GDPR

This week's reading, "*Defining Personal Data in the European Union and US*", sheds light on how the definition of PII or personal information varies across US and EU. The reading states that while the EU takes a broad approach to defining PII, where the same laws apply to whether the individual can be "identified" or is "indirectly identifiable", the US laws largely relate to instances where an individual can be certainly identified. As a result, this discrepancy in the definitions of PII ultimately affects businesses operating in EU region.

A study[1] shows that Big Tech companies operating in the EU ended up paying fines amounting to more than EUR 1 billion in Feb 2022 as a result of violations such as insufficient legal basis for data processing, or non-compliance with general data processing under the GDPR. For example, recently, Google[2] was issued a fine of EUR 10 million for "hindering the exercise of erasure of personal data" (which comes under EU's "right to be forgotten" law) and for "unlawfully transferring personal data to a third-party" (called The Lumen Project).

It seems obvious then, to come up with a universally applicable definition of personal information. The authors suggest the concept of PII 2.0 which places information on a continuum of identification risk. However, I have doubts on whether this solves the discrepancy in definitions. I believe the ambiguity lies in the type of information being classified as PII, given the context. For instance, though medical records are PII, a situation where these records are being shared with a fellow surgeon for further investigations differs from when they are shared for training a model. This was the criticism raised when the government announced that it would be partnering with Amazon to offer NHS medical advice through Alexa[3]. The fact that Alexa voice recordings are stored and also used for training machine learning models raises questions with regards to data protection policies. On the other hand, if the personal attributes about an individual are removed from a medical record or Alexa recording, it may very well be used for training purposes as there is no longer a risk of being identified. Loosely speaking, this method of using a continuum is merely increasing the number of categories of data classification, instead of solving the real problem.

However, I agree with the Schwartz and Solove's approach of applying differing policy regulations across the continuum, i.e., FIPPs applicable to the information should depend on the risk of identification. Yet I don't think one can clearly specify which or how many FIPPs should be applied to a certain category of data. For instance, in addition to security, transparency, and data quality, I believe that identifiable information should also be subject to data minimalization (i.e., limitation of data collection) in order to avoid the risk of turning this information into "identified" category. To elaborate, if the breadth of data collected is higher, there is a higher risk of identification.

Having said that, to some extent, PII 2.0 succeeds in bridging the gap between the definitions. The question remains: is it possible to place a certain piece of information on the continuum such that there is a universal status quo?

[1] *https://cms.law/en/deu/publication/gdpr-enforcement-tracker-report/numbers-and-figures*
[2] *https://www.dataguidance.com/news/spain-aepd-fines-google-10m-unlawful-transfer-personal*
[3] *https://www.stephens-scown.co.uk/intellectual-property-2/data-protection/amazon-alexas-use-of-nhs-information-raises-data-protection-concerns/*