Submission for IACAP 2015 (3000 words, excluding references)

Beyond Informed Consent – Investigating Ethical Justifications for Disclosing, Donating or Sharing Personal Data in Research

Markus Christen, Centre for Ethics, University of Zurich, Switzerland

Josep Domingo-Ferrer, UNESCO Chair in Data Privacy, Universitat Rovira i Virgili, Catalonia

Dominik Herrmann, Security in Distributed Systems Group, University of Hamburg, Germany

Jeroen van den Hoven, Philosophy Section, Delft University of Technology, Netherlands

Introduction

In the last few years, we have experienced a tremendous growth of the digital infrastructure in many countries, leading to an emerging web ecosystem that (i) involves a variety of new types of services, (ii) modifies the competition dynamics of companies, and (iii) blurs traditional boundaries both on the economical and societal level. A characteristic element of this web ecosystem is the massive increase of the amount, availability and interpretability (in terms of available tools) of digitalized information – a development for which the buzzword "Big Data" has been coined. The EMC Corporation estimated at the end of 2012 that the "digital universe" – the digital data created, replicated, and consumed in a single year – is currently doubling every two years and that the "zettabyte-mark" (10^{21} bytes) of the digital universe has been reached already in 2010. A significant part of these data is somehow related to people and can be used to infer otherwise private information about them.

The classic answer to this problem is to ensure that the individual has control over his or her personal data. For example, in 2012, the European Commission proposed a new legislation in the form of a regulation that is meant to replace the Data Protection Directive. The key changes include, for instance, increased responsibility and accountability for those processing personal data and a requirement for explicit consent for processing activities. Key provisions in this text – such as the Right to be Forgotten and the Right to Data Portability – clearly illustrate the goal to 'put citizens back in control of their data'. However, many of the new or modified provisions in the Regulation have been criticized. On top of that, uncertainties remain regarding the practical implementation of many of the new provisions. It is not entirely clear, for example, what the terms 'erasure' or 'anonymisation of personal data' mean exactly (Article 29 Working Party 2012), or whether they are even technically possible at all (Druschel et al. 2012).

Beyond these practical issues, however, remains the question whether this approach that focuses on control and consent is adequate to the deeper changes that result from Big Data and the associated digital technologies. After all, one of the novel ideas found in Big Data research is to work with data that has been collected for a different purpose in order to uncover surprising or valuable information. As Tene & Polonetsky (2012) observe, it can be very difficult to anticipate at the time of collection for what kind of analyses some data will be used in the future. The following considerations are based on the assumption that one of the most profound effects of this digitalization of information in all spheres of life is that the boundaries around which human beings used to conceptualize and organize their social, institutional, legal and moral world have been torn down, compromised or relativized. While the social online world tends to mirror the offline world, the traditional offline distinctions and demarcations of separate social realms (family and friendship, work, politics, education, commercial activity and production, health care, scientific research, etc.), each governed by context-relative norms, policies and rules, are threatened by the enhanced reproducibility and transmissibility of online data. What we had reasons to care about from a moral point of view in the offline world in these domains cannot be simply sustained and reproduced in a straightforward way in a digital age, which comprises both online and offline, and emergent interactions between both. Individual users of digital platforms are only partially aware of these effects, but they begin to appreciate the erosion of social meanings and the frailty of traditional social norms in the digital domain. Affected are core notions like 'informed consent', 'personal information', 'anonymity' or 'privacy' as well as their underlying foundational values like 'autonomy', 'responsibility' and 'fairness'.

The goal of this contribution is to briefly outline the possibilities and limitations of the classic idea of individual control and consent regarding the use of personal data and to investigate ethical justifications that may support disclosing, donating or sharing personal data, with a focus on using such data in research. This will be done in three steps: First, it is assumed – following several other scholars – that the practice of the 'art of separation' or the maintenance of 'contextual integrity' is a key moral issue that is at stake due to the recent developments in the field of Big Data. Second, it is argued that the core value of autonomy (which provides the moral foundation of control and consent) cannot support the defense of privacy by itself, but must be complemented with two other core values – responsibility and fairness – in order to sufficiently describe the moral landscape of the problem under investigation. Third, it is drafted how research relying on (potential) personal data could proceed in order to comply with these values.

Contextual integrity and its undermining

In 1983, the political philosopher Michael Walzer introduced the idea of *spheres of justice*, which proposes that societies consist of different social spheres (e.g., medical, political, market, family and educational) each defined by a different type of good that is central to that particular sphere. These different types of goods (e.g., medical treatment in the medical sphere, political responsibility and public office in the political sphere) and the meaning and significance they have in each of these spheres, have their own associated criteria, principles and mechanisms concerning their distribution and allocation. In order to prevent mixing up of distributional criteria and goods from different spheres (and prevent, e.g., allocation of seats in parliament on the basis of financial assets or family relationships or health condition, or making one's ranking on a waiting list in health care dependent on family relationships or college degrees) these spheres have to be kept separated. Walzer refers to the situation where advantages and positions regarding the distribution of a good in one sphere cannot be automatically converted in advantages in another sphere and each sphere and sphere internal moral considerations are given their due weight, with the term *Complex Equality*. This idea of complex equality captures an important aspect of what we mean by 'fairness' and it implies amongst other things that the distribution of access to particular goods tracks the sphere's specific normative considerations (e.g., 'need' in the medical sphere, 'democratic election' in the political sphere). Goods have to be distributed along the mechanisms of the corresponding sphere and goods from different spheres ought not to influence each other in terms of distribution. Put differently, this means that the exchange of goods between spheres has to be "blocked" in order to preserve complex equality. Walzer talks about "blocked exchanges" and the "art of separation". The same ideas regarding social differentiation and quasi-autonomy of social realms with their own internal goals, values and allocation schemes can be found in the work of many other political and social theorists.

Walzer's work has been applied to the realm of information systems by Van den Hoven (1997, 2008) and Nissenbaum (2004). Nissenbaum coined the term *contextual integrity* to refer to this idea, which she considers an "alternative benchmark for privacy, to capture the nature of challenges posed by information technologies" (Nissenbaum, 2004). Contextual integrity thus comprises a wider range of social spheres than the often-applied dichotomy of public and private. Instead, spheres are defined through the expectations and behavior of actors that differ per sphere. In order for contextual integrity and sphere separation to be achieved, the type of information that is revealed and the flows between different parties have to be appropriate for the context. Van den Hoven (2008) considers four different moral reasons to constrain flows of information. Next to the prevention of inequality based on Walzer, he points to information-based harm (e.g., through discrimination), the exploitation in markets and moral autonomy.

The challenge of "Big Data" is that since information produced within these spheres (health, politics, criminal justice, market) travels much faster and is more difficult to control (and to greater distances) than in the traditional offline world, we face a set of phenomena that threaten the integrity of social spheres and the cultural and social meanings expressed in them, including our values. Of course the boundaries between spheres are to a certain extent relative to time and culture, and not carved in stone forever, but it is important to note that every age, society and culture does in fact draw and treat these boundaries – construed as sets of constraints on the flow of information – as of high normative relevance. This implies that changes to them need to be morally justifiable.

From a purely technological perspective, it becomes more and more obvious that the integration of heterogeneous data describing the activity of individuals in different social spheres enable detailed inferences on the individual. As it is possible to merge different sources of data (e.g., this is the core business of data brokers, among others, see Anthes 2015), this requires studying new methodologies for privacy risk evaluation and the definition of privacy transformations suitable for addressing the multidimensional character of the data. In the literature, there exist some works on the identification of privacy risks in social network data. Examples include the problem of linking users across different platforms, e.g., Liu and colleagues (2014) who computed the similarity among users by analyzing both generated content and top-k friends. Kosinski et al. (2013) demonstrate that it is possible to infer demographic properties and traits from the set of pages a user "likes" on Facebook. Malhotra et al. (2012) studied a way to construct digital footprints using information retrieval for name disambiguation. Vosecky and colleagues (2009) proposed a method to identify users based on profile matching (either exact or partial). Nunes and colleagues (2012) collected user profiles and, for each dimension of the profile field (e.g., username, picture, location, occupation, etc.), they reduced the problem of user identification to a binary classification task. Jain and colleagues (2013) proposed identity search algorithms to find a user's identity on Facebook, given her identity on Twitter.

Based on such "reconstructions" of individuals, discrimination may occur, which refers to an unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category disregarding individual merits. Unfair decisions have been observed in a number of settings, including credit, housing, insurance, personnel selection and worker wages, web advertising and recommendation (Romei & Ruggieri 2014). Here, a first crucial problem is *discrimination discov-*

ery, i.e., defining methods capable of providing evidence of discriminatory behavior in activities such as the ones listed above. The legal principle of under-representation has inspired existing approaches for discrimination discovery based on frequent pattern mining (Ruggieri et al. 2010). A number of approaches have been recently proposed to tackle both privacy and non-discrimination risks in disclosing data and models (Hajian et al. 2014). Another source of complexity is when data do not explicitly contain an attribute denoting possibly discriminated groups. This case is known as indirect discrimination analysis (Hajian & Domingo-Ferrer 2013). A well-known example is redlining discrimination analysis, occurring when the ZIP code of residence is correlated with the ethnicity of individuals in highly segregated regions. The second crucial problem is *discrimination prevention*, that is preventing discriminatory decisions by automatic decision-making algorithms based on data mining. Discrimination prevention consists of extracting predictive data mining models, profiles, or recommendations that trade off accuracy with non-discrimination. There is a blooming research on this problem in the field of data mining, see e.g., the collection edited by Custers et al. (2013). A recent paper by Berendt and Preibusch (2014) has conducted a usability test methodology based on Amazon Mechanical Turk to assess the effectiveness of discrimination-aware approaches. These developments show that the technical capabilities for undermining the contextual integrity of data as well as detecting such integrity breaches are growing, although the former probably to a faster extent than the latter.

Affected values

The ideal of autonomy (a.k.a. informational self-determination, that is, the ability of persons to use digital technology in a self-determined and informed way) is often quoted as the indispensable precondition for personal data management. Closely associated to this value is thus the ideal of informed consent, in particular when disclosing information due to using some digital services or when sharing data with third persons. However, the recent developments make it questionable that the consent route is a sufficient and meaningful expression of autonomy in the context of Big Data, in which the amount of information extracted from data (including the elaboration of meta-data) might exceed ex-ante expectations of both users and platform administrators. Furthermore, when individuals use digital platforms, they are often in a position of informational asymmetry: they are not aware of the various informational links between social spheres that are generated in this way and that allow for unexpected benefits and control possibilities by platform providers. The orientation on autonomy puts the focus on the individual and disregards the moral obligations of the other players involved in Big Data.

In the following, it is proposed that the following three values provide a better outline of the moral landscape:

1. **Autonomy:** Users ought to be aware of how their data records are used in order to promote their values and gain control over privacy-related choices.
2. **Responsibility:** Users (both researchers and data providing research subjects) should be held responsible and accountable for the ways in which they use their personal information and the information about other people. If some subjects are wronged, it must be possible to attribute personal responsibility for the wrongs in question.

3. **Fairness:** The benefits of knowledge and information ought to be fairly apportioned to all participants in interactions, so as to rule out inequality of opportunity and exploitation by some at the expense of others.

These guiding values provide a broader in-depth analysis of the main types of moral concerns in the domain of data protection: informational harm, economic disadvantage, discrimination, and threats of self-presentation & identity (Van Den Hoven 2008; Van den Hoven et al. 2012).

Let us explain this point by some examples. Online behavior of users is tracked by advertisement agencies, in order to display more relevant ads. This so-called "behavioral targeting" is commonplace on the Internet today (Hoofnagle et al. 2012). Suppose that this service comes along with immediate benefits in non-material form (recommendations). One concern is that – based on consumer behavior –, the agencies learn habits and personal traits of users that can be used for price discrimination or "price gauging", or that some items might even not be offered (Turow 2011). For example, certain types of users, but not others, are offered special discounts for ordinary consumer products. Or in another case, it could be that an online health insurance provider offers a contract at a higher price.

This is a form of discrimination and relates to the value of *fairness*. Forms of discrimination are not necessarily unethical *per se*, but have to be addressed and analyzed with respect to their justification and counteracted if unjustified. It could be that if a consumer is facing price discrimination in ordinary consumer products, it is up to the user, considered as an autonomous agent, to strike a balance between the potential benefits and the harms of informational exposure. This ethical analysis emphasizing *autonomy* can be matched by a technology that enhances *awareness*, by measuring the informational exposure of the consumer, and other ways to help him or her understand the way his or her information might be used to predict potential harm that he or she faces. These are all necessary steps for promoting more informed decisions, related to the value of *responsibility*.

However, in considering the case in which a health insurance provider is involved, the ethical analysis might take a different course, since the *(contextual) integrity* of two spheres – shopping and healthcare – is violated. In this case the evaluation of the appropriate ethical response may be a form of *empowerment,* which could be promoted by a technology for anonymization and de-linking, or, alternatively, through a policy proposal, such as *extending* the rights of citizens in the digital domain, or by ensuring *accountability* of data mining by advertising agencies.

Disclosing data in online research

An in-depth ethical analysis based on this roughly drafted framework certainly strongly depends on the type of problem under investigation. In the following, the focus will be on research that relies on personal information emerging from individuals – either gained directly (e.g. through surveys or offering possibilities to donate data) or indirectly (e.g., by data mining in social networks). As research often aims to combine data emerging from different social spheres in order to answer specific research questions (e.g., the interrelation of social status and health), the issue of contextual integrity is of particular relevance for researchers that handle such data.

Using the framework above, it is claimed that a research infrastructure that harvests and manages personal data should provide the following functionalities:

- **Autonomy:** Enable research participants to gain awareness on what guides their choices (privacy preferences) and on what they potentially may disclose when providing certain types of data. Shift away the focus from (mere) informed consent towards empowering research participants and data donators.
- **Responsibility:** Ensure longer-term relations between participants and researchers through an infrastructure (social network) that allows for bidirectional relations (e.g., for suggesting new research questions by participants, participant-driven research). Empower the researcher both regarding legal / ethical requirements and technical instruments (e.g. for data anonymization) for doing responsible research with personal data. Empower the participant with the ability to verify how safe is the anonymization performed by the data collector/researcher.
- **Fairness:** Provide a broader set of utilities (not only monetary compensation) like visualizing the contribution of research participants, e.g. through donated data, to certain scientific results. Create novel types of interactions (using, e.g., co-private protocols, Domingo-Ferrer 2011, and, more generally, co-utile protocols, Domingo-Ferrer et al. 2015) that allow collaborative contribution to a common good (like ensuring each other's privacy).

An extended conference paper will outline these requirements in more detail.

Remark: This extended abstract is based on a research proposal that has been submitted in the context of Horizon 2020 and that includes researchers from the following institutions: Universität Zürich (PI), Technische Universiteit Delft (Netherlands), Universitat Rovira i Virgili (Catalonia), Università di Pisa (Italy), Universität Hamburg (Germany), Katholieke Universiteit Leuven (Belgium), Berner Fachhochschule (Switzerland), JonDos GmbH (Germany)

References

Anthes G (2015). Data brokers are watching you. Communications of the ACM 58(1):28-30.

Berendt B, Preibusch S (2014). Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. Artificial Intelligence and Law: 1-35.

Custers HM, Calders T, Schermer BW, Zarsky TZ (ed.) (2013). Discrimination and Privacy in the Information Society, vol. 3 of Studies in Applied Philosophy, Epistemology and Rational Ethics. Berlin/London: Springer.

Domingo-Ferrer J (2011). Coprivacy: an introduction to the theory and applications of co-operative privacy. SORT-Statistics and Operations Research Transactions 35: 25-40.

Domingo-Ferrer J, Soria-Comas J, Ciobotaru O (2015). Co-utility: self-enforcing protocols without coordination mechanisms. Proc. of the 2015 Intl. Conf. on Industrial Engineering and Operations Management, IEEE (to appear).

Druschel P, Backes M, Tirtea R (2012). The Right to Be Forgotten – between Expectations and Practice. ENI SA. Available at: https://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/the-right-to-be-forgotten/

European Data Protection Regulation (2012): The most recent version is available at the following website: http://ec.europa.eu/justice/data-protection/index_en.htm

Hajian S, Domingo-Ferrer J (2013). A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. IEEE Transactions on Knowledge and Data Engineering 25(7): 1445-1459.

Hajian S, Domingo-Ferrer J, Farràs O (2014). Generalization-based privacy preservation and discrimination prevention in data publishing and mining. Data Mining and Knowledge Discovery: 1-31.

Hoofnagle CJ, Soltani A, Good N, Wambach DJ, Ayenson M (2012): Behavioral Advertising: The Offer You Cannot Refuse. Harvard Law & Policy Review, 6:273–296.

Jain P, Kumaraguru P, Joshi A (2013). @i seek 'fb.me': identifying users across multiple online social networks. WWW (Companion Volume) 2013: 1259-1268.

Kosinski M, Stillwell D,Graepel T (2013): Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences. 110.15: 5802–5805.

Liu K, Terzi E. (2009). A framework for computing the privacy scores of users in online social networks. In Proceedings of ICDM 2009, The 9th IEEE International Conference on Data Mining: 288-297.

Malhotra A, Totti L, Meira Jr. W, Kumaraguru P, Almeida V (2012): Studying User Footprints in Different Online Social Networks. Proceedings of ASONAM 2012: 1065-1070. arXiv:1301.6870

Nissenbaum H (2004). Privacy as contextual integrity. Washington Law Review 79: 119-157.

Nunes A, Calado P, Martins B (2012). Resolving user identities over social networks through supervised learn-ing and rich similarity features. Proceedings of the 27th Annual ACM Symposium on Applied Computing: 728-729.

Romei A, Ruggieri S (2013). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review: 1-57.

Ruggieri S, Pedreschi D, Turini F (2010). Data mining for discrimination discovery. ACM Transactions on Knowledge Discovery from Data 4(2): Article 9.

Tene O, Polonetsky J (2012): Privacy in the Age of Big Data A Time for Big Decisions. Stanford Law Review Online 64 (63): 63–69.

Turow J (2011): The daily you: how the new advertising industry is defining your identity and your world. Yale University Press, New Haven.

Van den Hoven J (1997): Computer Ethics and Moral Methodology. Metaphilosophy 28(3): 234-248.

Van den Hoven J (2008). Information technology, privacy, and the protection of personal data. In: van den Hoven J, Weckert J (eds.): Information technology and moral philosophy. Cambridge, New York: Cambridge University Press: 301-321.

Van den Hoven J, Helbing D, Pedreschi D, Domingo-Ferrer J, Gianotti F, Christen M: FuturICT - The Road towards Ethical ICT. European Physical Journal - Special Topics 214: 153–181.

Vosecky J, Hong D, Shen V (2009). User identification across multiple social networks. Proceedings of the First International Conference on Networked Digital Technologies, NDT '09.

Walzer M (1982): Spheres Of Justice: A Defense Of Pluralism And Equality. New York City: Basic Books.