015) 124 Yale LJ 867.

social enterprises or even public authorities – could try to invoke the so-called essential facility doctrine.[25] Despite being defined rather narrowly by the European Court of Justice, some authors are beginning to argue that big data could be regarded – on a case-by-case analysis – as an essential facility, at least in some specific sectors, and as such could be the object of mandatory disclosure.[26] The scalability of this model of ad hoc data-sharing would, however, be inherently limited. A more promising avenue to unlock the power of private data seems today to be offered by the emergence of a new regulatory regime, notably the EU General Data Protection Regulation (GDPR) and by the supervisory authorities which will oversee its implementation.[27]

While its governing principle is that of purpose limitation, which confines the use of data by the data controller to one specific purpose, this Regulation offers some flexibility in given circumstances. It expressly allows a derogation to the principle in case of "further processing for archiving purposes *in the public interest*, scientific or historical research purposes or statistical purposes".[28] If interpreted favourably, and applied proportionally, this provision may be used so as to strike a balance between the interests of data controllers – who could overcome the personal data privacy concerns – and that of the data subjects, whose data could be used in the public interest, to the benefit of the many.

Although the debate over data as public good has not been able to bring about a generalisable model of data sharing, the new European data protection regulation may offer a promising entry point capable of breaking one of the major sources of corporate resistance against data-sharing: personal data privacy and its inherent risk allocation. By making data use accountable, GDPR enables companies (ie data controllers) to go beyond the principle of purpose limitation. As such, it may grant a supervised yet general power to use personal data beyond the original purpose for which they are collected, when it is for the public good. This could potentially be a game-changer in data-sharing practices for the public good. Yet this is not to suggest that fixing the supply side alone by data holders will magically turn data-sharing into a reality.

The demand-side from institutions – be they public authorities, civil society organisations, statistical agencies – also requires some deep structural and methodological work to take full advantage of released data. One indeed cannot assume that recipients are ready to use those data. Today, any data-sharing activity requires significant and time-consuming effort and investment of resources for both the data holders on the supply side, and the institutions that represent the demand.

There is, however, a need for both sides of the equation to understand one another, align expectations, and for the respective benefits to be understood and publicly disclosed. Indeed, both societal and business benefits must be declared and assessed. Unfortunately, for the time being, most of the efforts focus on getting the supply-side

[25] Initially developed by the US Supreme Court (*United States v Terminal Railroad Ass'n of Saint Louis*, 224 US 383 (1912)), this doctrine was applied by the European Commission for the first time in 1993 in a case related to harbour infrastructures, before being extended to matters related to non-material facilities.

[26] *Contra*, eg G Colangelo and M Maggiolino, "Big data as misleading facilities" (2017) 13(2–3) European Competition Journal 249–281.

[27] General Data Protection Regulation (GDPR) (EU) 2016/679.

[28] Art 5.1(b) GDPR.

https://doi.org/10.1017/err.2018.34 Published online by Cambridge University Press

ready for data-sharing – by developing data-share models, de-risking the data, preparing the skill-set for the personnel[29] – but omit the need to sensitise and prepare the demand side.

Today, establishing and sustaining these new collaborative and accountable approaches requires significant and time-consuming efforts and investment of resources for both the data holders on the supply side, and the institutions that represent the demand. Yet as more and more data-rich companies are set to realise the utility of their data and understand how the demand side may be using them, this might help to create an enabling environment for data-sharing. And it remains doubtful whether mandating release, on the one hand, or pricing data-sharing, on the other, would per se be capable of attaining such a goal. There is a clear case for continued experimentation.

Conclusions

It is almost a truism to argue that data holds a great promise of transformative resources for social good, by helping to address a complex range of societal issues. Yet as data are inaccessible, especially when it comes to those produced on commercial platforms, it is high time to unlock them for the social good.

The public debate has thus far opposed those who perceive data as commodities and those who believe they are the object of fundamental rights. As demonstrated by this article, the emergent discussion about the use of private data for the public good provides a welcome opportunity to enrich – by making it more complex – such a debate. The time has indeed come to embark on a less polarised conversation when it comes to the governance of data in our societies. To do so, it is essential to decouple the legal and ethical aspects of data-sharing and identify new approaches towards it. While the welfare-enhancing properties of data sharing make such a practice a moral necessity, we need technical and legal frameworks capable of translating such a growing moral imperative into workable and legally-sound solutions.

At a time in which conversations about tech companies tend to be negative – as signified by the abuses unveiled by scandals of global-scale tax avoidance or data-harvesting such as *Cambridge Analytica* which were initiated by some “data philanthropy” – the enduring refusal of the data-rich private sector, notably social media companies, to release part of their data under certain circumstances may trigger a further backlash against them.

Indeed, should the potential of real-time private data to save lives become public knowledge, the reputation of social media and other data-rich companies could be further tarnished.

How to explain to citizens across the world that their own data – which has been aggressively harvested over time – can’t be used? Not even in emergency situations?

Responding to these questions entails a fascinating research journey for anyone interested in how the promises of big data could deliver for society as a whole. In the

[29] One of the most promising, highly-specific initiatives is the one recently undertaken by GovLab with the Data Stewardship Portal, available at <thegovlab.org/tag/data-stewards/>.

absence of a plausible solution, the number of societal problems that won't be solved unless firms like Facebook, Google and Apple start coughing up more data-based evidence will increase exponentially, as will societal rejection of their underlying business models.

While embedding data-sharing values in the corporate reality is an opportunity for the few tomorrow, it is already a life-or-death matter for the many, today.

https://doi.org/10.1017/err.2018.34 Published online by Cambridge University Press