Elisa Bertino
Purdue University

The Quest for Data Transparency

Past and recent cases of data misuses and data breaches, and the central role that data plays today for artificial intelligence (AI) systems, are pushing the notion of data transparency at the forefront in several contexts. However, data transparency is still patchy and does not even have a comprehensive definition; rather, there are different definitions. Two notable definitions include: 1) the ability to access and work with data no matter where it is located and 2) the guarantee that the data being provided is accurate and from some official source. Obviously, those definitions are no longer adequate. The first one does not capture elements relevant to data privacy, trustworthiness, and fairness. The second one captures some elements of transparency that are today very relevant. However, it misses some elements, notably elements related to the use of data. Other recent definitions have been proposed that mainly focus on human data for legal purposes. As a result, they tend to favor the concept of informed consent for data providers, that is, subjects whose data are collected and used. While informed consent is vital for data transparency, a true general definition of data transparency must include other dimensions and consider groups of individuals who are often ignored.

Therefore, I would like to put forward the following broader definition:[1] "Data transparency is the ability of subjects to effectively gain access to all information related to data used in processes and decisions that affect the subjects." This definition covers several important cases. The first and more obvious case is the one in which data about an individual is collected and used in some processes that affect the individual. A relevant example of such of a process is the provisioning of the individual's data, perhaps together with data of other individuals, to a third party. Such a process may result in privacy breaches affecting the individual. Providing information about such a process is thus within the scope of data transparency.

The second case is the one in which decisions are taken about an individual, or processes are executed concerning the individual, and these decisions and processes are based on data that does not include the data of the individual. An example is a decision about an individual taken based on a recommendation by a classifier built using data from a population sample that does not include the individual's data. In such a case, providing information to the individual about the characteristics of the data set used to train the classifier is in the scope of data transparency even though the data set does not include the data of the individual in question. Information about such characteristics would allow one to assess important data ethical questions, such as whether the data set is biased against certain demographics or whether the data set collection and use was approved by some ethical board.

"Data transparency is the ability of subjects to effectively gain access to all information related to data used in processes and decisions that affect the subjects."

The third and perhaps less obvious case in the one in which data is used by an individual to make decisions and reach conclusions, such as scientific conclusions. Obtaining detailed information about the origin of the

Digital Object Identifier 10.1109/MSEC.2020.2980593
Date of current version: 14 May 2020

continued on p. 67

 Copublished by the IEEE Computer and Reliability Societies

data as well as changes made to the data allows the individual to better assess data trustworthiness and ethics and enhance data quality. Data transparency is thus extremely relevant also for data users and decision makers.

It is also important to notice that having data management infrastructures that enable data transparency would allow organizations with complex internal structures and processes to improve auditing activities for compliance with regulations and fix data errors as well as optimize data processes and usage. In addition, data transparency would allow organizations to more easily investigate data breaches. In such case, subjects for whom data transparency is critical are individuals that within organizations are responsible for compliance and data security and privacy.

Of course, just providing a definition is not enough. We need to understand the different types of data to be covered by transparency. Examples include:

1. raw data and derived data
2. direct data, that is, data directly provided by an individual to a data collection agent, and indirect data, data about an individual collected by an agent as a result of actions by the individual (for example, the latter is represented by data on web searches carried out by individuals; often individuals are not aware of the digital traces they leave around that are collected by a variety of parties)
3. decision-making data.

A comprehensive discussion about transparency should also identify the many dimensions of transparency, including data collection and record transparency, data use transparency, data disclosure and provisioning transparency, algorithm transparency, and policy transparency. The dimension of algorithm transparency is today particularly crucial given the widespread use of machine learning techniques, such as deep neural networks, used for recommendations, decisions, and other tasks.

Often individuals are not aware of the digital traces they leave around that are collected by a variety of parties.

We need to identify and design techniques and tools to be included in transparency infrastructures, based also on a detailed analysis of which information is relevant to different "transparency stakeholders." An important category of such tools is represented by transparency logs based on append-only authenticated data structures, such as the ones recently proposed for auditing certificate authority and thus achieving certificate transparency.[2] Techniques are also required supporting the tight association of use purposes with data and effectively preventing data misuses. Interesting solutions could be designed based on machine learning techniques. Last but not least, it is important to mention that data transparency may conflict with the common business practices of maintaining trade secrets and making only limited information available to outside parties, particularly regarding algorithms and consumer data collected and stored. It is clear that data transparency cannot be absolute in all circumstances and must be weighed against other competing interests. Ideally, such a compromise will be explicit and achieved based on the input of all involved stakeholders.

I have just mentioned a few ideas toward comprehensive data transparency definitions and challenges. However, the quest for data transparency will be long and challenging as novel technologies, like the Internet of Things, robots, and 5G, are further integrating the cyber and physical worlds and multiply the ability to collect detailed data and the opportunities to use this data. ■

References

1. E. Bertino, S. Merrill, A. Nesen, and C. Utz, "Redefining data transparency: A multidimensional approach," *Computer*, vol. 52, no. 1, pp. 16–26, Jan. 2019. doi: 10.1109/MC.2018.2890190.
2. A. Tomescu, V. Bhupatiraju, D. Papadopoulos, C. Papamanthou, N. Triandopoulos, and S. Devadas, "Transparency logs via append-only authenticated dictionaries," in *Proc. ACM SIGSAC Conf. Computer and Communications Security*, 2019, pp. 1299–1316. doi: 10.1145/3319535.3345652.

Elisa Bertino is a professor with Purdue University. Contact her at bertino@purdue.edu.