Governing machine learning models: Challenging the personal data presumption

Abstract

- This paper confronts assertions made by Dr Michael Veale, Dr Reuben Binns, and Professor Lilian Edwards in "Algorithms that remember: Model Inversion Attacks and Data Protection Law"[1], as well as the general trend by the courts to broaden the definition of 'personal data' under Article 4(1) GDPR to include 'everything data-related'.
- Veale *et al* use examples from computer science to suggest some models, subject to certain attacks, reveal personal data. Accordingly, Veale *et al* argue that data subject rights could be exercised against the *model* itself.
- A computer science perspective, as well as case law from the Court of Justice of the European Union, is used to argue that effective machine-learning model governance can be achieved without widening the scope of personal data and that the governance of machine-learning models is better achieved through already existing provisions of data protection and other areas of law.
- Extending the scope of personal data to machine-learning models would render the protections granted to intelligent endeavours within the black box ineffectual.

Keywords: data protection, data subject rights, governance, machine-learning, models, personal data, privacy

1. Introduction

There are growing calls for regulation of models used to make inferences about an individual. For example, in its final report on 'Disinformation and fake news', the United Kingdom's Department of Culture, Media and Sport specifically called for the extension of protections of privacy law "beyond personal information to include models used to make inferences about an individual."[2] Arguably, this reflects the general trend of the CJEU to extend the meaning of 'personal data' to almost everything data-related.[3] Widening the definition might be aimed at the benefit of data subjects in an exponentially-growing, technological-connected world and impose new obligations on controllers and processors in multiple 'smart' environments, but it imposes regulatory burdens on holders of information and may also reduce the effectivity of data subject tools for ensuring fundamental rights are protected.[4] In "Algorithms that remember: Model inversion attacks and data protection law"[5], legal and computer science scholars Veale, Edwards, and Binns outline how rapidly changing technology could theoretically render vulnerable machine-learning *models*[6], thought normally to be protected by obligations of confidence, as personal data.[7] The authors of "Algorithms that remember" offer a

---

[1] Veale, M., Binns, R., & Edwards, L. (2018). *Algorithms that remember: model inversion attacks and data protection law*. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180083.

[2] Department of Culture, Media and Sport, "Disinformation and 'fake news': Final Report", Available at https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/179105.htm#_idTextAn%20chor005 Para. 48, Accessed 02 March 2019.

[3] See Section 3 of Nadezhda Purtova (2018) *The law of everything. Broad concept of personal data and future of EU data protection law*, Law, Innovation and Technology, 10:1, 40-81, DOI: 10.1080/17579961.2018.1452176.

[4] See also the concerns of Bert-Jaap Koops: "the assumption that data protection law should be comprehensive…stretches data protection to the point of breaking and makes it meaningless law in the books" in "*The trouble with European data protection law*", Bert-Jaap Koops in International Data Privacy Law, Volume 4, Issue 4, November 2014, Pages 250–261 at Page 251,

[5] Veale M, Binns R, Edwards L. 2018 *Algorithms that remember: model inversion attacks and data protection law*. Phil. Trans. R. Soc. A 376: 20180083.

[6] Machine learning is defined as "the set of techniques and tools that allow computers to 'think' by creating mathematical algorithms based on accumulated data." See Landau, Deb. Artificial Intelligence and Machine Learning: How Computers Learn. iQ, Available at https://iq.intel.com/artificialintelligence-and-machinelearning/, Accessed 02 March 2019.

[7] Veale et al, citing Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October). *Model inversion attacks that exploit confidence information and basic countermeasures*. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322-1333). ACM.

cautionary tale that models in machine-learning systems can be classified as personal data under European data protection law and that the GDPR could be seen as an important tool in the governance of decision-making models. As a consequence of classifying models as personal data certain data subject rights (e.g. rights of access) and obligations on data controllers (e.g. to provide data subjects meaningful information about logic in decision-making/erasure/rectification) are activated. Certain type of 'attacks' on models "may leak data they were trained with"; as a consequence, "data protection rights and obligations might then apply to models themselves".[8] In the narrowest sense, the authors reflect upon research from computer science scholarship that highlights how machine learning models can be subject to a "range of cybersecurity attacks that causes breaches of confidentiality" and that some of these attacks reveal personal data.[9]

Although the authors do not expressly state as such, we have proceeded on the basis that Veale *et al*'s first claim, under certain circumstances models can be classified as personal data, is normative; and that their second argument: "cybersecurity attacks on machine learning models can reveal personal data; therefore, models are subject to data protection obligations", is descriptive.[10] This amalgamation of both normative and descriptive claims through their deployment of the condition "vulnerable" is deconstructed in this paper. We critique the claims made by Veale *et al* from a computer science perspective to determine whether member inference and model inversion attacks reveal 'any information' 'relating to' an 'identifiable' individual in the manner envisaged by Article 4(1) GDPR[11] and the jurisprudence of the Court of Justice of the European Union (CJEU). As opposed to *databases*, inversion and membership inference models can only ever contain unstructured, anonymous data. While an attack might "leak" data, this does not make the *model* personal.

Veale *et al* mitigate their claims about when models may become personal data by adding a condition of vulnerability to the models discussed. To determine whether an unauthorised, but successful, attack on a vulnerable model would trigger the rights and obligations of a data controller under EU data protection law, we examine the legal consequences of an attack on a model from both a computer science and legal perspective. We conclude that the attacks described are not within the law; accordingly, data protection law is disengaged and the other legal regimes are better placed to better govern models. Furthermore, jurisprudence from the CJEU makes clear that as it is not *legally* possible to reverse engineer the data subject, the legal regime for data subject rights is not applicable.

After providing a synopsis of the Veale *et al* argument, we start by providing a descriptive analysis of how training-models work inside the rather ubiquitous 'black box'.[12] We then review the case law from the jurisprudence of the Court of Justice of the European Union (CJEU) and 'meaningful logic' from computer science literature to the context of attacks on machine learning models. While Veale *et al* recognise training-models have long been regulated (and protected) by intellectual property laws, their approach to extending models the same protection as personal data requires re-identification of data subjects from anonymised data. We argue that this is not the case. We then analyse their descriptive claim, "inverted models" - a term derived from a body of computer science scholarship - are personal data and normative claims - that models are personal data. We examine the case law of the CJEU to determine the consequences of model inversion and membership inference attacks. We find there is no legal basis at present in the GDPR or the case law of the CJEU for regulating interpretations of knowledge and behaviour inferences before deciding how to act upon them.

We conclude that black boxes are places where personal and anonymous data, algorithms and intelligence can mix without becoming subservient to any one particular legal regime. Black boxes should therefore be seen as exchange points that are subject to a plurality of regulatory frameworks. This ensures that no one legal regime dominates over the black box. While data protection rules will still apply to personal data, machine-learning models cannot be classed as such. Extending personal data protection to models would activate non-scalable rights and obligations, potentially disrupting free movement and interactions among stuff inside the black box. Despite this, we argue that data

---

[8] Note 1, *Supra* at Page 4.
[9] Note 1, *Supra* at Page 6, citing Fredrikson and Ristenpart T. (2015).
[10] Normative claims assert that such and such ought to be the case. Descriptive claims assert that such and such is the case. "Normative claims make value judgements whereas descriptive claims don't", Available at https://criticalthinkeracademy.com/courses/moral-arguments/lectures/655333, Accessed 16/12/2018
[11] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.
[12] "Black box" is broadly defined as "anything that has mysterious or unknown internal functions or mechanisms". It has been more narrowly defined as a "usually complicated electronic device whose internal mechanism is usually hidden from or mysterious to the user"

protection law does have a role governing machine-learning models; in particular, ensuring deletion rights when personal data appears in the training data and ensuring compliance with the GDPR's security principle. We conclude that areas of the law outside of the GDPR are better suited to help ensure the protection of data subjects. We achieve this without expanding the scope of personal data to machine learning models.

2. Synopsis of the Veale *et al* argument

For the purpose of this paper, we have summarised Veale *et al*'s argument as follows: Machine-learning systems turn training data into a model that can provide predictions or classifications of new data on the basis of *patterns* distilled from those training data. While such a model does not contain the training data *per se*, a body of work in computer science has shown that reverse engineering techniques can, in certain cases, help to reveal information about the data in the training set. In cases where confidentiality of the training data is expected, in particular when it involves personal data, the application of such reconstruction technique is referred to as an 'attack', in terminology such as *model inversion attacks* and *membership inference attacks*.[13]

The authors suggest that, in those cases where such attacks can be used to infer information about natural persons, any *model* that allows for reverse engineering should be considered as personal data.[14] *Veale et al* envisage that the GDPR could become an important tool for the regulation and governance of machine-learning models. Both Purtova and Veale *et al* suggest that the judgements in *Breyer*[15], *YS and Others*[16], *Google Spain*[17], *Nowak*[18], and a series of working guidance documents on the applicability and scope of 'identifiable', 'relating to' and 'personal data' from Article 29 Working Party (A29WP) show both the regulatory and jurisprudential will to widen the scope of personal data.[19] As the scope of personal data law has been interpreted broadly by the courts[20] and with data protection authorities envisaging its meaning to encompass a variety of new, "factual situations"[21], including those when information enabling identification relating to an individual is not in the hands of one person.

In doing so, they align themselves with a school of data protection scholars that warn that the CJEU is slowly widening the material scope of personal data to include literally any data that "can be plausibly argued to be personal"[22] with the effect of engaging the GDPR's non-scalable regime of rights.[23] However, Veale *et al* add the following caveat:

> *"While we acknowledge this reductio ad absurdum argument concerning the current scope of personal data, and the consequences for it as making the law impracticably broad, our argument does not lean in this direction. We do not aim to support, oppose or resolve the dilemma raised by Purtova; but merely to note that the argument made here—that inverted models might fall under the definition of personal data—does not depend on the kind of expansive definition that gives rise to such absurdities*."

The authors acknowledge the consequences of widening the scope of personal data law and claim to discount the effect of classifying models as personal data. However, they add:

> "*Thus, even if the definition were to be somehow tightened in scope (indeed, the scope of personal data has changed even between recent cases such as YS and others and Breyer), the argument above concerning inverted models would still probably stand. In sum, model*

---

[13] X. Wu, M. Fredrikson, S. Jha, and J. F. Naughton. *A Methodology for Formalizing Model-Inversion Attacks*. Jun 2016, 2016 IEEE Computer Security Foundations Symposium (CSF).
[14] Note 1, *Supra* at Page 6.
[15] Case C-582/14, *Patrick Breyer v. Bundesrepublik Deutschland* [2016] ECLI:EU:C:2016:779.
[16] Joint Cases C-141/12 and C- 372/12 *YS and M. & S. vs Minister of Immigration, Integration &Asylum* [2016], ECLI:EU:C:2014:2081.
[17] Case C-131/12 *Google Spain SL, Google Inc v Agencia Española de Protección de Datos and Mario Costeja González* [2014] ECLI:EU:C:2014:317.
[18] Case C-434/16 *Peter Nowak v Data Protection Commissioner* [2017] ECLI:EU:C:2017:994.
[19] Article 29 Working Party opinion 4/2007 (Hereafter, WP136) on the concept of personal data, 20 June 2007.
[20] See for example, *Nowak*, Note 18, *Supra* at Para 46: "any information' [...] reflects the aim of the EU legislature to assign a wide scope to [the concept of personal data], which is not restricted to information that is sensitive or private, but potentially encompasses all kinds of information, not only objective but also subjective, in the form of opinions and assessments".
[21] *Google Spain*, Opinion of Advocate General Jääskinen, at Para 30.
[22] Purtova, Note 3, *Supra* at Page 2.
[23] Purtova, Note 3, *Supra* at Page 3.

inversion and membership inference attacks, where possible, do risk models being considered as personal data even without resorting to a maximalist reading of data protection law."

This claim is underpinned by the inclusion of the right to object to automated processing and the extent of its scope and functionality of the right.[24] Articles 21 and 22 of the GDPR are heralded as ideal mechanisms for data subjects to gain meaningful information about an algorithmic output that significantly affects them.[25]

Much of the data pumped into a 'machine-learning environment' (MLE) is 'personal data', i.e. linked to a particularly identifiable, living individual.[26] The GDPR defines "personal data" as "any information relating to an identified or an identifiable natural person; an identifiable natural person is one who can be identified, directly or *indirectly*, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".[27] Training data are used to build the models that record the internal 'logic' of the MLE, and if the MLE is designed to process personal data, the training data will also contain personal data. Veale *et al* justify their classification of models as personal data on the basis that "the data returned from model inversion attacks are quite easily construed as personal data, insofar as they resemble a training set or can be used to identify individuals".[28]

We acknowledge that this claim does not say that all machine learning models should be classified as personal data: only those models that are vulnerable to attacks like the model inversion or membership attacks, as Veale *et al*. mention in their paper. In the next section we counter that even those models should be considered as personal data. Models are in fact not like datasets or databases, where information about individuals can be located and – at the request of the data subject – be altered or deleted. Models are amalgamations of identified patterns within training data, that also contained fully anonymous data, and for example weights and parameters for the algorithm etc., into a binary in which the training data or other information is not identifiable – which is also why the reconstruction attacks are done on the basis of observing the behaviour, as a black box. We argue in the next section that the information in the models is anonymous data, from which it follows they should NOT be classified as personal data.

3. Conceptual discussion of models as data

The GDPR endorses a system-based view to automated decision making: the right applies at the point a data controller acquires personal data for the purposes of automated decision making and after the output via automated decision making. This interpretation gives rights to "meaningful information about the logic involved", as well as "the significance and the envisaged consequences of such processing for the data subject."[29] The information should consist of "simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision without necessarily always attempting a complex explanation of the algorithms used or disclosure of the full algorithm."[30]

---

[24] Edwards, L., & Veale, M. (2017). *Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for*. *Duke L. & Tech. Rev.*, *16*, 18.

[25] Roig, A., "*Safeguards for the right not to be subject to a decision based solely on automated processing (Article 22 GDPR)*", in European Journal of Law and Technology, Vol 8, No 3, 2017.

[26] Article 4(1) GDPR.

[27] *Id*.

[28] Note 1, *Supra* at Page 7.

[29] Article 13(2)(f), GDPR.

[30] In Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/67a published in early 2018, the Article 29 Working Party stated: "Article 15(1)(h) says that the controller should provide the data subject with information about the envisaged consequences of the processing, rather than an explanation of a particular decision. Recital 63 clarifies this by stating that every data subject should have the right of access to obtain 'communication' about automatic data processing, including the logic involved, and at least when based on profiling, the consequences of such processing", at Page 27. There is also significant academic debate about the meaning and scope of the right to meaningful information; See Edwards, Lilian, and Michael Veale, Note 24, *Supra;* Wachter, S., Mittelstadt, B., & Floridi, L. (2017). *Why a right to explanation of automated decision-making does not exist in the general data protection regulation*. International Data Privacy Law, 7(2), 76-99. However, Wachter also suggests that the GDPR could be used to provide data subjects with counterfactuals, see Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "*Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR*." Harv. JL & Tech. 31 (2017): 841 and used to provide a right to reasonable inferences, Wachter, S., & Mittelstadt, B. '*A right to reasonable inferences: re-*

In order to assess the normative claim that some machine learning models should be classified as personal data, it is useful to expand on how those models are constructed and how they work. Generally speaking, a machine learning model is a system that answers questions directed at classifying items (i.e. predicting discrete values), or predicting continuous values (such as risks, price development etc.). This is mostly done by machine learning algorithms, where the algorithms reconstruct relationships and dependencies between the characteristics in the training data and the target output. The learning is about constructing a model, that can be applied as a function on new input data to output a corresponding prediction (optimised according to a chosen accuracy metric).[31] The resulting model then contains a "logic" of the dependency of the output on the input for the given task, which it has derived from the training data.

Machine learning models have an important role in modern data-driven decision systems. In traditional decision or prediction systems ("expert systems"), the output would be based on a handcrafted model (Figure 1): a theory that makes a logical interpretation on (presumably causal) relations between attributes based on earlier observations. In this case, the contents of the black box referred to by Veale *et al* are *anonymous, not anonymised, data*. Not because generalization and suppression techniques have been used to anonymise data, but because the information inside the model is nothing more than gobbledygook.

Figure One:

In systems based on machine learning, this ("handcrafted") theory is replaced by the machine-learning model (Figure Two), which is a set of pure correlations without explicit pointers for human interpretation or causality. It is important to realise that the "logic" of the machine learning model, replacing the "handcrafted" model, is of a non-substantive nature: it merely codifies correlations (according to a chosen metric and parameters that turn out to be effective and produce an acceptable margin of error), rather than causal dependencies, i.e. logical dependencies as we may understand them.

Figure Two:

Much of the confusion about so-called automated decision-making can be traced back to misunderstanding what is meant by "algorithmic decisions". In fact, algorithms merely produce a network of numerical interrelations between variables, captured in the machine-learning model, which *we interpret* according to how we expect them to relate to our logic. If people give an explanation of the outcomes of a machine model, this should be recognised as an interpretation, a *rational reconstruction*, of the outcomes of a purely numerical optimization process representing correlations, optimised towards producing an expected outcome (e.g. classification of images, or predicting behaviours). It is also important not to confuse a machine-learning model with a database: training data does not get stored in the model. While the patterns found in the training data and encoded in the model may be characteristic enough to infer membership of certain data (with enough context information), this does not mean that personal data can be located within the model.

To illustrate how the correlations identified by the system heavily depend on the training data, and how interpretations of the working may be misled, Ribiero *et al* trained a model designed to distinguish pictures of huskies from pictures of wolves. While the model looks successful when tested, it could be demonstrated that it in fact had learned to distinguish pictures with snow from pictures without snow - given that the training data that was biased in the sense that most pictures of wolves contained snow, while the pictures of huskies did not.[32] If the animal in the picture is what we intend

---

thinking data protection law in the age of Big data and AI'. Columbia Business Law Review, 2019(1) (forthcoming).

[31] Cathy O'Neil and Rachel Schutt, Doing Data Science. O'Reilly 2014: "In supervised machine learning we have to know beforehand what the intended answer is for the training data (label the data), in order to let the model reconstruct the dependencies based on the features. Unsupervised machine learning on the other hand is the type of machine learning that is useful to detect patterns or rules, in other words: the resulting models are descriptive rather than predictive. In semi-supervised machine learning, unsupervised machine learning may for example be used to cluster training as a labelling to reduce the cost of doing it by hand"

[32] Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin: "*Why Should I Trust You?": Explaining the Predictions of Any Classifier*. KDD 2016: 1135-1144; https://arxiv.org/abs/1602.04938; Note that the bias in the training set was intentionally inserted for the research result presented in the paper: the LIME framework. Using LIME, the

the system to distinguish, we project this onto the classification made by the algorithm and say the system distinguishes huskies from dogs. Yet the system may base its distinction on characteristics of the images that are not completely salient to our logic. It is this projected interpretation that constitutes the decision upon which we then act. Ribiero *et al*'s example informs us that the meaning of "the logic involved" is not referring to a feature of the system, as such, but to the logic we project on it - and that that logic may be incorrect (for example, because the system was trained on biased data).

This is why Article 22 is not applicable to "the logic involved". Often it is not the model itself, but our interpretation of it according to a logic that is NOT in the model, that affords the re-identification or makes membership inference sensitive. In their argument on how models can reveal personal data, Veale *et al* cite the following example:

> "*Faces from facial recognition systems* [can be reconstructed] *to the point where skilled crowdworkers could use the photo to identify an individual from a line-up with 95% accuracy".*[33]

Note that it is *not* the machine learning model that establishes the re-identification. It is the *skilled crowdworker* who may identify a face from a line-up on the basis of a picture that was synthesised through an elaborate exploitation of the model. This crowdworker may in that case infer that this person's picture was in the training data of the facial recognition system. This could reveal sensitive information about the data subject if the intended use of the model is known (for example, the system was trained on a data set of pictures of people with a certain disease), but that information (on intended use) does not leak from the model, but comes from contextual information. Why then should the model be treated as personal data?

We conclude our conceptual analysis of the question whether models can be conceived of as personal data, by looking at the concepts of data, information and the GDPR's definition of personal data. The concepts of data and information are often presented as layers in an epistemic hierarchy: "Wisdom is the ability to increase effectiveness. Intelligence is the ability to increase efficiency. Knowledge is know-how, and is what makes possible the transformation of information into instructions. Information provides answers to who, what, where and when questions. Data are defined as symbols that represent properties of objects, events and their environment. They are the products of observation."[34] Following this conceptualization, data and information are regarded as distinct categories. In the context of Machine Learning, data transforms into information, for example, through labelling and storing into something structured that is not a database; they would rather classify as knowledge (like more traditional statistical models would), in that they make "possible the transformation of information into instructions".[35]

While the very definition of personal data in Article 4(1) of the GDPR as "any information [...]", opens the door for conflating the layers of this hierarchy, it would be an incongruence to put models in the category of personal data, in particular, because the models themselves as data are basically anonymous. In that sense, machine learning models are very different from traditional statistical models: they also potentially allow to derive sensitive information from people, even if their own personal data was not at all involved in the process of creating the model. While statistical models are equations built on (anonymised) data of a population, they do not contain any data, l*et al*one personal data. Such statistical models capture information on the relationship between certain variables (e.g. allowing to predict a person's weight on the basis of their length, in the case of the Body Mass Index). In the case of statistical models, neither technical attacks nor the involvement of a subject's data in the establishment of the equation is needed to infer information about a person. And indeed, statistical models are NOT considered to be personal data - neither does it follow from Veale *et al*'s line of reasoning grounded in the existence of sophisticated attacks that they should be considered to be personal data (or even as data).

---

system can indicate the salient elements on which it bases its classification, and it highlighted patches of snow in the example.

[33] Note 1, *Supra* at Page 6, citing Fredrikson and Ristenpart T. (2015).

[34] Rowley, 2007, p. 166, paraphrasing Ackof 1989, taken from: Baskarade, Sasa; Koronios, Andy. *Data, Information, Knowledge, Wisdom (DIKW): A Semiotic Theoretical and Empirical Exploration of the Hierarchy and its Quality Dimension*. Australasian Journal of Information Systems, [S.l.], v. 18, n. 1, Nov. 2013.

[35] *Ibid.*

4. Attacks/Models

Regulators have taken a binary approach to categorizing data: it is either personal and subject to the GDPR or anonymous and not subject to the EU's data protection regime.[36] This has led to ongoing debates among data protection and computer science scholars about whether the *process* of full anonymization can ever be achieved.[37] In this context, 'anonymised' data means data that does not identify an individual from the data itself or from that data in combination with other data, "taking account of all the means that are reasonably likely to be used to identify them".[38] If data is anonymous or is anonymised, then it is not covered by data protection legislation[39]:

> "*The principles of data protection should not apply to anonymous information, namely information that does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable*".

The scope of the GDPR, alongside Article 11, encourages data controllers to abandon personal identifiers altogether in order to encourage a data flow that is not encumbered by the scale of obligations attached to personal data.[40] Anonymization processes are often criticised as being ineffective in large datasets.[41] A report from the White House even went as far as to claim that data cannot be reliably de-identified:

> "*When data is initially linked to an individual or device, some privacy-protective technology seeks to remove this linkage, or 'de-identify' personally identifiable information—but equally effective techniques exist to pull the pieces back together through 're-identification'.*"[42]

However, Cavoukian and Castro criticised a study on which the White House report was based that claimed to be able to re-identify people on their mobile usage.[43] They argue that the re-identification methods used did not actually re-identify any individuals at all; on the contrary, all they did was show how an individual's mobility data is highly unique.[44] El Enam makes similar claims about an MIT study that claimed to identify 90 percent of the people in the dataset. While researchers were able to show unique patterns of spending, they did not actually identify any individuals.[45]

Veale *et al* suggest a "direct analogy can be made to personal data which have been 'pseudonymised'."[46] This argument is built on two premises: first, that the *model,* opposed to the *personal data* used in the training set, is pseudonymised. Recital 26 makes clear that pseudonymised data remains personal data. Second, a single data controller in possession of both the model and the key required to re-identify personal data should be treated the same in law as a data controller that releases a model wherein the key is held by a *different entity*. The authors correctly recognise that

---

[36] See Article 29 Working Party's Opinion 05/2014 on Anonymization Techniques.
[37] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010).
[38] ICO, *Big data, artificial intelligence, machine learning and data protection*, 20170904 at Page 58.
[39] Recital 26, GDPR.
[40] Article 11(1), GDPR and the caveat found in Article 14(5)(b).
[41] Jain, P., Gyanchandani, M., & Khare, N. (2016). *Big data privacy: a technological perspective and review*. Journal of Big Data, 3(1), 25; See also president's council of Advisors on Science and Technology. *Big Data and Privacy: A technological perspective*. White House, May 2014. Available at http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf Accessed 03/03/2018.
[42] Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values*, May 2014,Available at http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf, Accessed 19 July 2019.
[43] Cavoukian, Anne and Castro, Daniel. *Big data and innovation, setting the record straight: de-identification does work*, Office of the Information and Privacy Commissioner, Ontario, June 2014. https://www.ipc.on.ca/English/Resources/Discussion- Papers/Discussion-Papers- Summary/?id=1413 Accessed 1 June 2018.
[44] See "The Risk of Re-identification Has Been Greatly Exaggerated" in Cavoukian, Ann, and Daniel Castro. "Big data and innovation, setting the record straight: De-identification does work." White Paper, Jun (2014): 20.
[45] El Emam, Khaled' *Is it safe to anonymise data?* BMJ, February 2015, Available at http://blogs.bmj.com/bmj/2015/02/06/khaled-el-emam-is-it-safe-to-anonymise-data/, Accessed 03 March 2019.
[46] Article 4(5) GDPR defines pseudonymization as "*the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.*"

Recital 26 imposes a test of reasonable likelihood of re-identification, but rather disappointingly assume that model inversion and membership inference attacks would automatically render the *double-objective test* found in Recital 26 irrelevant as long as the key to re-identifying the data is held by another entity.

The discussion on reconstructing personal data from the training data through model inversion and/or membership inference bears resemblance to questions about the two types of non-personal data: anonymous and anonymised data. By analogy to de-anonymization attacks on anonymised datasets, the model is not directly revealing personal data, as such. It requires an intentional process of reconstructing the functionality of the model in order to gain information from the training data that can be related to a natural person. However, treating the model like pseudonymised data, viz. *as* personal data, is too quick a conclusion: the model *does not constitute* the crucial information referring to a natural person. Rather, the model, containing codified correlations of numeric parameters of training data, is the tool in the process. The crucial information referring to the natural person, or even the judgment whether it actually refers to the natural person is dependent on our interpretation, a rational reconstruction by a human that depends on a lot of contextual information e.g. about the purpose of the model - which is important to note, is not encoded, as such, in the model.[47] Reclassifying an entire model as personal data after an illegal attack that hypothetically results in an inference by a crowd worker with a specific skillset goes beyond any CJEU ruling on the extent of "personal data".

5. "Model inversion" & "membership inference" attacks

When it comes to making knowledge claims on the basis of machine learning outcomes, there are several caveats. As demonstrated by the Wolf-Huskie example above, the very selection of the training data may highly influence the resulting model. What is relevant for the attacks described by Veale et al is that models behave measurably differently if they operate on data that was present in the dataset they were trained on, compared to when they operate on 'fresh' data. Model inference attacks are about revealing a specific inference: using certain data about a person as a baseline, was this person's personal data in the original training dataset? Determining where someone's personal data was in the training data could, by contextual information on what the model is trained for, reveal something of significance about that person (e.g. if the model is aimed at predicting treatment success for a certain disease). Does the possibility of such type of "attack" make a machine-learning model itself personal data, and therefore fall under the regulatory remit of the GDPR?

Membership inference "attacks" are not necessarily attacking in a "breaching encryption" sense, but in the fact that information which was assumed to remain confidential, can be revealed if someone puts in the effort of reverse engineering. The use of "attack" in "model inversion attack" or "membership inference attacks" does not refer to breaking some intentional protection (e.g. for protecting the trade secret). It refers to a technique using the (black box) model (through an API, so still protecting trade secrets) to generate a series of shadow models that mimic the original model.[48] The result is a so-called "attack model" that behaves similar enough (as validated in the cited works) to the original model - and then exploit the fact that trained models generally show different behaviour when fed data from the training data set versus data "they see for the first time".

However, there are malicious scenarios thinkable, especially in ML-as-a-service (MLAAS), where machine-learning service providers make their algorithms available in a way that training data get stored more explicitly than necessary in the model, and can also be retrieved from it.[49] The ML-AAS provider becomes a data controller of the training data their client provides them with while using their service. Not because the model is personal data, but because of the way they process the personal data in the service they provide. This triggers the need to provide reasons for this choice and inform

---

[47] Ribiero et al at Note 32, *Surpa* demonstrate that we can build tools and frameworks that provide information for reconstructing a logic and validating it vis à vis the models output - and how those reconstructions can also point us to errors in our initial interpretations: in the wolf-husky example the system showed to actually use the environment in the picture as salient for the classification instead of the canine - as most people (with the context information that this system was built to distinguish wolves from huskies) would have interpreted it.

[48] For example, the works cited in Veale et al: Reza Shokri et al. "*Membership Inference Attacks Against Machine Learning Models*". In: 2017 IEEE Symposium on Security and Privacy (SP). IEEE, May 2017, 3–18 and Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. "*Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*". In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM, 2015, pp. 1322–1333.

[49] See the work cited in Veale et al: Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. "*Machine Learning Models that Remember Too Much*". In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17).New York: ACM, 2017, 587–601.

the users of the platform (as data controllers themselves). In the facial recognition systems example discussed above, the blurred image only related to a specific person whose identity is not apparent from the data and the data is not directly linked with data that identifies the person.[50] Re-identification can only occur when matched to additional identifying data. There is no known, systematic way for the data controller to reliably create or re-create a link with identifying data. The outcome may be subject to potential re-identification attacks that could create a possibility of *inference* in *some* records to an identifiable individual with *some* degree of confidence. But this relies on a model being so rich that it "leaks" personal data.

When a model does in fact "leak" personal data, Veale *et al* suggest that data subjects exercise their Article 17 GDPR erasure right to achieve one of two outcomes. They suggest to retrain the model on an amended data set, no longer including the personal data of the data subject. The retrained model no longer allows this particular data subject to be re-identified as the model's functionality no longer contains its traces. Alternatively, they suggest the model should be amended to remove the traces of the training data in the functionality of the model, blocking the potential for reconstruction of the model. We agree this is a theoretical solution, but at present, not a practical one, given relevant techniques for that are not yet operational. Notably, both suggested measures do not involve deletion of the model, but avoiding constructible traces of training data within the logic of the model. It seems Veale *et al* treat models as personal data in the way databases can be personal data for multiple data subjects. Yet, one data subject cannot ask for deletion of an entire database to avoid infringement of the rights of the other data subjects. Deleting entire models should *not* be a right, but not because models should be treated analogously to databases *as* personal data. Models need not to be classified as personal data in order to require countermeasures to provide protection from the unlawful reconstruction of what is within both the model or the training data set.[51] This would be a far more appropriate outcome than giving data subjects the right to demand that a model is retrained on amended training data.

6. The Governance of ML-Models

6.1 Data Protection and the "Legal Means" test

According to the case law of the CJEU, the GDPR contains four elements to determining whether data is personal: (1) any "information" must (2) relate to (3) an identified or identifiable (4) natural person.[52] As the following paragraphs outline, this is particularly apt for any discussion about whether data subjects can exercise their rights against model controllers. For the most part, our own critique maps neatly onto Purtova's analysis[53], with some exceptions: First, we concur with her conclusion that 'any information' should be interpreted broadly to include information, regardless of its nature or content. However, it is important to note that the *Nowak*[54] court did not explicitly endorse the Article 29WG's advice note on the meaning of "all information" and has yet to provide clear guidance on what the term "information" really means. The consequence of leaving this unanswered is instrumental to the present debate among data protection scholars. At one end of the spectrum of data protection advocacy, some argue that "all information" must include "all data"; at the other, "any information" can only include information that **could** contain information that could be personal data, provided that the other requirements of Article 4(1) GDPR are satisfied.[55]

The court in *Breyer*[56] gives broad meaning to the term, "identifiability", but that scope is limited in situations when identifying can only come about from merging additional data held by data controllers or 3rd parties by *legal means*.[57] To determine whether non-personal data could become personal data when combined with data in the hands of another, the *Breyer* court had to determine whether the *legal means* existed between an Internet services provider and an online media services provider for an exchange of information that would render non personal data identifiable.[58] The court determined that the latter could take steps necessary to obtain information from the internet service provider in order to bring criminal proceedings against those accused of cyber-attacks, piracy, etc.[59] Importantly,

---

[50] *Id.*
[51] *Id.*
[52] Article 4(1), GDPR.
[53] Purtova at Note 3, *Supra*
[54] Case C-434/16 Peter Nowak v Data Protection Commissioner [2017] ECLI:EU:C:2017:994.
[55] Note 4, Supra (Purtova) at Page 23.
[56] Case C-582/14, Patrick Breyer v. Bundesrepublik Deutschland [2016] ECLI:EU:C:2016:779.
[57] Breyer at Para 49.
[58] Breyer at Para 47.
[59] *ibid.*

it was not enough to classify dynamic IP addresses as personal data after identifying a third party, irrespective of who it may be, capable of using those dynamic IP addresses to identify network users.[60] The re-identification argument deployed by Veale *et al* can, in fact, only apply to situations where the combination of the various elements of information constitutes a "legal means likely reasonably to be used" to identify the data subject. The *Breyer* decision makes clear that the theoretical possibility of recombining relevant pieces of information to enable the identification of relevant individuals is insufficient. If identification would be *illegal* or practically impossible on account of the fact that it would require a disproportionate effort in terms of time and effort then individual non-identifying pieces of information would not constitute personal data:

> *"….a dynamic IP address registered by an online media services provider when a person accesses a website that the provider makes accessible to the public constitutes personal data within the meaning of that provision, in relation to that provider, where the latter has the* ***legal means*** *which enable it to identify the data subject with additional data which the internet service provider has about that person*"[61] **[Emphasis added]**

The *Breyer* decision refers to a triangular relationship between a data subject, an IP address that cannot be tied to his/her name, and an Internet service provider that can identify the name of the person behind the IP address. The court found that an IP address could constitute personal data for the website publisher if the publisher has the *legal means* to obtain additional information that enables the publisher to identify the visitor. Because of the legal relationships between website publisher and the internet service provider, re-identifiability was deemed possible. The internet service provider has additional information that could be combined with the IP address to identify the website visitor.[62]. However, model inversion and membership inference attacks are not exactly the *legal means* envisaged by the *Breyer* court: *"if the identification of the data subject was* ***prohibited by law*** *or practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant.*"[63] **[Emphasis added]**

This suggests that 'the means likely reasonable to be used' referred to in Recital 26 requires a double objective test to determine whether a data subject is identifiable:

> *"account should be taken of all the means* ***reasonably likely to be used****, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person,* ***account should be taken of all objective factors,*** *such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments."* **[Emphasis added]**

But this does not include identification *prohibited by law*. The difference here is that attacks only reveal something that could infer personal data. The model owner may not have the legal means to obtain, from another party, extra information that enables the model owner to identify the person. The examples provided by Veale *et al* are limited to situations a) when the model and the key are held by the same data controller and b) when the key and the model are held by different entities. What is not considered by Veale *et al* is where the model is attacked to reveal information that allows re-identification by someone in a way that no key is required.

It would also be an absurdity for someone that has anonymised data to escape the obligations of data protection law or to leave data subjects without any protections after an illegal breach is used to re-identify personal data. It is, however, also an absurdity to impose the GDPR's obligations on someone working with anonymous data that falls victim of an illegal attack. The victim would be burdened with obligations arising from processing, despite never handling personal data and third-party inferences about the identity of data subjects. Accordingly, the appropriate action would be against the hacker, whose illegal actions have made them a data controller without a legal basis for the processing of personal data. This would appear to contradict the GDPR's framework for damages for unlawful

---

[60] Judgment in Case C-582/14 Patrick Breyer v Bundesrepublik Deutschland; See also Opinion of Advocate General Sanchez-Bordona General in Breyer at Para 50.
[61] Breyer at Para 49.
[62] Zuiderveen Borgesius, Frederik, Breyer Case of the Court of Justice of the European Union: IP Addresses and the Personal Data Definition (Case Note) (June 6, 2017). European Data Protection Law Review 2017, Volume 3, Issue 1 . Available at SSRN: https://ssrn.com/abstract=2933781.
[63] Para 46. The CJEU refers to para 46 of the AG's opinion.

processing. Article 82(2) states: “Any controller involved in processing shall be liable for the damage caused by processing which infringes this Regulation.” But Article 82(3) clearly limits controller liability in situations like those described by Veale et al: “A controller or processor shall be exempt from liability under Paragraph 2 if it proves that it is not in any way responsible **for the event** giving rise to the damage”. Theoretically, data subjects could exercise their rights against a controller if they have not taken reasonable means to secure their model, but responsibility is a two-way street. When determining whether a controller has lived up to their obligations, it is only right that regulators use a risk-based assessment rather than imposing strict liability. Presuming that the data controller had lived up to their Article 5(1)(f) and Article 5(2) obligations, the effect of Article 82 is clear: data subjects would not be able to seek damages as a result of an *illegal* attack.

Where does this leave data subjects then? It should be axiomatic that a data subject wishing to exercise their rights should be able to force a data controller to identify personal data relating to them and delete upon request. Herein lies another problem with the Veale et al argument. The training data that enters a machine-learning environment might contain personal data (ensuring a role for data protection law); however, once the MLE starts updating with new data, the training data no longer relates to the data subject. There would be no identifiable data within a machine-learning model that relates to an identifiable living individual. The data subject would only be able to realise the erasure of identifiable information against the controller of the training data and against the skilled crowdworker.

6.2 Criminal law as deterrence

One way of holding both attackers and model owners accountable can be seen in the criminal law measures found in the United Kingdom’s Data Protection Act 2018. The attacks Veale *et al* describe are criminal offences under Section 170 and S171. Section 170 makes it an offence for a person to knowingly or recklessly a) obtain or disclose PD without the consent of the controller b) to procure disclosure without the consent c) after obtaining PD to retain it without the consent of the controller. Even if a handcrafted model is made up of pseudonymised data, then Section 171 states it is an offence to knowingly or recklessly re-identify information that is de-identified personal data. The bottom line is that model inversion and membership inference attacks are illegal. It is a criminal offence to hack a model for the purposes of obtaining personal data or to re-identify pseudonymised data.

When developers started using integrated functionality from third-party providers, API usage quickly became the *de facto* standard for developing “interactive digital experiences users have gotten used to and are fundamental to a business’ digital transformation”.[64] Undoubtedly models are vulnerable to hacking and other forms of cybersecurity attacks. Models exist inside systems, and systems are protected through various legal mechanisms against both unauthorised access and unauthorised purposes after authorised access. The Cybercrime Convention defines "computer system" as “any device or a group of interconnected or related devices, one or more of which, pursuant to a program, performs automatic processing of data”.[65]

The Convention also requires Member States to establish in their domestic law a criminal offence for “intentionally accessing [a computer system] without right”.[66] The stated objective of the Cybercrime Convention is “to deter action directed against the **confidentiality, integrity and availability** of computer systems, networks and computer **data** as well as the misuse of such systems, networks and data by providing for the criminalisation of such conduct” **[emphasis added]**.[67]

Observing leaked data in a data stream without authorisation would amount to authorised access for ‘unauthorised purpose’ and could be offences under Sections 1, 2, and 3 of the UK’s Computer Misuse Act 1990.[68] In *R v Bow Street Magistrates, ex parte Allison*[69], Lord Hobhouse made clear that

---

[64] *The problem of API abuse*, Approov, 18 Oct 2016, Available at https://www.securityweek.com/nextbig-cyber-attack-vector-apis, Accessed 26 Jan 2019.
[65] Article 1 definitions, Convention on Cybercrime (Details of Treaty No.185), Available at https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/185, Accessed 19 July 2019.
[66] Article II, Illegal Access.
[67] Budapest, 23/11/2001 - Treaty open for signature by the member States and the non-member States which have participated in its elaboration and for accession by other non-member States.
[68] Section 2 is sometimes referred to as the ‘ulterior intent’ offence. This involves unauthorised access to data with the intent of using access to commit another crime.
[69] [1999] 4 All ER 1.

someone with authorised access to one area of a computer system is not immune from prosecution for accessing other pieces of data of the same kind.[70]

Just as using phishing techniques or a Trojan horse to obtain identity data or to acquire any other data from an unauthorised source, model attacks could also amount to unauthorised modification of computer materials.[71] Even if a system is designed to be used in a certain way, this does not mean abusing that system to reveal information about a data subject is not an offence. Authorization to view data does not extend to authority to copy or alter that data.[72] Furthermore, the Section 171 offence where an individual "knowingly or recklessly re-identifies information that is de-identified personal data without the consent of the controller responsible for de-identifying the personal data" specifically acts as a deterrent to these types of hacks – and criminalizes every subsequent actor in the chain. Section 171(5) makes it an offence for a person knowingly or recklessly to process personal data that is information that has been re-identified where the person does so without the consent of the controller responsible for de-identifying the personal data, and in circumstances in which the re-identification was an offence under Section 170(1). Collectively, these provisions will be important tools to ensure the responsible control of models. Beyond criminal and data protection law, the market also provides the means for ensuring models that have value through their commercial sensitivity are properly governed.

6.3 Market as modality for model governance

Ambiguous terms like "ambient computing" and "big data" create a lack of clarity for businesses wishing to extrapolate economic value from the data and marketable uses of those insights. Unsurprisingly, business leaders sought to develop the "volume, variety, and velocity" of big data into monetization opportunities[73], with API use and data analytics becoming two forms of intellectual capital. More companies began contemplating the challenges associated with accounting for models as "corporate assets".[74] While some see an economic benefit in the 'free flow' of data, some sceptics claim benefits are only actually derived from mined insights.[75] Others argue the machine-learning algorithm inside the black box is what is actually valuable; hence, one of the reasons markets for trading machine-learning models have developed.[76] Models exhibit unusual characteristics when compared to other balance sheet assets. This is because most assets depreciate over time; however, machine-learning models should appreciate or gain in value with usage; that is, the more the model learns, the more complete and more valuable the model becomes to the asset owner. These characteristics also apply to analytics, where analytics is basically "data" that has been "curated" into customer, product or operational insights. As models learn, they can become both valuable and proprietary in nature, whereby keeping the model's secrecy is of the utmost importance to the business's survival.

The EU regime for protecting commercially sensitive secrets has been overlooked for its ability to provide additional protection for data subjects *beyond* the GDPR.[77] Trade Secrets protect value in knowledge for a business in competition, without bestowing a property right on that knowledge. The business value in machine-learning comes from the combination of data sets, computing power, the choice of a certain combination of algorithms (with chosen settings for parameters), together generating the models that are of service to customers for certain tasks. Businesses are going to want to protect those elements to generate the models.

On the other hand, the GDPR focuses predominantly on ensuring data controllers deploy appropriate "technical measures" to secure data with the latest technology[78] or use pseudonymisation or encryption[79] and take reasonable "organisational measures" that focus on how personal data is

---

[70] Ibid, at 7. See also Section 5 of the UK's Computer Misuse Act 1990. This can be applied to data as well as computer systems.
[71] Section 3 Offence, see the interpretation found in Section 17(5)(b) Computer Misuse Act 1990.
[72] [2000] 2 AC 216 at 223-4.
[73] Mayer-Schönberger & Cukier, *Big Data: A Revolution That Will Transform How We Live, Work and Think* (2013).
[74] Dimitrovska, I., & Malinovski, T. (2017). *Creating a Business Value while Transforming Data Assets using Machine Learning*. Computer Engineering and Applications Journal, 6(2), 59-70.
[75] Custers, B.H.M. (2018) *Data Mining and Profiling in Big Data*, in B.A. Arrigo (ed.) The SAGE Encyclopaedia of Surveillance, Security, and Privacy, p. 277-279, Thousand Oaks: SAGE Publications, Inc., Available at https://ssrn.com/abstract=3183286, (visited 15 March 2019).
[76] Veale *et al* citing Zhou ZH. (2016*) Learnware: on the future of machine learning*. Front. Comput. Sci. 10, 589–590.
[77] M.R Leiser and F. Dechesne "*Trade Secrets, Models & Personal Data: Resolving the Conflict*", (forthcoming).
[78] Article 32 (1), GDPR.
[79] Article 32 (1)(a), GDPR.

processed. For example, the GDPR points to the implementation of risk assessments,[80] processes[81], code of conducts and certification mechanisms[82]. The latter is assessed by supervisory authorities and not by the company itself[83]. This begs the question whether the "reasonable protective measures" that an organization takes to protect their trade secrets under Article 2(1)(c) of the Trade Secrets Directive equate to the requirement that a data controller take "technical and organizational measures" under Article 25 GDPR?

As a rule of thumb, it seems realistic to assume that the more controlled the access to and/or use of the secret information has been within a business, the more likely it will be that the information is protected. The GDPR recommends the pseudonymisation of personal data, and *transparency* with regard to the *functions* and processing of personal data. On the other hand, the trade secret holder should take contractual, physical and organizational measures to ***prevent*** *unauthorised access*, use, disclosure, loss and modification of trade secrets. The TSD suggests protocols that include everything from restricting access (need-to-know) to physical barriers and other technical safeguards (including passwords, firewalls, automated intrusion detection systems and authentication measures), none of which are suggested in the GDPR or the Working Guidance on Technical Measures.[84] The reasonable measures requirement does not take into consideration the fluidity required to ensure data flows smoothly.[85] Recital 76 of the GDPR requires an objective test to determine whether an organisation has taken appropriate technical and organisational methods for ensuring the rights and freedoms of data subjects: (1) the likelihood, and (2) the severity of the risk should be determined by reference to the nature, scope, context and purposes of the processing: "The likelihood and severity of the risk to the rights and freedoms of the data subject should be determined by reference to the nature, scope, context and purposes of the processing. Risk should be evaluated on the basis of an objective assessment, by which it is established whether data processing operations involve a risk or a high risk". Article 25 also states: "Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures…".

Taken together a cyber-security attack on high-risk processing would not fall foul of the Article 25 requirement if an objective assessment of the *likelihood* of an attack was low. On the other hand, in addition to technical and organisational measures, the owner of a machine-learning model protected by trade secret law will required to prove they took other practical steps, outside the scope of technical and organisational measures, including (reviewing) employment and contractual provisions (nondisclosure and confidentiality agreements, effective dispute resolution clauses), the use of non-compete agreements, a critical review of the company's IT infrastructure and HR and supplier policies with a focus on appropriate confidentiality clauses. Therefore, if the legally responsible owner of a commercially-sensitive model has met the requirements of the Trade Secrets Directive, they will likely exceed their obligations in the GDPR to implement "technical and organisational measures to protect personal data" as far as they are contained or can be reconstructed from the model. This is particularly apropos in processing, whereby the security protocols are only *proportionate* to the level of risk. Commercially sensitive models that meet the threshold of protection as a trade secret will have also satisfied the legal requirements for security under the GDPR and should be seen as a mechanism for the protection and governance of machine-learning models.

7. Conclusion

The GDPR has been said to be a powerful tool "that intends to strengthen and unify data protection for all individuals within the European Union (EU)" and "the most important change in data privacy regulation in 20 years."[86] Others have claimed that the GDPR provides data subjects with significant new rights and provides data protection authorities with new powers to prevent the harms often associated with the Facebook and Cambridge Analytica scandal. Yet classifying models as personal data blurs the lines between personal and anonymised data in a way the GDPR drafters never intended and exacerbates Purtova's primary concern: expanding the scope of personal data at scale would create

[80] Article 35, Recital 83, GDPR.
[81] Article 32(1)(d), GDPR.
[82] Article 32 (3), GDPR.
[83] Article 41, GDPR.
[84] Recital 78, GDPR.
[85] Bone, Robert G. "*Trade Secrecy, Innovation, and the Requirement of Reasonable Secrecy Precautions*." (2010): 09-40.
[86] https://eugdpr.org/

a system so vast and so large, it would be impossible for data subjects to enforce their rights or for data protection authorities to do any meaningful supervision or enforcement. The distinction between personal and anonymous data underpins responsible innovation and development, but also facilitates trade and creates new markets and opportunities. Model trading is a natural consequence of the GDPR's permissive approach to barrier-free markets, the volume of information associated with big data mining, and enterprise developing insights to swaths of data for valuable public, social, or commercial benefits. In some instances, the model has been trained to such extent that it provides valuable performance prediction (e.g. stock market, marketing and advertising, and even architecture[87]) to the extent that it gains protection as commercially sensitive.

Due to concerns that ML systems are fast becoming part of our critical societal infrastructure[88], much of the "evangelical compulsion"[89] on algorithmic decision-making has been dedicated to finding fault in automated outputs. Concerns about unfairness and discrimination have led to suggestions that widening the scope of personal data to include all things information-related is the only way to reconcile fundamental rights with the increasingly ambient, smart, and connected society in which we live. For example, Edwards and Veale have previously argued that one way to address these issues is to widen the scope of the GDPR to include ML-Systems[90] and Purtova welcomed "the broad interpretation of the concept" to justify widening the scope of personal data to include any data that has a potential to "impact people".[91] These claims can be questioned from two perspectives:

1. **Effectiveness:** Why should the solution for protecting against impacts be sought in widening the scope of personal data, when the resulting scale of the GDPR makes rights and obligations unmanageable? There are other legislative frameworks that apply more directly and more effectively to protect the fundamental rights at stake. Data protection is not designed to be the penultimate fundamental right that can be deployed to protect all others.

2. **Conceptually:** The GDPR already contains obligations for data controllers to assess impacts of data-processing decision systems and to implement protective technical and organisational measures. Does classifying parts of these systems as *personal* data add any further protections? Regulating the effects of improper use when it impacts the data subject is a better strategy. Veale *et al* propose models should fall under the definition of personal data on the basis of it technically being possible to reconstruct information on persons, using the normal functionality of the model, if their data were used in the training dataset. However, it is very important to be clear on the fact that the model itself does not *contain* personal data *per se,* nor reveal it. It requires purposeful action of a nefarious actor seeking to reveal the personal data, e.g. the reconstruction of the black box model into a shadow model for the purpose of revealing information from the training data. This should count as improper use of the model. We do not have to look for a solution in terms of strict data protection if we already have a body of work regulating bad behaviour.

Independent of the previous point, we raise the question to which extent the outputs of the shadow model are actually directly relating to an identified or identifiable person. Veale *et al* cite model reconstruction for a facial recognition system that delivered blurred versions of pictures in the training data of the original model - the actual re-identification is not done by the model but by a skilled human. The data in itself only become personal data on the basis of external inferences. And while data can certainly impact people, not all data that impacts people is personal data.

We hope this is not the start of a troubling trend in digital law whereby academics, specializing in privacy and data protection, use 'proof of concepts' to advocate increasing the material scope of 'personal data', to the extent that the rest of information technology jurisprudence relevant to the very 'concept' is overlooked; in this case, ignoring laws that make unauthorised access in computer systems illegal or existing offences within data protection law. Data protection law is not needed to protect victims from unlawful attacks, but already provides sanctions if need be. The 'right' outcome comes from other, existing legal frameworks.

By stretching the scope of the concept of data, and with it of data-protection mechanisms, we run the risk of reducing their ability to protect against impacts of data processing systems effectively. By using

---

[87] Elizabeth Stinson, Wired, 01/12/2017, "*What happens when algorithms design a concert hall? The stunning Elbphilharmonie*", Available at https://t.co/RVohGotkIc, Accessed 25 Jan 2018. 120

[88] *Slave to the Algorithm*, Note 24*, Supra* at Page 26

[89] We have purposely borrowed this phrase from Edwards and Veale's earlier work, *Slave to the algorithm,* Note 24, *Supra* at Page 26.

[90] *Slave to the Algorithm* at Note 24, *Supra*

[91] Note 3, *Supra.*

non-data protection mechanisms, we can provide remedies for when “data impacts people” while leaving the scale and objectives of the GDPR as the drafters intended. It would be wise to remember that if all information was meant to be ‘personal data’, then there would be no need for ‘data’ to have the ‘personal’ qualifier.

An attack that uses data in the hands of a 3rd party to re-identify data subjects does not make a legally responsible owner of a model a data controller. Furthermore, even if no technical measures are in place to secure the model, if there was no commercially-sensitive protection available for the model, and the hacker was able to re-identify the data subjects as existing in the training data, the model itself would still not amount to personal data. Online machine learning systems update their decision rules after every query, with the effect that any disclosure will be obsolete as soon as it is made. The very nature of machine learning systems renders any query to algorithmic processes pointless; each query updates the existing decision rule, becomes outdated and rendering disclosed information purposeless.

The effect of Veale *et al*’s argument would result in the protection of models as personal data as soon as created and before they have even been used. This is the German concept of self-determination or *selbstbestimmung*. This approach certainly might explain why there is not any mention of identified harms in *not* labelling models as personal data. However, without any identified harms, would it be logical to regulate our own interpretations of a model’s outputs? Are Veale *et al* arguing that inputting personal data into a model is an infringement of the right to informational self-determination? If so, this right is not absolute. One cannot control all aspects of one’s ‘self’; but, especially not in a model, where one’s self is always in relation to others. The "model" is just an interpretation of the self; data protection does not prohibit others to interpret the ‘self’ that you determine. For all the reasons above, we find the authors’ argument to be the *reduction ad absurdum* they claim to avoid.