EDPS

01 June 2023

EUROPEAN DATA PROTECTION SUPERVISOR

The EU's independent data protection authority

"ENISA Annual Privacy Forum"

Keynote Speech

Wojciech Wiewiórowski,
European Data Protection Supervisor

Given the latest advances in Artificial Intelligence, it would be almost impossible to talk about privacy in 2023 and not consider the effects that Artificial Intelligence is having and will have in data protection and privacy.

I am among those who are attracted and fascinated by the capabilities of the latest versions of artificial intelligence systems. The results obtained by systems such as OpenAI's ChatGPT, GitHub's Copilot or Google's Bard, to name a few, are amazing. I have tinkered with some of them myself to testing and checking their virtues and limitations.

Some latest developments in artificial intelligence such as Low-Rank Adaptation of Large Language Models have democratized AI development by greatly reducing the resources need to train, test and validate Large Language Models.

In principle, limiting technological development has never been fully successful. The Non-Proliferation Treaty (NPT) is a good example of it. The NPT is an international treaty aimed at preventing the spread of nuclear weapons and promoting cooperation in the peaceful uses of nuclear energy. Controlling the means necessary to develop nuclear weapons is far easier than controlling the means to develop artificial intelligence.

Cloud computing services make it possible for anyone to have the computing power to start developing artificial intelligence systems based on open source libraries and publicly available data sets.

A different question is whether we should ban or limit certain uses of artificial intelligence. The European Commission got it right and included a section on prohibited uses of AI when drafting the proposal for an Artificial Intelligence Regulation. Among those prohibitions are the of use artificial intelligence for social scoring or for exploiting any of the vulnerabilities of a specific group of persons due to their age, physical or mental disabilities.

Already in July 2021, the EDPS and the EDPB published a Joint Opinion in which we recommended to add some AI uses to the list of high-risk AI systems and criticized some of the exceptions added to the prohibited uses, notably the ones regarding the use of facial recognition technology in publicly accessible spaces.

In order to thrive, artificial intelligence must overcome a number of challenges which are legal, ethical and societal. It is evident that AI systems must comply with existing legislation. However, given their foreseeable widespread deployment and deep impact they must also behave in line with the ethical standards that societies have agreed upon. AI systems must also be perceived by society as trustworthy and beneficial or they will face a severe pushback.

I am not a Luddite, but I am not either a techno-solutionist. There is no technology whose effects are only positive. The gunpowder we use in our quarries to mine stones is the same technology that is used to destroy lives in bullets, shells and bombs. Artificial intelligence systems used in the context of social media can help us in content moderation, but they can also discriminate content produced by minority groups. If we choose to take stock of the positive and negative effects that artificial intelligence systems can have, how are we going to do it?

We need to have standardized frameworks to conduct risk assessments on AI systems. In this regard, I am well aware of the efforts by the NIST in their Risk Management Framework and the

CHECK AGAINST DELIVERY

recent ISO 23894 guidance on risk management, but there is a gap in between the publication of standards and their widespread application. Few days ago, the G7 leaders of the G7 called for the development and adoption of international technical standards for trustworthy artificial intelligence. We need to close the gap for high-risk AI system and we need to close it quickly.

I would now like to go into some more detail taking as example large language models. As you probably know, these models are AI systems intended to understand and generate human-like language. Their main objective is to be able to use language in a correct way both syntactically and semantically. In contrast with the expert systems developed in the 90s, large language models' design does not pursue the goal of making them experts in any field of knowledge. Its development is not aimed at learning facts or assessing the coherence and reliability of those facts.

Large language models learn facts as part of the learning process, but are not designed with the goal of being a "fact-checker". Therefore, the process of learning facts is a kind of by-product of the learning process through which the system acquires its competences.

This brings us to the challenge of ensuring data quality, personal and non-personal. The learning process of large language models requires a huge amount of training data. One of the biggest challenges currently is to prepare the training data so that the artificial intelligence systems is conscious of the reliability of the data it is trained, tested and validated on.

Unless carefully setup to use specific knowledge sources, large language models respond as if they had knowledge about any topic, but they don't. They learn what is a suitable answer from the "facts" in the information used to train them. In principle, for a large language model, the information coming from Donald Trump's Twitter profile is as relevant and trustworthy as the one coming from the European Central Bank's Twitter profile.

Taking into account the huge sizes of some AI system training datasets, ensuring the accuracy, relevance and completeness of these datasets is going to be a though job.

I can tell you a little anecdote serves to illustrate the extent to which an artificial intelligence system can give a result that makes sense and seems correct but is completely wrong.

A few days ago, I was testing what ChatGPT knew about some of us at the EDPS and started asking who was X and who was Y. According to ChatGPT, one of our deputy heads of unit had been promoted to head of unit. Considering that I am informed of the institution's promotions I found it somewhat suspicious, but then ChatGPT told me that one of our heads of unit was actually the current the European Data Protection Supervisor.

While this mistake may be fun, others may not be so fun and less un-consequential. In a world where "fake news" and disinformation are increasingly common, the use of artificial intelligence systems whose results are unreliable can easily lead to increased political and social polarization.

Artificial intelligence systems and more specifically large language models will have a great impact on our lives as they will be used for countless tasks ranging from filtering and sorting the information we look for on the Internet to estimating the risk when applying for a loan. For this reason, they must be transparent in different ways.

On the one hand, training, testing and validation data sets have a major influence on the development of artificial intelligence systems. The greater the risk that an artificial intelligence system poses to people and their fundamental rights, the greater the transparency with respect to

CHECK AGAINST DELIVERY

the training data should be. The latest trends go right in the opposite direction. Two examples: OpenAI was far more transparent about the training data of GPT-2 or GPT-3 when compared to GPT-4. Google was not much more transparent in this regard when presenting few days ago their PaLM 2 model.

Transparency is not at odds with the legitimate interests of those who develop these systems to maintain their intellectual property rights. Both rights can and should be protected, but it should be clear that, in the event of a clash, the fundamental rights of individuals as they stand will prevail. That does not mean that training data from any AI system has to be public. A company is obliged to inform its customers how their products affect them but does not have to publish full details of the way in which it manufactures them. However, it is accountable to the competent authorities and will have to provide full details to them. In the same way, the designers of artificial intelligence systems would not have to publish all the details about the form or data with them developed by their systems, but they will have to be accountable to the competent authorities to supervise these systems.

On the other hand, systems that, like large language models, can impersonate a human being must be transparent about their nature as an artificial intelligence system. This is one of the requirements set out in the proposed Artificial Intelligence Regulation and I think it is makes a lot of sense.

In this context, I believe that the "interpretability" of the results and the development of explainable artificial intelligence systems (Explainable AI) are going to be extremely relevant. So far, most of the efforts in this field have been directed to explain the autonomous decisions of systems that decided things like the granting of a credit or whether the person who is shown in front of a camera is the one who is authorized to access a device or enclosure. I think that soon many of the efforts in this field of artificial intelligence will be dedicated to being able to explain why a generative artificial intelligence system created or chose a certain response from among the many possible.

At the EDPS we believe that explainable AI will become increasingly relevant and that is why we devoted our IPEN event yesterday precisely to this topic. In the three panels of the event we had participants from academia, public authorities and private companies. For those of you who missed the event but have an interest in the subject, I invite you to watch closely our website because we will publish soon the recording of the event panels.

The challenge of curating the training, validation and testing datasets, jointly with the AI output reliability problem (remember the "promotions" awarded by ChatGPT at the EDPS) gives rise to a related challenge: how to deal with some of the data subjects rights. How are the stakeholders in the AI ecosystem going to ensure effective means to implements rights such as the right to be forgotten, the right of access or the right to objection when considering the development and use of artificial intelligence systems? AI developers and users are accountable and they will have not only to build the tools to make the exercise of these rights possible, but also to be capable of demonstrating that these tools work properly.

Another challenge that artificial intelligence faces is how to deal with bias. We all agree that AI designers, developers and users must fight bias leading to unfair discrimination. However, how are AI developers or users expected to measure biases related to special categories of data without processing big datasets of those types of data?

I am afraid of a situation in which AI designers, developers and users could end up having a legal requirement that could force them to process big datasets of special categories of data such as

CHECK AGAINST DELIVERY

ethnic origin or gender for the only purpose of detecting and mitigating bias. How to deal with bias without processing special categories of data is a very complex question for which I do not have a definitive answer.

In some cases, the use standardized benchmark datasets could help us reducing the processing of special categories of data. In some other cases, explainable AI might allow developers and users to fight bias by explaining which are the features that produced an anomalous output (e.g. why a job applicant who speaks French, German, English and Arabic gets lower rank job offers when compared to another applicant who is equally skilled but does not speak Arabic).

I can already hear the complaints about the GDPR not being prepared for new technologies and artificial intelligence being a good example of its limitations, but I would like to remind you that the GDPR is technology agnostic, or neutral, and its principles should apply no matter the technology at hand. Should we renounce our principles because some technology make their application very difficult or impossible? I do not think so.

You have the right to know which of your data are being processed by a data controller and the answer to that question cannot be: we are sorry but our datasets are too big for us to know. You have the right to request that your data are not used for certain purposes and the answer cannot be: we are sorry but retraining my system so that it does not process your data is too resource consuming.

Data protection by design and by default means that those processing personal data with any means, including emerging technologies, should make themselves this kind of questions. A lack of clear answers means that the technology might not be mature enough to process personal data. The days of moving fast and breaking things must come to an end.

In conclusion, it would be unsound to renounce the benefits that AI can bring to our society. The increased automation and the improvement of the decision-making are only a couple of those benefits.

However, the risks to the rights of individuals are very real and can have a dramatic impact on our democracies.

I believe that by addressing the risks proactively, and with a long-term vision, we can harness the potential of AI while safeguarding privacy rights.

But that means acting now!

CHECK AGAINST DELIVERY