Physical sciences / Engineering / Electrical and electronic engineering URI /639/166/987

Scientific community and society / Scientific community / Policy URI /706/648/453

Data protection in the age of Big Data

Europe's data protection laws must evolve to guard against pervasive inferential analytics in nascent digital technologies such as edge computing.

Sandra Wachter

With modern interconnected digital technologies, data is not so much knowingly created by the user as it is observed or 'captured' by devices and services in the course of normal use. Networks of sensors, for example, can discreetly collect usage and behavioural data, often in previously unobserved private settings (such as in the home or when we are asleep) in order to anticipate and respond to our needs (adjusting temperature or lighting, for example) using machine learning and artificial intelligence (AI). Data from fitness trackers, internet browsing, mobile phones (geolocation), and countless other devices can also be seamlessly and tirelessly captured. The extent and potential value of the data collected remains unclear to users and, thus, the establishment of a data protection regulation that seeks to provide individuals with oversight and control over their personal data in order to protect privacy could not have arrived at a better time. This new legal framework, called the General Data Protection Regulation (GDPR), was adopted in Europe in April 2016 and enforced in May 2018.

Unfortunately, many nascent digital technologies seem destined to undermine the aims of GDPR. Digital services and distributed devices now increasingly operate on a linked-up basis, in which information is shared between networks of devices and service providers, making use of unique user identifiers to provide seamless data sharing and personalised experiences using machine learning and AI. Such seamless experiences are rooted in identification technologies used to manage authentication, data access and transfer, and link together disparate sources of user data for inferential analytics. Clearly, these hubs can be of significant commercial value. In 2017, a consortium of some of the biggest tech firms recently announced the Data Transfer Project, an initiative designed to facilitate exercise of users' right to data portability under the GDPR via common data interoperability standards and data transfer mechanisms.

Edge computing, for example, realised via the Internet of Things (IoT), cyber-physical systems (such as autonomous or connected cars) or smart cities are built on precisely the premise of linked up data**.**. Health tracking devices collect data that can be used to infer health status, sleeping habits, levels of exercise or general wellbeing. In smart cities, sensor data, WiFi data (used for smart transport services, traffic management, contactless payment) and location data allow insight into movement patterns as well as business and leisurely activities. Drivers of cars with networked sensors are subject to behavioural profiling (shopping habits and social networks) based on their journey information. It is also possible to infer health status and wellbeing via eye tracking, facial recognition software or heart rate measurements used in fatigue detection systems[1]. Similarly, smart phone sensors and data can predict stress levels, eating habits, mental illness and demographics (such as age and gender). This can be achieved via smart sensors used in accelerometers, GPS, microphones, cameras or fingerprint identification to infer exercise levels, health status, and mood, or by analysing phone usage patterns such as browsing habits, preferred apps or social interactions[2]. Data describing the private lives of users can increasingly be pervasively captured, shared, and analysed to infer characteristics, behaviours, and unmet needs of users.

The types of data produced by such technologies can be used to draw non-intuitive and unverifiable inferences and predictions about the behaviours, preferences, and private lives of individuals. These inferences draw on highly diverse and feature-rich data of unpredictable value, and create new opportunities for discriminatory, biased, and invasive decision-making, often based on sensitive attributes of individuals' private lives[3]. The common factor among these risks is the fear of privacy pervasive data collection that allows sensitive inferences to be drawn. These inferences can lead to discrimination, especially when shared with third parties such as insurance companies, financial institutions, or employers.

Arguably all these areas fall within the scope of the new EU's General Data Protection Regulation. However, the law uses ineffective and outdated strategies to address them. First, the GDPR focuses mainly on protection at the input stage when data is collected, but hardly during or after analysis. The law thus ignores the fact that unforeseen threats to privacy can arise *after* data collection owing to inferential analytics. Second, even though the goal of data protection law is to protect privacy and identity, the law hardly regulates how and according to which parameters the data is assessed and evaluated. Assessments of individuals (e.g. predictions on work performance, financial liquidity, life expectancy) thus mostly fall outside the scope of the GDPR. Instead, the law grants varying standards of protection, defined against artificial and fluent categories reflecting the status of the data at the point of collection.

Anonymised data. None of the rights in the GDPR will apply to anonymised data. This is problematic for two reasons. First, any de-identified data can theoretically be reverse engineered and linked back to an individual.[4]Second, even truly anonymised data can be used to build user profiles and thus privacy and discrimination harms still occur[5], without the need to identify a particular individual. Here again data protection law focuses on the stage of data collection rather than on how the data is used, and its effects on relevant individuals and groups[6].

Personal data. Whilst the GDPR is designed to protect personal data, the European Court of Justice is not clear whether or not inferences fall under this definition. The jurisprudence is inconsistent: a 2014 judgment[7] clearly excluding inferences from the safeguards of data protection law, while a later judgement[8] in 2017 attributes the status of 'personal data' to inferences. However, even in the latter case the Court did not grant all the rights associated with this status and made it clear that data protection law does not include a right over how individuals are assessed. Rather, the law is designed only to ensure that the input data is lawfully obtained.

Sensitive data. European data protection law affords greater protection (e.g. higher standards for consent, limited permitted uses) to processing of sensitive data, or 'special categories', describing characteristics such as health, ethnicity, or political beliefs. When personal data can be shown to allow for sensitive attributes to be inferred, or 'indirectly revealed', the source data from which sensitive inferences can be drawn can also be treated as sensitive data. However, the Court[9] again limits the application of Art 9 GDPR (which defines 'special categories' of data) to cases where there is an intention to infer sensitive information, and when the data source is reliable to draw that inference. Both these conditions are detrimental to privacy protection in the age of Big Data. The intention to infer sensitive attributes is irrelevant. Proxy data such as postcodes contain sensitive data (for example, gender, sexual orientation or race) regardless of whether these attributes are intentionally or explicitly inferred. Everything is potentially sensitive data, we just do not know it yet[10]. Furthermore, whether or not a data source is reliable a reliable basis to draw sensitive inferences is irrelevant for the person concerned. If someone is incorrectly categorised as a woman and experiences discrimination as a result, the accuracy of the classification is irrelevant to its impact.

In a world of seemingly ubiquitous data collection via edge computing, and expansive knowledge generation using inferential and predictive analytics, as well as greater sharing of this knowledge

between public and private parties, the remit of data protection law must be redefined. Outdated, ineffective and fluid categorisations of data as personal or non-personal and sensitive or non-sensitive must be abandoned. These categories reflect only the nature of the data at the time it is collected, but ignore its subsequent usage and potential transformations (for example, inferring sexual orientation, health status or gender). The potential risks to data subjects do not end at the time data is collected and, therefore, data protection laws fail to guard against the potential harms of inferential analytics. As I have recently argued elsewhere, the age of Big Data calls for a "right to reasonable inferences."[11] that governs the responsible and normative acceptable use of data, ethical data sharing practices, and appropriate legal remedies in cases of harms. As the benefits and risks greatly differ depending on the sector (private, public health, transport, or finance) and the specific application, further research is required to flesh out societal acceptable standards of reasonableness. The future of edge computing requires a dialog between developers and society that does not only focus on what is technically possible, but also what is reasonable.

Sandra Wachter

Research Fellow

University of Oxford

Oxford Internet Institute

1 St Giles, Oxford, OX1 3JS, UK

The Alan Turing Institute

96 Euston Road, London, NW1 2DB, UK

e-mail: sandra.wachter@oii.ox.ac.uk

References

1. Wachter, S. The GDPR and the Internet of Things: a three-step transparency model. *Law Innov. Technol.* **10**, 266–294 (2018).
2. Peppet, S. R. Regulating the internet of things: first steps toward managing discrimination, privacy, security and consent. *Tex Rev* **93**, 85–176 (2014).
3. Wachter, S. Normative Challenges of Identification in the Internet of Things: Privacy, Profiling, Discrimination, and the GDPR. *Comput. Law Secur. Rev.* **34**, 436–449 (2017).
4. Ohm, P. Broken promises of privacy: Responding to the surprising failure of anonymization. *Ucla Rev* **57**, 1701–1777 (2009).

5. Hildebrandt, M. Profiling and the identity of the European citizen. in *Profiling the European citizen* 303–343 (2008).
6. Mittelstadt, B. From Individual to Group Privacy in Big Data Analytics. *Philos. Technol.* **30**, 475–494 (2017).
7. European Court of Justice. *YS, M and S v Minister voor Immigratie, Integratie en AsielJoined Cases C-141/12 and C-372/12*. (2014).
8. European Court of Justice. *Peter Nowak v Data Protection Commissioner Case C-434/16*. (2017).
9. *Kathleen Egan and Margaret Hackett v European ParliamentCase T-190/10*. (2012).
10. Zarsky, T. Incompatible: The GDPR in the Age of Big Data. *Seton Hall Law Rev.* **47**, (2017).
11. Wachter, S. & Mittelstadt, B. A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI. *Columbia Bus. Law Rev. Forthcom. 2019* (2018).