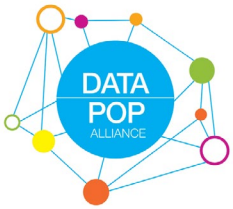
Vodafone Institute for Society and Communications

In partnership with:



HARVARD HUMANITARIAN INITIATIVE

mit media lab

HOW TO USE BIG DATA?

LEADING EXPERTS' ROADMAP TO DATA-DRIVEN INNOVATION PROJECTS

Key results from the Digitising Europe Initiative

Emmanuel Letouzé, Director and Co-Founder, Data-Pop Alliance
David Sangokoya, Research Manager, Data-Pop Alliance
In Cooperation with the Vodafone Institute for Society and Communications

Foreword by Alex ‘Sandy’ Pentland

Towards a data-enabled social contract

This paper is the result of a 2-year collaboration between the Vodafone Institute for Society and Communications, self-described as Vodafone‘s “European thinktank”, and Data-Pop Alliance, a partnership between the MIT Media Lab, the Harvard Humanitarian Initiative, the Overseas Development Institute and the Flowminder Foundation I co-founded. In many ways, this collaboration and its outcome show the way of the future, towards a data-enabled social contract.

We created Data-Pop Alliance with the aim of “promoting a people-centred Big Data revolution” in 2014, at a time we felt discussions and initiatives in this still nascent field were often too polarized, and too theoretical. As a scientist, I of course have nothing against theory. “Social physics” aims to provide elements of a computational theory of human behaviour. What I think we need to collectively work towards, are better theories about how human systems work to help them achieve more ambitious societal goals. And as has been the case since the invention of science, this requires observing, hypothesizing, conceptualizing, measuring, testing, inferring, suggesting, discussing, deciding (on what should and could be changed), and so forth. This is how human progress has always happened.

„We will move towards social systems where human interactions are less confrontational; decisions more rational ...“

The big difference between today and then is indeed the level and speed at which ‘we’ can now measure, monitor, and even manipulate human behaviour—as always, for both better and worse. The history of technological innovations and revolutions is one where the new tools are, typically, first in the hands of the powerful, who, for the most part, seek to use them to further entrench their power. It takes time and effort for the benefits of the new technology to spread. And so, it is good and necessary that critical voices be heard, red flags waved, opposing perspectives confronted, competing interests unpacked. Our collaboration with the Vodafone Institute, and the 4 public events that structured it, reflected this conviction. This is democracy as government through discussion. This is development as co-design. This is how some form of consensus emerges and societies can move forward. This is hard because by and large current incentive structures and systems do not reward cooperation very well.

But in the future, this should and I suspect will change. We will move towards social systems where human interactions are less confrontational; decisions more rational; prejudice less inescapable; and our economic and political models more sustainable—because the effects of the various alternatives will appear more clearly to more people. The key to this new social contract will be data—not just the bits, but the systems and architectures enabled by data—controlled by their primary producers and users. This is of course a toll order, a bold vision—some will say a fantasy; others a fallacy. And yet I think thriving for a world where data is at the heart of new social contracts based on greater transparency, trust, and accountability is both a desirable and accessible objective. The Open Algorithms project (described in Figure 1) is certainly a step towards this vision. And I view this paper and collaboration with the Vodafone Institute, Vodafone’s “European Thinktank” as another small but important step.

Acknowledgements

Data-Pop Alliance and the Vodafone Institute for Society and Communications are especially grateful to the following institutions and individuals for their participation in and support to the four public events that have constituted the backbone and raw material of this paper [2].

In Berlin, November 12, 2015: Andrew Keen, entrepreneur and author, and Alex 'Sandy' Pentland, Professor, MIT, and Academic Director, Data-Pop Alliance.

In Brussels, January 25, 2016: Kenneth Cukier, Data Editor, The Economist and co-author of "Big Data: A Revolution That Will Transform How We Live, Work, and Think", Dr Linnet Taylor, Marie Curie Research Fellow, "Data for Development: the Implications of New Types of Digital Data for International Development", University of Amsterdam; Senior Research Affiliate, Data-Pop Alliance, Kevin O'Connell, Member of the Cabinet of EU Commissioner for Justice Věra Jourová, Joe McNamee, Executive Director, European Digital Rights (EDRi), Nico van Eijk, Professor of Media and Telecommunications Law, University of Amsterdam, Nicklas Lundblad, European Public Policy Director, Google, Rob Shuter, CEO, Vodafone Netherlands, Yves-Alexandre de Montjoye, Post-Doctoral Fellow, MIT and Harvard University; Senior Research Affiliate, Data-Pop Alliance, Matthew Kirk, Chairman of the Advisory Board, Vodafone Institute, and Frances Robinson, freelance journalist and moderator

In Madrid, June 2, 2016: José Luis Zimmerman, Director de la Asociación Española Economía Digital, Professor Esteban Moro, Associate Professor, Universidad Carlos III de Madrid and Member of the Joint Institute UC3M-Santander on Big Data, Marco Bressan, Chief Data Scientist, BBVA, Jose Luis Melero, Client Service Director, TNS, Daniel Noguera, CEO, RED.ES, Pedro Peña, Director of Legal and Regulatory, Vodafone Spain, Pilar Torres, Director of Operations and Marketing, Microsoft, David Alandete, Deputy Director, El Pais, Francisco Roman, President, Vodafone Spain, as well as our partners for the event Vodafone Spain and El Pais.

In Dublin, September 1, 2016: Heather Savory, United Kingdom Deputy National Statistician, Ronald Jansen, Deputy Director, United Nations Statistical Division, Amparo Ballivian, Economist, The World Bank and Mikko Niva, Group Privacy Officer, Vodafone.

[2] Institutional affiliations and titles are those at the time the events were held.

Content

The Digitising Europe Initiative

Through a series of regional stakeholder dialogues and debates in Berlin, Brussels, Madrid and Dublin, the Vodafone Institute for Society and Communications and Data-Pop Alliance initiated the digitising Europe initiative as an opportunity to identify ongoing tensions between privacy protection and big data-driven innovation, and provide insights for European companies and public sector institutions on initiating big data-driven projects and promoting responsible use and governance of big data for public good.

Taking a multi-stakeholder approach, the initiative convened the stakeholders of the European data revolution—academic experts, private companies, government, civil society and the public from across Europe—to discuss its social, economic and ethical dimensions and identify both practical solutions and policy recommendations for shaping a people-centred, growth-driven data revolution for European companies and citizens.

Industry, government and institutional representatives have included: BBVA, Betterplace Lab, DG Connect, El País, European Digital Rights, EuroStat, European Commission, Facebook, Google, Harvard University, Massachusetts Institute for Technology, Stiftung Neue Verantwortung, University of Amsterdam, University de la Madrid, Vodafone Group and the World Bank.

Each of the country roundtables highlighted a specific dimension of big data-driven innovation:

Digitising Europe Roundtables

- **Berlin** – the potential of big data-driven innovation: The discussions in Berlin largely centred on issues of anonymization and data protection; objectivity and validity of big data's claims; and the diversity of expertise needed to explore existing potential solutions and develop new tools and safeguards towards optimal data use.

- **Brussels** – the practice of big data-driven innovation: With experts across government, academia and the private sector, this roundtable specifically highlighted the new role of the private sector in the responsible collection, governance and use of big data for societal purposes. Panellists discussed the need for all stakeholders to "do no harm" by co-creating responsible governance structures and test cases on data use; understanding current notions of privacy; and assessing both individual and group privacy implications of big data-driven innovation.

- **Madrid** – the promotion of big data-driven innovation: Overall, these discussions focused largely on issues of transparency, accountability, consent and literacy. Panellists emphasised the need to redefine partnerships between the public and private sector around personal data; the need to set up clear and transparent internal governance rules, standards and procedures for companies (who gets access to which data and under which conditions); create shared language around big data use and communities with infomediaries; and invest in data literacy efforts and developing language for the public through media.

- **Dublin** – the positioning of big data-driven innovation: The last roundtable in Dublin focused on the potential, practice and promotion of Big Data for Official Statistics and the evolution of use of non-traditional data sources by national statistical offices. Serving as a side event to the International Conference on Big Data for Official Statistics, the Dialogue on Access to Big Data and related Privacy issues was organized by the UN Global Working Group on Big Data for Official Statistics, Data-Pop Alliance and Vodafone Institute. Panellists and experts discussed the positioning of these new kinds of innovations among government and policymakers and a roadmap towards their incorporation with traditional systems.

These efforts have culminated in this position paper that reflects the major takeaways from these events and provides insights for both the European private and public sectors in unleashing big-data driven innovation and promoting responsible data governance for public good.

Executive Summary

Big data has increasingly been viewed as a critical lever of innovation for society to understand evolving societal behaviour and improve policy and practice at large. Much of the societal usefulness of big data comes from discoveries in secondary or alternative uses of passively collected data from companies, such as the use of call detail records (CDR) or location data for humanitarian response and disease tracking. Through new data-driven partnerships and activities, companies have shared big data through various modalities with the public sector towards social benefit.

Privacy has often been seen as a challenge or tension in sharing big data. Although anonymization and 'privacy by design' have been described as potential solutions, academic research has revealed that, with a few points of identifiable information, assumingly anonymised data could be "de-anonymised," calling into question common measures and solutions taken by companies and governments.[3]

Europe faces a multitude of socio-political and environmental crises, and the use of big data in the data revolution provides opportunities towards greater efficiency and better decisions with scarce resources.

Ultimately, how can the stakeholders of the European 'big data revolution' — governments, companies, civil society and the public — both derive social value through big data-driven innovation, while preserving privacy and emphasising transparency and accountability?

This position paper highlights overall takeaways and recommendations in the areas of privacy protection, responsible data governance, transparency and accountability for unleashing big data-driven innovation:

1. Putting 'privacy by design' into action: privacy-preserving technical procedures and standards for data sharing and use

Privacy must be baked into the technology and infrastructure of data protection efforts, not as an afterthought or simply a set of guiding principles. This paper suggests four categories of privacy engineering solutions and design strategies for big data use:

- Distributed Data Repositories: Model in which different kinds of data are stored in separate repositories and tools are deployed from an external or remote human query that can send queries to the correct data repository.
- Move the Algorithm to the Data (Distributed, Privacy-Maximizing Algorithms): Model in which each data repository enables remote queries to send their query statements or algorithms to the repository.
- Data Always in Encrypted State (at Rest and in Computation): Model in which data is always in an encrypted state and new cryptographic algorithms and approaches allow operations to be carried out without need to decrypt first
- Encode Data Usage Agreements in Legal Trust Networks: Trust network model for large scale data sharing in an ecosystem, combining computer network tracking user permissions for each piece of data within a specific legal framework.

2. Focusing on responsibility in data use: establishing internal responsible data governance standards

Companies and other big data stakeholders—including privacy experts, policymakers, data scientists, data users, private sector and ethicists—need to co-create infrastructure and spaces for discussion, safe experimentation and transparent findings. Best practices for responsible data use include: adaptation of ethical modalities such as internal review boards; internal ethical review processes and committees involving external actors ("Bring the ethicists to the table"); creation of advisory boards and ethical review processes involving external actors; and sending a data scientist-in-residence to work within another company instead of sending data.

3. Keeping transparency, trust and user control at the centre: engaging all data stakeholders

New data public private partnerships must make transparent for all data stakeholders where and how the data is shared and what kinds of public problems will be solved through these data innovations in order to build public trust. Best practices for transparency and accountability measures include inclusive and participatory governance's mechanisms (including participatory frameworks, opt-in and opt-out consent options), multi-stakeholder public dialogues and data-driven user engagement campaigns, and broader data literacy efforts.

Taking into account these key areas, the roadmap (see p. 21) highlights three main areas of investment for companies aiming to initiate big data-driven innovation projects:

[3] de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. „Unique In The Crowd: The Privacy Bounds Of Human Mobility".Sci. Rep. 3. Nature Publishing Group. doi:10.1038/srep01376.

Background: The big data revolution and social impact in Europe

The data revolution: the emergence of new data from today's technology-infused, sensor-laden society. The data revolution describes the flood of opportunity for companies to learn, disrupt and create value from the digital interactions and traces from consumers. In the last decade, society has witnessed a digital revolution that has transformed the way in which we interact, work, learn, and do business. As consumers participate within a technology-infused, sensor-laden society, their personal data follows along as well and is often used and collected by businesses, public agencies and other entities across inter-connected platforms and devices.

Big data can help provide insights and evidence for answering key questions towards economic and societally valuable insights.

In addition to open government data and national statistics, big data has increasingly been viewed as a critical lever of innovation for society. Big data includes data from credit cards transactions, transportation (transit card data and vehicle GPS), online searches, physical sensors (electrical meters, weigh sensors on a truck) or remote sensors (satellites, cameras). Compared to traditional instruments such as surveys and censuses, these data "crumbs" can provide information with more frequency, often with more geographical precision, and at a lower cost because the data is passively collected. In the use and analysis of big data on how people actually behave, big data helps provide resources towards a scientific and systematic exploration of the human societal experience towards societally valuable insights, new forms of economic and social value, better resource allocation, and ultimately greater public good.

In discussing big data, it is essential to understand big data as more than just data sources, but as an ecosystem of new data sources, capacities and stakeholders.

Originally framed as the "3 V's"—volume, velocity and variety—in the early 2000s, big data has emerged as an ecosystem of "3 C's": digital "crumbs" (digital translations of human actions and interactions captured by digital devices); powerful capacities to collect, aggregate and analyse data; and communities involved in generating, governing and using data, including data generators, end users, policy-makers, experts, privacy advocates and civic hacker communities.

Several case studies have emerged describing the potential of big data shared by the private sector for social good. **Much of the societal usefulness of big data comes from discoveries in secondary or alternative uses of passively collected data from companies, such as the use of call detail records (CDR) or location data for humanitarian response and disease tracking.** In the last seven years, several case studies, BBVA's Big Data Innovation Challenge and Vodafone's work with local government in the UK have emerged involving companies sharing big data through various modalities with the public sector towards social benefit; these experiments have been presented as exciting new opportunities to understand evolving societal behaviour and improve policy and practice at large.

Privacy has often been seen as a challenge or tension in sharing big data.

The rules governing the Big Data ecosystem have been a source of constant debate in light of widespread corporate and government use of data that can potentially counter an individual's right to privacy. In today's increasingly connected, big data world, the emergence of big data problematizes several established governing principles around data collection, sharing and consent. Individual users are largely unaware of how and through which channels their data is used and processed, and the mechanisms used by companies to provide notice and consent (e.g. terms and condition agreements, privacy policies, etc.) have often failed to provide meaningful choice.

Privacy protection continues to remain at the centre of data policy conversations in Europe.

Brussels has largely been at the centre of EU data protection conversations—specifically around the adopted General Data Protection Regulation (GDPR)—in which EU policy-makers and officials have debated for years how to legally find the balance between the opportunities of big data use and the risks to individual privacy protection. Although anonymization and 'privacy by design' techniques have been described as potential privacy-preserving solutions, academic research has revealed that, with a few points of identifiable information, assumingly anonymised data could be "de-anonymised," calling into question common measures and solutions taken by companies and governments. [4]

Europe faces a multitude of socio-political and environmental crises, and the use of big data in the data revolution provides opportunities towards greater efficiency and better decisions with scarce resources. Ultimately, how can the stakeholders of the European 'big data revolution'— governments, companies, civil society and the public—both derive social value through big data-driven innovation, while preserving privacy and emphasising transparency and accountability?

[4] de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. „Unique In The Crowd: The Privacy Bounds Of Human Mobility".Sci. Rep. 3. Nature Publishing Group. doi:10.1038/srep01376.

SECTION 1:

Putting ‘privacy by design’ into action: Privacy-preserving technical procedures and standards for data sharing and use

Risks and Challenges

Risks and challenges in privacy protection. Privacy protection has often been illustrated as the opponent of data-driven innovation within current policy conversations. The risks to privacy described in the GDPR have mainly focused on the collection of personal data and the protection of individual privacy. Protection of individual privacy mainly involves the prevention of unintended release of personally identifiable information (PII), lack of meaningful notice and consent mechanisms by companies for consumers and lack of user transparency in how personal data is being collected, shared and used.

Big data elicits risks to groups as well as individuals.

Individuals’ belonging to specific social groups — in terms of their gender, ethnicity, sexual orientation, etc. — tend to show in big data including CDRs, and may be used for targeting purposes whether or not the individual’s identities are known, raising concerns over ‘group privacy’. Described in Nathaniel Raymond’s work, demographically identifiable information (DII) describes information that is “either individual and/or aggregated data points that allow inferences to be drawn, enabling the classification, identification, and/or tracking of both named and/or unnamed individuals, groups of individuals, and/or multiple groups of individuals according to ethnicity, economic class, religion, gender, age, health condition, location, occupation, and/or other demographically defining factors.” [5] The release of both PII and DII can lead to a kind of group association that, within a spectrum of harms to consumers, may result in effects from unwanted targeted ads to demographic based discrimination against actual and perceived group members.

Pseudo-anonymization techniques are uncertain at best in preventing re-identification in the age of big data.

A method of data protection traditionally used during the past decade has been pseudo-anonymization: pseudo-anonymising a dataset by removing personally identifiable information (PII) during data processing; however, in the age of big and linked data, this technique can render re-identification more difficult, but not impossible. In fact, the high degree of predictability and unicity of human behaviour is what makes CDRs both valuable and also at risk of ‘re-identification’ of an individual in the dataset. For instance, in his work on the intrinsic re-identification risk of a data set, Yves-Alexandre de Montjoye, now Assistant Professor at Imperial College, showed that 4 spatio-temporal points are enough to re-identify 95% of people in a mobile phone database of 1.5M people and 90% of people in a credit card database of 1M people. This work also highlighted that coarsening the data–by lowering the resolution of the dataset through spatial and temporal aggregation- only required additional data points to uniquely identify people, meaning that “even coarse datasets provide little anonymity.” Though some researchers have developed a methodology that injects ‘noise’ in CDRs to make re-identification more difficult, it still only requires a few more data points to single out individuals.

[5] Raymond, Nathaniel. “Beyond ‘Do No Harm’ and Individual Consent: Reckoning with the Emerging Ethical Challenges of Civil Society’s Use of Data” in L. Taylor, L. Floridi, & B. van der sloot Eds., Safety in numbers? Group privacy and big data analytics in the developing world. Springer. (forthcoming)

Data protection reform: the General Data Protection Regulation (GDPR).

Following years of discussion and debate, the adopted GDPR aims to harmonise data protection in the member states and ease business activities for foreign companies. The General Data Protection Regulation (GDPR), for which a political agreement was reached in December 2015, will apply to most or all digital interactions involving humans as a result of the broad definition of personal data, which also includes pseudonymous processing. As the common standard for the EU, the GDPR aims to harmonise data protection in the member states and ease business activities for foreign companies. However, the **GDPR remains vague across several elements,** leaving room for interpretation and not always providing sufficient guidance for businesses and supervisory authorities. For instance, the rules applying to profiling based on big data are only vaguely articulated. The underlying principles of the regulation involve informing the data subject and, at times, getting their consent, purpose limitation and data minimisation. Critics note the limits of these policy tools in a big data world where the use cases often evolve after the actual collection of the data. Also, in view of the amount of data sources and networked nature of data processing, it seems inevitable that, while recognising the need for transparency and reasonable choice, the emphasis will have to move from the actual collection of data into ensuring responsible and accountable use of data.

Solutions and Strategies

Operationalising “privacy by design” requires a set of strategies addressing privacy across the project’s architectural, governance and operational dimensions.
While the GDPR’s emphasis on privacy by design and default underlines the value of privacy, privacy must be baked into the technology and infrastructure of data protection efforts, not as an afterthought or simply a set of guiding principles. Drawing from decades of research on data and encryption, as well as recent work on privacy-conscientious models for mobile data use renowned scientist and MIT Professor, Alex “Sandy” Pentland, with colleagues from MIT Connection Science & Engineering, has detailed four categories of privacy engineering solutions and design strategies for big data use.

FIGURE
OPAL as an example of Public-Private-People partnership

The Open Algorithms (OPAL) project is a socio-technological innovation developed by Data-Pop Alliance, Imperial College London, MIT Media Lab, Orange and the World Economic Forum to leverage private sector data for public good purposes by “sending the code to the data” in a privacy preserving, predictable, participatory, scalable and sustainable manner. It is designed to provide a far better picture of human reality to official statisticians, policymakers, planners, businesses, and citizens, while enabling greater inclusion and inputs of all on the kinds and use of analysis performed on data about themselves. The first piece is to send the algorithms to the data, through a secured platform, not the other way around, so that data is not exposed to theft and misuse.
The second step is to co-design of how big data algorithms are used, with inputs from local advisory Committees for the Orientation of Development and Ethics (CODE) so that these algorithms served local needs and respect local standards, instead of imposing external perspectives and expertise. OPAL is currently being deployed in Senegal and Colombia with funding from the French Development Agency (AFD).

Applied data “privacy by design” technical solutions

1. Distributed Data Repositories

Technical Design and Infrastructure
Model in which different kinds of data are stored in separate repositories and tools are deployed from an external or remote human query that can send queries to the correct data repository. This kind of model enforces individual privacy in allowing for the tracking and auditing of metadata associated with patterns of communication between repositories and queries.

Key design strategies and considerations for privacy
- Move data as little as possible.
- Unalterably record patterns of communications between databases and human operators, making these patterns publically auditable
- All data is encrypted wherever possible
- Use P2P architectures whenever possible.

2. Move the Algorithm to the Data (Distributed, Privacy-Maximizing Algorithms)

Technical Design and Infrastructure
Model in which each data repository locally provides query processing and data computation capabilities that enable remote queries to send their query statements or algorithms to the repository.

Key design strategies and considerations for privacy
- Keep data at its repository - never expose raw data.
- Use distributed query processing to send/route queries and sub-queries to correct repositories
- Each data repository returns only anonymous aggregated results
- Repository owner controls degree of privacy by controlling the granularity of answers

3. Data Always in Encrypted State: at Rest and in Computation

Technical Design and Infrastructure
Model in which data is always in an encrypted state (both in storage and during computations) and new cryptographic algorithms and approaches (such as Enigma and homomorphic encryption) allow operations to be carried out on encrypted data without need to decrypt first

Key design principles and considerations for privacy
- Raw data must remain encrypted during transit and storage
- Computation performed on encrypted data
- Provide controls to data owner

4. Encode Data Usage Agreements in Legal Trust Networks

Technical Design and Infrastructure
Trust network model for large scale data sharing in an ecosystem, which combines computer network tracking user permissions for each piece of data within a legal framework specifying actions and violations of data use.

Key design strategies and considerations for privacy
- Develop and deploy operational trust networks as the legal foundation for data access and data sharing
- Maintain strong audit of all data access and usage modes, with a tamper-proof history of provenance and permissions, to ensure that data usage agreements are being honoured
- Ensure a high degree of interoperability with existing trust frameworks that address relevant aspects of data sharing

Source: Thomas, Hardjono and Alex „Sandy“ Pentland, Connection Science & Engineering, Massachusetts Institute of Technology, “Preserving Data Privacy in the IoT World.” July 2016. http://resources.getsmarter.ac/other/preserving-data-in-the-iot-world-an-mit-report/

SECTION 2:

Focusing on responsibility in data use: Establishing internal responsible data governance standards to ‘do no harm’

Risks and Challenges

A needed shift towards data governance frameworks. While big data invites innovation towards new opportunities to understand behaviour and society, policymakers and practitioners must consider questions of governance—who categorises, who sorts, who intervenes, and who ultimately should be held accountable or responsible as personal data is used in these projects? Though several case examples have emerged where companies have experimented in sharing data with the public sector for social efforts, little has been proposed on what kind of governance is needed to allow such data sharing and use.

Principle of ‘do no harm’ and responsible data governance.

In ongoing global conversations on responsible data governance and sharing, researchers, policymakers and data sharing entities often bring up a common central question in regards to conceiving governing principles and minimum standards for responsible data collection and use: how can we, similar to doctors, “do no harm?” Just as Big Data inherently problematizes the use limitation principle of the former OECD fair information principles, understanding the effects of all such future uses of data and determining harm in the present has proved to be unrealistic and difficult. Researchers are attempting to create frameworks to understand the boundaries of what they can understand, while policymakers feel the administrative pressure to respond quickly and create protective regulations.

Solutions and Strategies

In order to even begin to gain ground on tackling this issue of “do no harm,” companies and other big data stakeholders—including privacy experts, policymakers, data scientists, data users, private sector and ethicists—need to co-create infrastructure and spaces for discussion, safe experimentation and transparent findings.

Internally, best practices that some companies have experimented regarding structures for responsible data use include:

- **Adaptation of ethical modalities used in other fields of research:** the use of institutional review boards (IRBs) (as in the case with Flowminder and the Karolinska Institute); development of codes of conduct (BBVA);
- **Internal ethical review processes and committees involving external actors (“Bring the ethicists to the table”):**
- **Creation of advisory boards and ethical review processes:** UN Global Pulse Privacy Advisory Group for Big Data for Development and Humanitarian Response; the development of Facebook’s internal ethics review processes and systems following its emotion contagion study [8]
- **Sending a data scientist-in-residence** to work within another company instead of sending data

[8] Jackman, Molly, and Lauri Kanerva. „Evolving the IRB: Building Robust Review for Industry Research.“ Washington and Lee Law Review Online 72.3 (2016): 442.

SECTION 3:

Keeping transparency, trust and user control at the centre: Engaging all data stakeholders

Risks and Challenges

EU citizens are willing to share their data if the benefits derived from the use and impacts of big data-driven innovation are transparent.

Need for transparency in public-private data partnerships to build public trust. Public-private partnerships generally involve the outpouring of capital and resources from the private sector; governance, oversight, and regulatory mechanisms led by the public sector; and some form of participation and feedback from the public. However, participants stressed that these partnerships in the age of Big Data require new kinds of participation. The nature of the resource--typically the collection and use of passively generated personal data--warrants greater stakeholder participation in directing the vision, framing and rationale for these partnerships. The framing needed to build such partnerships must go beyond solely where and how the data is shared, but what kinds of public problems should be solved together with all parties involved; if big data entails a certain amount of risk to personal data, to what extent do companies and the public voice what kinds of public projects warrant the use of their data? In his past experience directing the BBVA Innova Challenge, Professor Esteban Moro stressed that a key factor in building trust and public interest within these partnerships is focusing on solving significant public problems as communicated by the public.

As companies continue to navigate their role in the data revolution, transparency in their efforts remains essential. A study from the Vodafone Institute highlighted European user uncertainty and scepticism of the benefits derived from the use of their information as a part of big data initiatives. Surveying more than 8,000 European digital users, the Vodafone Institute study suggests that users are "sceptical of the Big Data phenomenon because public and private organisations are failing to explain clearly how and why their data is analysed, and do not give them adequate control over how their data is being used."

However, when asked about sharing their data and impacting privacy, users were willing to share personal data if the personal or societal benefits were made clear; for instance [9]:

- "53% said that they wouldn't mind their data being analysed if it would help them or other people to improve their health;"
- "68% stated that they were in favour of smart meters to record data on building residents' usage behaviour so that more eco-friendly heating practices could be introduced; and
- "55% said they were happy about data from their cars being transferred in order to receive personalised traffic reports."

Lack of shared, simple language to facilitate data literacy among citizens and encourage stakeholder collaboration and participation in data-driven projects. Infomediary-led initiatives, platforms and engagement events have an important role in fostering shared language around big data use. Marco Bressan, Chief Analytics Officer at BBVA, highlighted the diversity of actors involved in big data and development projects and emphasised how a lack of shared language among companies, regulators, and universities inhibits effective collaboration and citizen participation. Data infomediaries can play a role in helping gathering cross-sector insights, create platforms for meaningful discussion, and identify potential areas for further collaboration.

[9] http://www.vodafone-institut.de/2015/09/vodafone-institute-survey/

Solutions and Strategies

Best practices discussed during the roundtables for promoting transparency and accountability in big data-driven innovation projects include:

- **Inclusive and participatory governance frameworks** (including participatory frameworks, opt-in and opt-out consent options): In addition to providing information and involving government and civil society representatives in project scoping and planning, providing an understandable consent experience for users remains an essential transparency and accountability mechanism for empowering and including users. This must include communicating the user's right to withdraw consent at any time and providing meaningful options for opt-out without losing service.
- **Multi-stakeholder public dialogues and data-driven user engagement campaigns:** Big data is often described in the extremes within popular media: either as a kind of panacea solving major global problems or harbinger of government surveillance, unwanted advertising and the end of privacy. Public dialogues allow citizens to hear from experts and have their concerns heard. In addition to government, academic and industry experts, greater inclusion of data infomediaries, the media and other stakeholders in the European data revolution will help sustain new data partnerships by raising the level of public discourse towards areas of needed civic dialogue. While many citizens may not be able to specifically work with data, infomediaries can play a role in helping foster a new kind of civic literacy in the age of big data that focuses on their ability to constructively engage in discussions about their data, their digital rights and the kinds of projects and partnerships being established in the name of public good. Additionally, these dialogues can provide opportunities to discuss project impact and results, and provide citizen-oriented data outputs to help citizen's engagement with the value added by their participation in the project.
- **Promotion of broader data literacy efforts to raise level of awareness (as technologies and possibilities evolve):** Universities and civil society actors continue to lead global data literacy efforts through initiatives, platforms and engagement events on building data skills and providing training. These activities have an important role in fostering shared language around big data use. The diversity of actors involved in big data and development projects often results in a lack of shared language among companies, regulators, and universities, which inhibits effective collaboration and public participation. Data infomediaries can play a role in helping gathering cross-sector insights, create platforms for discussion and identify potential areas for further collaboration. Companies can work with data infomediaries to help promote data literacy efforts of their project and broadly raise the level of awareness around data and information.

ROADMAP

How to initiate big data-driven innovation projects?

Building on the three main areas discussed in the workshops, the following roadmap highlights the requirements and milestones for companies aiming to initiate big data-driven innovation projects.

Invest in building big data internal knowledge and support

1.
Focus projects and resources on solving public problems with clearest ethical imperative and problem definition.
Identify issue areas and public problems in which your data can play an important role towards decision-making and/or resource optimisation. Kenneth Cukier during the Brussels public event described "not using data [as] the moral equivalent of burning books," and described several case studies in which big data could have been leveraged by the right actors. Additionally, look for "low hanging fruit" – data from your company could even play a small but critical role in a specific issue area rather than solve a much larger problem.

2.
Analyse current risks and trade-offs in various forms of data sharing to make the case for internal support.
Evaluate existing case studies in which companies have previously shared similar data. What modalities did they use to share data? How have others handled privacy issues in using data towards this particular issue? What kinds of risks and trade-offs would exist in the creation of this specific project? How can this project both make social impact as well as support the company's overall objectives? This could take the form of sponsoring or hosting a series of expert workshops, participating in larger forums on big data and social impact or soliciting help from external consultants or researchers who would be embedded into your organisation and teams. Ultimately, answering these questions internally is critically both for gaining support as well as building internal momentum and enthusiasm.

3.
Recognise organisational challenges and constraints in initiating innovation.
What resource constraints exist for you company? How much funding or support is available and what kind of projects will be possible?

4.
Invest in knowledge management and implementation of privacy engineering solutions.
Evaluate the existing legal framework around privacy and learn more through existing solutions how to incorporate privacy-preserving elements into your project. Invest in developing mutually beneficial big data partnerships.

5.
Catalyse mutually beneficial partnerships with the right stakeholders with diversity of expertise and capacity for thought leadership.
Several participants stressed that many of the existing big data projects would not exist outside of years of building trust and taking steps (and risks) with the right stakeholders. Identifying the right stakeholders involved evaluating expertise and thought leadership, as well as willingness to take on risks and process alignment.

6.
Focus partnerships on determining best use of private sector data to enhance existing traditional data.
Using data sources such as call detail records, transaction data and other new data sources will require some form of ground truth data from existing traditional sources. Identify partners that can also provide data or capacity to analyse data from public or open sources.

7.
Adapt features of existing ethical frameworks and guiding principles toward initial "do no harm" governance model.
While companies may aim to "do no harm" in the development of their projects, what guiding principles or ethical frameworks govern their use, and what happens when problems arise? While researchers continue to develop and assess ethical frameworks for data use (e.g. the Menlo Report), several companies have experimented with various models to incorporate ethical standards into their projects. This has involved the development of codes of conduct (in the case of BBVA, for example), organising ethical roundtables , fostering research on ethical data use (such as the report on the implications of big data on the right to informational self-determination by philosopher Anna Wehofsits of Ludwig-Maximilians-University, published by the Vodafone Institute) and incorporating an ombudsman to oversee data sharing projects.

8.
Create and encourage mechanisms for public feedback and consultation.
Consult stakeholder groups to help inform the project design and provide opportunities for the public to give feedback.
Invest in supporting meaningful civic engagement through multi-stakeholder big data literacy promotion

9.
Communicate project results through data visualisations.
Evaluate how to best communicate project results for both intended beneficiaries and the public, using infographics and data visualisations for example.

10.
Enable meaningful mechanisms for opt out.
Assess applicable notice and consent solutions and provide option for users to remove their data from use in the project.

11.
Promote opportunities for shared language and principles among stakeholders through participation in cross-sector, cross-disciplinary collaboration, platforms and events led by intermediaries.

12.
Invest in long-term civic engagement and public understanding of how data is being used through data literacy efforts through media, government and civic outreach.

Glossary

CDR	Call detail records
DII	demographically identifiable information
GDPR	General Data Protection Regulation. A regulation by which the European Parliament, the Council of the European Union and the European Commission intend to strengthen and unify data protection for all individuals within the European Union (EU).
GPS	Global Positioning System
PII	personally identifiable information

About

Data-Pop Alliance

Data-Pop Alliance is a global coalition on Big Data and development created by the Harvard Humanitarian Initiative (HHI), MIT Media Lab and Overseas Development Institute (ODI) that brings together researchers, experts, practitioners, and activists to promote a people-centred Big Data revolution through collaborative research, capacity building, and community engagement.

www.datapopalliance.org

Vodafone Institute

The Vodafone Institute for Society and Communication explores the potential of future technologies to improve society. The Institute is a thinktank that fosters dialogue between science, business and politics. It initiates projects and research, and publishes reports as a source of practical recommendations for decision makers.

www.vodafone-institut.de

Imprint

Authors David Sangokoya and Emmanuel Letouzé, Data-Pop Alliance.
Editor Vodafone Institute for Society and Communications, Behrenstraße 18, 10117 Berlin, Germany. **Chairman of the Advisory Board** Joakim Reiter.
Project Lead Alice Steinbrück. **Editorial Management** Friedrich Pohl.
Research Support Vivian Weitzl. **Layout** Nick Böse, Stefanie Rosemeyer.
Initiated i.a. Dr. David Deißner.

www.vodafone-institut.de
www.facebook.com/VodafoneInstitute
@vf_institute