Personal data's ever-expanding scope in smart environments and possible path(s) for regulating emerging digital technologies

Raphaël Gellert*

Key Points

- The goal of this contribution is to present some reflections concerning Purtova's 2018 paper on 'the law of everything', by providing a multidisciplinary analysis combining insights from data protection law scholarship and from technical literature on machine learning.
- Purtova convincingly sketches a 'maximalist' definition of the notion of personal data. It relies upon a double scenario, where (i) everything becomes information, and (ii) all this information qualifies as personal data. Hence data protection law as 'the law of everything'.
- She relies upon two sets of assumptions. The first one is legal and based on the EU official interpretation of the notion of personal data. The second one is of a more hypothetical nature insofar as it relies upon a number of assumptions concerning the development of certain technologies and their affordances.
- This contribution therefore describes the way in which artificial intelligence is deployed in smart environments, with a view of discussing Purtova's paper. The paper puts three main points forward. Overall, Purtova's analysis is very sound and cogent. However, at times she seems to rely upon too linear accounts of technology.
- This is the case as far as the current and future capabilities of data-driven technologies are concerned (first point).
- This is also true as far as her view of anonymization is concerned, which she seems to conceive as a purely technological phenomenon, whereas technical literature has on the contrary put forth a more nuanced understanding of anonymization which is grounded in its socio-technical nature (second point).
- Finally, even though she rightly points to the challenge of 'system overload' as a result of the expanding scope of the notion, the issue of knowledge inference—at the heart of data-driven technology—is absent from her analysis. Yet, the latter also presents key challenges for data protection law, and should be at the core of any regulatory framework concerning data-driven technologies (third point).

Introduction: personal data: a discussed notion, and 'the law of everything'

The EU debate—both at policy and academic level—around the notion of personal data has been a steadily ongoing one.

Beyond the differences with the narrower US-based notion of PII (personally identifiable information),[1] discussions have taken place at various levels.

At the policy level, one can think of the shift from the narrower version of biographical information that existed in a number of national statutes to the broader, current notion of personal data as enshrined in the EU

* Raphaël Gellert, Radboud University, Nijmegen, The Netherlands, Email: r.gellert@jur.ru.nl
This contribution reports on the results of the project 'Understanding information for legal protection of people against information-induced harms' ('INFO-LEG'). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 716971). The article reflects only the author's view and the ERC is not responsible for any use that may be made of the information it contains. The funding source had no involvement in study design, in the collection, analysis and interpretation of data, in the writing of the report, or in the decision to submit the article for publication.
I would like to thank Prof. Paul De Hert for earlier comments on this paper and Mara Paun for discussing the over- and under-inclusivity of rules, as well as the anonymous reviewers. All remaining mistakes are the sole responsibility of the author.

1 See, eg Paul M Schwartz and Daniel J Solove, 'The PII Problem: Privacy and a New Concept of Personally Identifiable Information' (2011) 86 New York University Law Review 1814.

data protection Directive,[2] and confirmed in the General Data Protection Regulation (GDPR).[3]

At the academic level, the discussions have been manifolded. The technical literature has often focused upon the possibility of linking a dataset to an individual so that the latter be identifiable as per the EU definition of personal data, and has therefore been characterized by technical prowess attempting to re-identify supposedly anonymized datasets.[4]

The legal literature has often focused on the impact of modern data processing operations on the definition of personal data. This has mostly been the case with profiling, a debate that emerged more than a decade ago already. Some academic contributions have argued that profiles do not always amount to personal data.[5] Other, non-binding policy instruments, have argued in the opposite direction.[6] To the knowledge of the present author, this matter has not been firmly settled yet,[7] which is probably why it has also led to the emergence of a brand of contributions focusing on the concept of 'group privacy',[8] and to Barocas and Nissenbaum's notion of 'reachability' as an alternative to identifiability.[9]

In a 2018 contribution,[10] Purtova has put back the issue of the definition of personal data at the centre of the academic debate.[11] In her paper, she provokingly yet quite convincingly argues that the scope of the notion of personal data is an ever-expanding one. She sketches a double scenario, where (i) everything becomes information (because of advances in technology allowing for the digitization of our environment), and (ii) all this information qualifies as personal data, thereby falling under the scope of data protection law. Hence data protection law as 'the law of everything'. In order to reach this conclusion, she relies upon two sets of assumptions. The first one is of a legal nature and is based on the official interpretation of the notion of personal data in EU law as has been defined in the authoritative opinion of the former Article 29 Working Party (Art 29 WP), and confirmed in the case-law of the Court of Justice of the EU. The second one of a more hypothetical nature insofar as it relies upon a number of assumptions concerning the development of certain technologies and their affordances, more particularly, machine learning in the context of so-called smart environments.

The goal of this contribution is to take Purtova's 2018 paper and her 'law of everything' scenario seriously. It will therefore proceed along the following steps. It will first summarize Purtova's argument, in both its legal and technical assumptions. Secondly, it will explore more in detail Purtova's technical assumption by providing an in-depth analysis of the way in which machine learning technology is deployed in smart environments. Thirdly, it will discuss whether this more detailed analysis of machine learning can bring some more detailed insights into Purtova's argument. Crucially, it will find that Purtova's account of data-driven technological affordances is a bit over-optimistic: there are still technical hurdles before our environment is fully datified. Similarly, whereas data can be said to always relate to individuals in the context of smart environments, the contribution argues that her account of 'the end of anonymisation' relies upon a seemingly purely technological account of anonymization (or its counterpart identifiability), which sits ill at ease with the socio-technical dimension of anonymization, which the technical literature has evidenced.[12] Such socio-technicality entails a contextual, case-by-case analysis of anonymization (even though some cases are obviously clearer than others),[13] which is at odds with claims that

2 See, Jessica Eynard, *Les Données Personnelles: Quelle Définition Pour Un Régime de Protection Efficace?* (Michalon Paris 2013).

3 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 (General Data Protection Regulation), [2016], *OJ L 119/1*.

4 See, eg Luc Rocher, Julien M Hendrickx and Yves-alexandre De Montjoye, 'Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models' (2019) 10 Nature Communications 1, and the references contained therein.

5 See, Wim Schreurs and others, 'Cogitas, Ergo Sum. The Role of Data Protection Law and Non-Discrimination Law in Group Profiling in the Private Sector' in Mireille Hildebrandt and Serge Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer Dordrecht 2008).

6 See, Council of Europe, 'The Protection of Individuals with Regard to Automatic Processing of Personal Data in the Context with Regard to Automatic Processing. Recommendation CM/Rec(2010)13 and Explanatory Memorandum' (2010).

7 See, eg Irish Data Protection Commissioner, 'Annual Report' (2017) 16.

8 See, eg Brent Mittelstadt, 'From Individual to Group Privacy in Big Data Analytics' (2017) 30 Philosophy and Technology 475.

9 Solon Barocas and Helen Nissenbaum, 'Big Data's End Run around Anonymity and Consent' in Julia Lane and others (eds), *Privacy, Big Data, and the Public Good* (CUP Cambridge 2014).

10 Nadezhda Purtova, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10 Law, Innovation and Technology 40.

11 A number of publications and other academic contributions have explicitly built upon her paper, see, eg Michael Veale and others, 'Algorithms That Remember: Model Inversion Attacks and Data Protection Law' (2018) 376 Philosophical Transactions of the Royal Society 1. Not to mention the ongoing references to her paper each time this issue is at stake, see, eg Sandra Wachter and Brent Mittelstadt, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' (2019) 2 Columbia Business Law Review 494.

12 Mark Elliot and others, 'Functional Anonymisation: Personal Data and the Data Environment' (2018) 34 Computer Law and Security Review 204; Miranda Mourby and others, 'Are "pseudonymised" Data Always Personal Data? Implications of the GDPR for Administrative Data Research in the UK' (2018) 34 Computer Law and Security Review 222.

13 See, Elliot and others, ibid 215–17.

advances in technology have made anonymization impossible. A similar too linear account of technology and technological development underpinning the onlife scenario at the basis of her hypothesis is at odds with the various AI winters that have taken place in the previous decades.

In other words, whereas this contribution discusses the most hyperbolic aspects of Purtova's description of the expanding scope of data protection on the account that it sometimes relies upon a too linear view of technology and technological development as well as upon a too technological view of identifiability/anonymization, it still agrees with Purtova's point that the scope of data protection law is steadily expanding given the advances in information and data-driven technology as well as their increasing and pervasive uptake at every level of the societal fabric.

However, and contrary to Purtova who sees this expanding scope as creating a risk of 'system overload' for data protection law, this contribution points to another issue which is absent from her hypothesis. Namely, the capacity of machine learning algorithms to infer knowledge and therefore to create vast amounts of data that relate to individuals in terms of content. The latter puts a different and parallel type of pressure on the data protection legal framework. This key issue, which is absent from her analysis might nonetheless be a good starting point for her search into an alternative material scope for data protection law, which is more in line with data-driven technology and the type of 'information-induced harms' it leads to.[14]

Purtova's hypothesis

Legal assumptions: a broad notion of personal data

The EU notion of personal data is defined in Article 4(1) of the GDPR.[15] According to the latter, personal data is defined as 'any information relating to an identified or identifiable natural person ("data subject")'. According to the Art 29 WP, the notion has four elements: (i) any information; (ii) relating to; (iii) an identified or identifiable; (iv) natural person.[16] Of relevance for the present discussion are points (i–iii).

Concerning the notion of information, Purtova rightly notes that there has not yet been a single Court of Justice of the European Union (CJEU) judgment determining what counts as information for the purposes of the definition of personal data.[17] Furthermore, the notion of information as defined in the Art 29 WP opinion is a broad one, even though it is not entirely clear.[18] As Purtova puts it, the Art 29 WP 'does not examine what information means, probably considering it self-evident, and focuses immediately on *what kinds of information* would fall [under the definition]'.[19] For instance, it does not clarify what exactly is meant by information, nor how it relates to other related concepts such as data, meaning, knowledge or what Purtova refers to as 'information artefacts' and which include information carriers such as books, cd, etc.).[20] The Art 29 WP starts by emphasizing the EU legislator's intention to work with a broad concept of personal data, and goes on to say that any information can fall under the legal definition, regardless of its nature (ie true or untrue, objective or subjective . . .), its content (ie it does not need to relate to the private life of individuals), or of its format, medium, or form (ie alphabetical, numerical, graphical, photographical or acoustic, kept on paper or stored in a computer memory, binary code either structured or unstructured, video, voice recordings . . .).[21] As far as the 'relating to' element is concerned, data can relate to someone because of its content (eg name, address, etc.). It can also relate to someone regardless of its content because of the purpose for which it is processed (ie to evaluate, treat in a certain way or influence the individual). Similarly, it will also relate to the individual when the processing has a result/impact thereupon (and such impact need not be 'major' , a different treatment will suffice).[22] Lastly, data will be considered personal if the individual is identified on the basis of the existing data (identified), or if he/she can be identified through additional means (identifiable). The possibility of identifiability must be assessed given 'all the means reasonably likely to be used . . . either by the controller or by another person'.[23] This requires a case-by-case

14 See, Purtova (n 10) 80.

15 Regulation (EU) No 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 ('General Data Protection Regulation') [2016] OJ L 119/1.

16 See Art 29 WP, 'Opinion 4/2007 on the Concept of Personal Data' (2007).

17 Purtova (n 10) 72.

18 For a further analysis of the meaning of information in data protection law, see Lee A Bygrave, 'Information Concepts in Law: Generic Dreams and Definitional Daylight' (2014) 35 Oxford Journal of Legal Studies 1.

19 Purtova (n 10) 48. Original emphasis.

20 ibid 49. With the only exception covering the issue of human tissue samples, which, accordingly, are the source of biometric information, rather than the biometric information itself. See art 29 WP (n 16) 8–9.

21 Art 29 WP (n 16) 6–9. See also, Purtova (n 10) 48–49.

22 Art 29 WP (n 16) 9–11. The CJEU upheld such definition of "relatability" in Case C-434/16 Peter Nowak v Data Protection Commissioner [2017] ECLI:EU:C:2017:994, para. 35 (*Nowak*).

23 Recital 26 GDPR. In a similar sense, see ibid 12–17; WP29, 'Opinion 4/2007 on the Concept of Personal Data' (2007), 12–17.

assessment taking into account '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'.[24] Thus, identifiability is a contextual criterion that can be hard to be determined in an absolute way.[25]

As one can see, it is quite easy for a piece of information to legally qualify as personal data. The EU definition goes well beyond the traditional biographical data,[26] and also encompasses data that has no 'personal content' as long as it is used for the purpose of affecting the individual or ends up doing so in practice.

Technological assumption: data-driven environments or the *onlife* condition

On top of this broad definition of personal data, which in practice leaves little data/information out of its scope, advances in digital technology only further increase the scope of personal data notion.

Purtova describes the current and future technological landscape from the perspective of the so-called onlife manifesto, which envisages a future where our daily existence is mediated through, and transformed into information technology.[27] This scenario is similar to the datification hypothesis put forth by Mayer-Schönberger and Cukier,[28] and points to the fact that our whole environment is being 'datified', that is, transformed into data and information. These technological scenarios, can be said to be variations on the Ambient Intelligence vision of technology, which itself has been associated to the so-called Web 3.0, also known as 'semantic web',[29] or Internet of Things (IoT).[30] This scenario has also been more recently referred to as one of data-driven agency.[31] It is one characterized by smart infrastructures, or cyber-physical infrastructures[32] where objects are mutually connected, and are endowed with decision-making capabilities, hence their (data-driven) agency. As Hildebrandt puts it, data-driven agency can be defined as 'a specific type of artificial intelligence, capable of perceiving an environment and acting upon it, based on the processing of massive amounts of digital data'.[33] In other words, the environment is transformed into digital data and objects in turn are endowed with agency-like capabilities.

In other words, one of the explicit hypothesis upon which she bases her technological assumption is one where the 'upgrading' of physical objects with digital capabilities endows them with a type of *sui generis* agency, in line with their condition as object.[34] This agency is characterized by the fact that such digital or smart environment tends first and foremost to adapt in real-time to its users', that is, 'a frictionless world that surreptitiously adjusts the environment to the needs and desires of its users'.[35] Purtova takes the example of a smart city where individuals are subject to real-time individualized treatment thanks to the massive collection of data rendered possible by the datification of the whole environment. This allows for instance to regulate the speed of public escalators in order to promote physical exercise, or to adjust the warmth and intensity of street lightning in order to prevent undesirable behaviour.[36] In this sense, this real-time adaptability which characterizes this data-driven agency can also be seen as the taking of real-time decisions that impact individuals one way or another.[37]

Convergence of the two assumptions: data protection law as 'the law of everything'

In other words, data protection law becomes the law of everything for the following reasons. In a data-driven context, where our very infrastructure and materiality becomes digital, everything becomes information, precisely as a result of being part of such 'cyber-infrastructure'.[38] And since the legal definition of personal data is such a broad one, it is very likely that everything will be personal data. Hence data protection as the law of everything. She takes the example of weather data in a

24 Recital 26 GDPR.

25 On this matter, see Michèle Finck and Frank Pallas, 'They Who Must Not Be Identified — Distinguishing Personal from Non-Personal Data under the GDPR' (2020) 10 International Data Privacy Law 1.

26 See, Eynard (n 2).

27 Purtova (n 10) 41.

28 Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data A Revolution That Will Transform How We Live, Work and Think* (John Murray New York 2013).

29 Tim Berners-Lee, James Hendler and Ora Lassila, 'The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities' (2001) 284 Scientific American 34.

30 The expression was first coined by Kevin Ashton in 1999 during a presentation he made whilst working for Procter & Gamble, see Kevin Ashton, 'That "Internet of Things" Thing: In the Real World, Things Matter More than Ideas' [2009] RFID Journal 1.

31 Mireille Hildebrandt, *Smart Technologies and the End(s) of Law* (Edward Elgar Publishing Cheltenham UK, Northampton, MA, USA 2015).

32 See, Purtova (n 10) 55.

33 quoted in ibid 55–56; See Mireille Hildebrandt, 'Law as Information in the Era of Data-Driven Agency' (2016) 79 Modern Law Review 1.

34 Or to speak like Latour, their 'objectity', see Bruno Latour, *Reassembling the Social An Introduction to Actro-Network-Theory* (OUP Oxford 2005).

35 Hildebrandt quoted in Purtova (n 10) 56.

36 Ibid.

37 see, ibid 56–57.

38 see, ibid 55.

smart city context, namely, the Stratumseind Living Lab (SLL).[39] One of the goals of this project is to support local business in the HORECA sector, notably by fighting criminality and vandalism on the streets. It therefore relies upon a smart/data-driven environment in order to prevent such behaviour from happening (and/or de-escalating it). It does so by gathering extensive information from the datified environment. This includes street-relevant information collected through video, acoustic cameras, sound sensors, WiFi tracking, and a weather station. All this data is stored in a database, serves as the back-end for this data-driven environment to adapt in real time, eg by dimming the light of the light poles as a way to de-escalate violence.[40] In this scenario, weather data can be considered as personal data.[41] In this scenario the various information that are constitutive of 'the weather' (rainfall per hour, temperature, wind direction and speed) are datified in order to be processed by the digital environment. Furthermore, even if weather data does not relate to the individuals at stake in terms of content, in relates to them either in purpose or in impact since it is embedded in the real-time adaptability of the environment (which in this particular case is explicitly predicated upon the prevention of undesirable behaviour). Furthermore, in the context of the high-dimensional database at stake (ie one that contain many data points as a result of the datified environment), the identification of the individuals at stake is reasonably likely at the very minimum.[42]

In sum, in the scenario that Purtova describes, everything is personal data because of the real-time adaptability of the data-driven environment, which can be conflated to the taking of decisions about individuals, and therefore, impacts people by default. Furthermore, in such data-rich environment, the threshold of reasonable likeliness for the identifiability of the individuals ceases to be an issue.[43]

AI and smart environments

Whereas the previous section provided a summary of Purtova's hypothesis, the present section will look more into detail into the technological aspect of smart environments. The main findings of this analysis will then be used in the next section in order to discuss Purtova's hypothesis. This section will start by clarifying what is machine learning and how it operates, before looking more into detail into how it is deployed in smart environments.

Defining AI

As Jordan argues, most of what is referred to AI nowadays, is in fact machine learning (ML), a sub-branch of AI.[44]

The whole point of machine learning is to enable computers to achieve specific tasks. As Samuel, one of the founders of the discipline (and who actually coined the expressions 'machine learning') put it, the idea is to programme computer to achieve specific tasks which, 'if done by human beings or animals, would be described as involving the process of learning'.[45] In other words, computers achieve and do things by learning from experience,[46] and as they learn they improve their performance.[47] In this sense, machine learning has been described as focused on a set of two interrelated questions. First, 'how can one construct a computer system that automatically improves through experience?' And secondly, 'what are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organisations?'.[48]

39 On the SLL, see Maša Galič, 'Surveillance and Privacy in Smart Cities and Living Labs: Conceptualising Privacy for Public Space' (Optima Grafische Communicatie Rotterdam 2019).

40 Purtova (n 10) 59.

41 Weather data is constituted of 'rainfall per hour, temperature, wind direction and speed' ibid 58. For a more nuanced analysis of this issue, see Maša Galič and Raphaël Gellert, 'Data Protection Law beyond Identifiability? Atmospheric Profiles, Nudging and the Stratumseind Living Lab' (2021) 40 Computer Law & Security Review 1.

42 Ibid 58–59.

43 Not to mention that when the purpose of a processing operation is to impact the individual, the art 29 WP argues that such individual is identifiable because identification is intended per definition. see Art 29 WP (n 16) 16.

44 Michael I Jordan, 'Artificial Intelligence — The Revolution Hasn't Happened Yet' [2018] Medium 1, 3. Of course in the context of smart environments one could also make reference to computer vision or natural language processing. However, with the advance of deep learning it becomes increasingly complex to neatly separate between these fields, see eg Yoav Goldberg, 'A Primer on Neural Network Models for Natural Language Processing' (2016) 57 Journal of Artificial Intelligence Research 345; N Sebe and others, *Machine Learning in Computer Vision* (Springer Dordrecht 2005).

45 Arthur L Samuel, 'Some Studies in Machine Learning Using the Game of Checkers' (2000) 44 IBM Journal of Research and Development 206, 7.

46 Ibid.

47 This is the canonical definition of Mitchell for whom a machine is said to learn if its performance at some defined task or tasks improves with experience, see Tom M Mitchell, *Machine Learning* (McGraw-Hill Redmond, Ithaca 1997).

48 Michael I Jordan and Tom M Mitchell, 'Machine Learning: Trends, Perspectives, and Prospects' (2015) 349 Science 255, 255. Experience and learning for computers can indeed be amenable to statistics and probabilities, see Samuel (n 49); Ronald A Fisher, 'Theory of Statistical Estimation' (1925) 22 Mathematical Proceedings of the Cambridge Philosophical Society 700.

The various machine learning tasks

The tasks that computers can achieve are usually divided into various sets of tasks, or what Belaidouni and Miraoui refer to as 'machine learning paradigms'.[49] In other words, human defined tasks are redefined as pertaining to one of these 'machine learning paradigms'.

The first task is known as supervised learning. It mainly has to do with the predictive abilities of machines. It involves receiving labelled data (ie input data as well as corresponding output data) in order to make predictions on future instances of unlabelled data.[50] This would be the case for instance with an algorithm inferring users' viewing preferences in order to develop a personalized TV recommendation system.[51] The second task is known as unsupervised learning. It is concerned with unlabelled data (ie input data without corresponding output data). The main goal of this learning paradigm is therefore to extract knowledge from the dataset in order to label the data, that is, to classify it. It can be used for instance for the purposes of recognizing and/or assess in real time human actions (or spam detection, etc.).[52] The third type of task is known as reinforcement learning. Contrary to the two previous paradigms, which are in both cases concerned with the gaining of knowledge, this is not the case for reinforcement learning, which is why it is described as 'model-free'.[53] Instead of learning new information and knowledge, its purpose is to learn to 'control a system so as to maximise a numerical performance measure that expresses a long term objective'.[54] In other words, the maximization of performance simply depends upon learning the expected utility in a given situation.[55] Reinforcement learning therefore is concerned with issues of 'action', and is at work with any autonomous computing such as self-driving cars, robots, and more generally, with the 'autonomous' dimensions of things.[56]

From machine learning paradigms and tasks to 'real world' tasks

The previous section has provided a brief overview of the way that machine learning internalizes what it means to perform a task. These various tasks or paradigms are then mobilized in order to perform a number of tasks which are defined from a human viewpoint.

This is crucial because the technological scenario of data-driven smart environments upon which Purtova relies for her hypotheses is not simply based on machine learning as if it were a 'monolithical thing'. On the contrary, the literature on smart environments has shown that a smart environment relies upon a number of tasks. There are variations among the literature as to the exact type of tasks and their mutual interactions, [57] but it is safe to ground the description upon the following four main tasks. Namely detection, recognition, prediction, and optimization.[58] These various tasks each involve and entail some of the machine learning paradigms. As Belaidouni and Miraoui argue, data-driven or smart environments are predicated on their capability to be context-aware. By capturing a number of contextual attributes (eg user's current positions or activities, the surrounding environment), they are better able to understand what the user is trying to do and what services they might need. This entails that smart environments depend upon their ability to learn and extract knowledge from their environment in view of their decision-making and optimization thereof.[59]

According to Stenudd, detection problems are meant to recognize the occurring activities at stake. They therefore rely upon unsupervised data-mining algorithms,[60] and usually involve using various sensors.[61] Recognition problems aim to classify the recognized activity within a

49 Somia Belaidouni and Moeiz Miraoui, 'Machine Learning Technologies in Smart Spaces' in Sergey Balandin, Michele Ruta and Moeiz Miraoui (eds), *Ubicomm 2016: The Tenth International Conference on Mobile Ubiquitous Computing Systems, Services and Technologies* (International Academy, Research, and Industry Association Wilmington, New York 2016) 53. Note that each of these tasks relies upon specific algorithms. Neural networks are very popular for supervised learning for instance, see ibid.

50 See, eg Belaidouni and Miraoui (n 53) 53.

51 Ibid.

52 Ibid.

53 Ibid. In the context of machine learning, knowledge is referred to as the model (cf actionable) which has precisely been learn from the data. This is why Wu and others argue that knowledge is 'a broad concept that includes the general principles and natural laws related to every object' Qihui Wu and others, 'Cognitive Internet of Things: A New Paradigm Beyond Connection' (2014) 1 IEEE Internet of Things Journal 129, 136.

54 Belaidouni and Miraoui (n 53) 53, and the internal references contained therein.

55 Belaidouni and Miraoui (n 53) 53.

56 Ibid 54.

57 Belaidouni and Miraoui for instance divide the tasks in terms of recognition, prediction, adaptation, optimization adaptation ibid 52, whereas Wu and others conceive things in terms of perception-action, massive data analytics, semantic derivation and knowledge discovery, intelligent decision-making, and on-demand service provisioning Wu and others (n 57) 132.

58 Sakari Stenudd, 'A Model for Using Machine Learning in Smart Environments' in Mika Rautiainen and others (eds), *Grid and Pervasive Computing Workshops - International Workshops, S3E, HWTS, Doctoral Colloquium, Held in Conjunction with GPC 2011, Revised Selected Papers* (Springer Berlin, Heidelberg 2012).

59 Belaidouni and Miraoui (n 53) 53.

60 Stenudd (n 62) 26.

61 Belaidouni and Miraoui (n 53) 53.

set of pre-defined activities. This is also an issue of classification. But contrary to detection, this is a matter of supervised learning.[62] The next problem at stake is prediction, which, as its name indicates, should enable the smart environment to foresee 'the most probable event or subsequent activity'.[63] This is also a matter of supervised learning (and can be thought of as a classification or regression problem).[64] The last machine learning problem is that of optimization, which is about the control of the smart environment, and further, about endowing the environment with decision-making abilities, which is the ultimate goal of data-driven environments. It therefore consists in selecting the path in a smart traffic system,[65] or preventing the water from flowing from a domestic tap.[66] It is based upon reinforcement learning.[67] As Wu and others put it, optimization (or decision-making) is about 'choosing an action from an action set', and more in particular, to being able to intelligently adjust its decisions based on past actions.[68]

Regardless of the framework adopted for understanding the role of machine learning in smart environments, it remains undisputed that for all the steps leading to the, final, decision-making one, the key concern consists in analysing the collected data, inferring useful information therefrom. These are classical data analytics, or data mining/KDD tasks,[69] which is why they rely upon supervised and unsupervised learning. The optimization/decision-making step is not concerned with the analysis of data, but rather with autonomic actions, which is why it relies upon reinforcement learning.[70] As Mohammadi & Al-Fuqaha put it, in reinforcement learning, 'there is no output for the training data',[71] meaning that there is no learning or analysis to be made from the data since there is no output to be found or learned. Instead, the point is to reward the choice of the right action. The point therefore is to find an action for each state of the system so that the total reward of the learning agent is maximized on the long term.[72] Or to put it differently, 'reinforcement learning aims to imitate the learning process of humans', meaning that the machine learning agent generalizes on the basis of past experience of action in order to confront new and unknown situations.[73]

AI in smart environments: back to Purtova's hypothesis

The rapid overview of the way in which machine learning is deployed in smart environments is considered useful because it allows us to consider Purtova's hypothesis from a finer-grained perspective. More in particular, it will allow to make two of the key points of this contribution.

First, it will discuss accounts on the current capabilities of data-driven technology, which would have it that datification is a reality (and hence, 'everything is information'). Existing algorithms and infrastructures do not have just yet the capacity to transform everything into information (first key point). Secondly, it will look into the notion of personal data. It confirms her point that in a smart environment context, data will always relate to people. Irrespective of their content, whether individuals are specifically targeted by the smart environment (ie purpose), or simply happen to be affected by its transformation and optimization as a by-product (ie impact), the data at stake will relate to them. However, it will discuss her point that advances in technology make anonymity 'a thing from the past' by emphasizing that anonymization is more than a property of dataset: it is a complex, contextual, socio-technical phenomenon (second key point).

Everything is information

As one remembers, Purtova's hypothesis rests upon the assumption that our very material reality becomes information as a result of being part of such 'cyber-infrastructures' and smart environments.[74]

This assumption is still nuanced by the technical literature on machine learning in smart environments (first key point). In a recent paper, Mohammadi and Al-Fuqaha for instance have discussed the fact that at present the huge majority (over 90 per cent) of the data generated in the context of a smart city remains wasted, insofar as no information and/or knowledge can be derived and extracted therefrom.[75] This they argue, can be imputed to the fact

62 Stenudd (n 62) 26; Since the output data is indeed already known. See also Belaidouni and Miraoui (n 53) 52.

63 Ibid; Stenudd (n 62) 26.

64 Stenudd (n 62) 26–27.

65 Wu and others (n 57) 137.

66 Mehdi Mohammadi and Ala Al-Fuqaha, 'Enabling Cognitive Smart Cities Using Big Data and Machine Learning: Approaches and Challenges' (2018) 56 IEEE Communications Magazine 94, 99.

67 Stenudd (n 62) 26–27; As indicated Belaidouni and Miraoui add the intermediate step of adaptation as a preliminary to optimization Belaidouni and Miraoui (n 53) 52.

68 Wu and others (n 57) 137.

69 On the difference and entanglements between machine learning and data mining/KDD, see <https://www.educba.com/data-mining-vs-machine-learning/> last accessed 14 June 2019.

70 Mohammadi and Al-Fuqaha (n 70) 95.

71 Ibid.

72 Ibid.

73 Ibid 97.

74 See, eg Purtova (n 10) 55.

75 Mohammadi and Al-Fuqaha (n 70) 94.

that current machine learning algorithms still rely upon fixed training models, and a static context.[76] This stands in stark contrast with smart cities environments, which are optimized for scalable, real-world data and real-time data analytics.[77] In other words, no information or knowledge can be derived therefrom,[78] which means it is pure noise and therefore unusable.[79]

Every information relates to an individual

Purtova argues that it matters little that the data relates to the data subject content-wise. Whether individuals are specifically targeted by the smart environment (ie purpose), or simply happen to be affected by its transformation and optimization as a by-product (ie impact), the data at stake will relate to them. The present analysis confirms this point.

Dalla Corte has offered a rebuttal of this stance.[80] He reminds of the contextual nature of personal data by having regard to the life-cycle of a processing operation. As he puts it, '[data] gets created, collected, processed, re-shaped, aggregated, stored, and eventually deleted'.[81] In the case of data relating in terms of purpose and impact/result, Dalla Corte therefore argues that the data will in fact be relating to individuals only at specific stages of the processing life-cycle. This argument finds weight in the European Data Protection Supervisor's (EDPS) analysis of statistical data. It has shown that the personal nature of a piece of statistical data will vary according to the stages of the statistical processing.[82] The weather data from Purtova's example will indeed relate to individuals at the life cycle stage when it is used to regulate traffic or impact individuals one way or another. Yet, from the perspective of Dalla Corte's analysis, there are reasons to believe that it would not relate to these individuals when it is cleaned or aggregated for instance.[83] However, and as a direct rebuttal of Dalla Corte's hypothesis one can also argue that if the purpose of the processing is to impact the data subject from the very beginning, then there is no reason to differentiate between the various stages of the processing, and the data can be considered personal since its collection. This finds confirmation in the previously mentioned Opinion of the EDPS. When statistical data is aggregated, it is potentially anonymous because it is only used for statistical purposes to the exclusion of any form of decision-making. This stance is also confirmed by the Art 29 WP. In its opinion on the concept of personal data it makes it very clear that the data will relate in purpose (and in impact/result) even in cases where the data is 'used or likely to be used, taking into account all the circumstances surrounding the precise case'.[84] In other words, it suffices that at the beginning of the collection the data will likely be used to affect the individual (in purpose or result) for it to relate. This directly contradicts Dalla Corte's point, and confirms Purtova's.

Every information relates to an identified/able individual

Another discussion of Purtova's hypothesis can take place at the identifiability level (second key point). As one remembers, she argues that because of the existence of high-dimensional databases that underpin smart environments, in practice, individuals are always at least identifiable.[85] However, one can make a number of points regarding this claim.

Identification through additional information

First, as Dalla Corte argues, in determining whether the data subject is identified/able, it is necessary to distinguish between data that relates to individuals in content and data that relates in purpose/impact. In the former case, there is an overlap as the relating content also functions as an identifier (eg name, address, etc.). In the

76 In a similar sense, see also Nikolas Zygouras and others, 'An Active Learning Framework Incorporating User Input For Mining Urban Data: A Case Study in Dublin, Ireland', *UrbComp'16* (2016); Alaa E Abdelhakim and Wael Deabes, 'Handling Missing Annotations In Supervised Learning' (2020) arXiv:2002 arXiv 1.

77 Mohammadi and Al-Fuqaha (n 70) 100–101.

78 see, Wu and others (n 57).

79 Of course one can argue that noise is data and therefore this does not contradict Purtova's hypothesis. But the definition of personal data talks about 'information that relates', and noise clearly is not information. This is not only a semantic issue since machine learning relies upon the Data Information Knowledge Wisdom Pyramid, entailing that for the algorithmic process to be successful it cannot be stuck at the data level, see Russell Ackoff, 'From Data to Wisdom' (1989) 16 Journal of applied systems analysis 3. Secondly, if this data cannot be processed, it is data only in name. Purtova's hypothesis is obviously about data or information that can be processed.

80 Of course Dalla Corte's point can also be read from the identifiability viewpoint, and the need to collect additional identifying information. This point however is treated separately in the next section. For this reason, this section only draws on Dalla Corte's argument as far as the 'relating to' requirement is concerned.

81 Lorenzo Dalla Corte, 'Scoping Personal Data: Towards a Nuanced Interpretaton of the Material Scope of EU Data Protecton Law' (2019) 10 European Journal of Law and Technology 1, 11. See also the GDPR definition of the processing life-cycle embedded into the definition of processing, art 4(2) GDPR: 'collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction'.

82 EDPS, 'Opinion 10/2017 on Safeguards and Derogations under Article 89 GDPR in the Context of a Proposal for a Regulation on Integrated Farm Statistics' (2017).

83 See Dalla Corte (n 85) 11.

84 Art 29 WP (n 16) 11. See also p 12 for the relation in impact.

85 Purtova (n 10) 59.

latter case however, additional identifying information will necessarily be required (provided that the data is encoded in an identifiable form of course).[86] In this sense, data that relates to the data subject other than in content leads by definition to a higher threshold of identification since it requires additional auxiliary information in order to render the individual identifiable. Therefore, data about things –or in Purtova's example data about the weather,[87] may very well become personal data, but only when it can be tied through sufficient additional auxiliary data to the individual.[88] Purtova might object to this argument by saying that finding auxiliary identifying information is not an issue in a smart environment context that relies upon high dimensional databases (cf above). However, the argument of high-dimensional databases can also be taken as a rebutting one. Technical literature for instance sees the high dimensionality of data sets as a challenge more than anything else (and actually refers to it as 'the curse of dimensionality').[89] Powerful pre-processing algorithms are indeed needed in order to transform the high dimensionality into more than mere noise, and further, into valuable data allowing for semantic derivation and knowledge extraction. The challenge therefore with high dimensional datasets is precisely to reduce the dimensionality so that the data can be usable.[90]

Identification as a socio-technical process: the data environment

Furthermore, and critical to the argumentation, it seems as though Purtova reduces the issue of identifiability to a merely technological one (ie provided sufficiently powerful and dimensional databases, data are identifiable per definition). However, this point seems to overlook the socio-technical dimension of anonymization, which the literature has underlined.[91] Rather than simply a property of data (or a dataset),[92] anonymization is a socio-technical concept that can be captured through the concept of the 'data environment'. Such data environment is indeed constituted of four factors that are at play in the identification process. Namely: 'other data', 'infrastructure', 'governance', and 'agency'.[93] 'Other data' and 'infrastructure' respectively refer to the existence of additional data, and to the structure of databases both at hardware and software level. 'Agency' is a factor insofar as 'there is no re-identification risk without human agency – this may seem like an obvious point but it is one that is often overlooked and the least understood'.[94] Governance refers to the existing policy and legal framework governing the use of data. This is why they argue that in order to adequately assess the chances of identification, one must understand 'who the key protagonists are, and how they might act'.[95] This also means that anonymity is not a question that can be answered in absolute terms. On the contrary it is always context-dependent.[96]

Traces of the socio-technical dimension of anonymization can also be found in the GDPR itself since Recital 26 GDPR seems to underscore human agency when it refers to 'the means reasonably likely to **be used** (. . .) **either by the controller or by another person** to identify the natural person'.[97] The role of human agency was also confirmed (at least implicitly) in the *Breyer* case, since the Court concluded to the identifiable nature of the data at stake not solely on the basis of the existing technical means (as had been suggested by the Advocate General), but rather, because it was reasonably likely to consider that the technical means at hand could be used by the party at stake.[98] Similarly, the Art 29WP seems to confirm the socio-technical nature of anonymization by putting forth a number of organizational factors to determine the identifiable nature of a dataset such as the context, type of processing, or the purpose of the processing (ie when the purpose of the processing

86 Dalla Corte (n 85) 9; See also ibid: 'When the information relates to the data subject due to its content, the very substance of the information may lead to the identifiability of the data subject'.

87 Also referred to as 'attribute data', see Gerald Spindler and Philipp Schmechel, 'Personal Data and Encryption in the European General Data Protection Regulation' (2016) 7 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 163, 168.

88 Dalla Corte (n 85) 10.

89 Michel Verleysen, 'Machine Learning of High-Dimensional Data: Local Artificial Neural Networks and the Curse of Dimensionality' (Université Catholique de Louvain 2000); Piotr Iwo Wójcik and Marcin Kurdziel, 'Training Neural Networks on High-Dimensional Data Using Random Projection' (2019) 22 Pattern Analysis and Applications 1221.

90 Wu and others (n 57) 133.

91 Mourby and others (n 12); Elliot and others (n 12).

92 Elliot and others (n 12) 206.

93 Mourby and others (n 12) 231.

94 Ibid.

95 Ibid. This should not be taken to mean that a data controller or processor 'must want to identify'. This could also happen by coincidence for instance when the person behind the screen happens to look at two datasets, and in doing so identifies someone. This is referred to as spontaneous identification by Elliot and others (n 12) 213.

96 See Elliot and others (n 12) 212–15, who talk about functional anonymization.

97 Emphasis by the author. See also The Art 29 WP which counts purpose as one of the parameters to take into the determination of the reasonable likeliness of the means, Art 29 WP (n 16) 15.

98 CJEU, Case C-582/14, *Breyer* [2016] ECLI:EU:C:2016:779 para 45. For a discussion of this issue, see Spindler and Schmechel (n 91) 167–68; Mourby and others (n 12) 226.

is to treat a data subject in a certain way, then the identifiability of the individuals is implied by the very processing purpose).[99]

Profiling and the socio-technical dimension of identification

It should be clarified however that if from a technical perspective the socio-technical dimension of anonymization leaves little doubt, this is not yet the case within legal discussions. This is maybe best illustrated with legal debates around profiling, which it is interesting to briefly mention since smart environments heavily rely upon thereupon (cf., steps 1-3 as the construction of a profile and step 4 as its application).[100] As a matter of fact, there are two competing views. Those who argue that a profile always amounts to personal data, and those who argue that this is not necessary always the case.

Schreurs and others have argued that profiling does not involve the processing of personal data when the individual cannot be identified on basis of the existing data. This could be the case of a profiling operation the goal of which is to infer the type of customer at play (e.g., low purchasing power, 'pickier', etc.) on the basis of behavioural biometrics such as the way a shopping trolley is driven in a supermarket . Unless additional identifying information is used, they argue that the profiling operation will not involve personal data.[101] One can argue that this view purports a strictly technical view of identification (or its counterpart, anonymization): it is dependent upon the existence of identifying data in the dataset.

On the other end of the spectrum, there are those who argue that profiling amounts to the processing of personal data even when the individual is not identified on the basis of existing data. This is the case of the Council of Europe Recommendation CM/Rec(2010)13 on profiling. Therein it argues that even if based on an anonymous profile, the application of the profile to individuals entails as such that these individuals are at least identifiable, meaning that personal data is processed since the moment of collection.[102] This is clearly a socio-technical view on identifiability, which takes the actors' intentions into account.[103]

From the perspective of the present contribution, the socio-technical dimension of anonymization leaves little doubt, meaning that anonymization requires a contextual analysis (irrespective of whether it leads to a broader scope of personal data).

Advances in machine learning technology and the definition of personal data

The last section discussed two of the key points put forth in this contribution. Namely, a sometimes too linear account of existing data-driven technology, and a too technological account of anonymization. This section addresses a related point, namely a similar linear account of technological developments, which in turn would address existing caveats (ie full 'datification' of the environment, and the 'end' of anonymous datasets), and make her hypothesis come true (ie future technological capabilities). This section will therefore look at shorter term technological evolutions concerning the datification of the environment and anonymization, before taking a longer-term perspective on the future of machine learning technology and what this means for the advent of the onlife scenario that is at the core of Purtova's hypothesis.

Technical solutions to datification

The technical literature has put forth a number of technical solutions in order to improve the datification process, and in particular the issue of wasted information.

One solution put forth is known as fog architecture. This new type of architecture would replace the current cloud architecture. Contrary to the cloud, the fog allows for the performing of local analytics, which in turn would allow processing more data in a meaningful way (ie improve the analytics capacity).[104] Another way to address these challenges is by resorting to novel machine learning tasks such as semi-supervised deep reinforcement learning. These new and hybrid algorithms could enable to make use of big quantities of unlabelled data for tasks that would normally require labelled data. Yet, and however promising, these solutions are not fully

99 Art 29 WP (n 16) 15. In a similar sense, see also, Elliot and others (n 12) 213–14.

100 See Mireille Hildebrandt, 'Profiling: From Data to Knowledge. The Challenges of a Crucial Technology' (2006) 30 Datenschutz und Datensicherheit 548, 550.

101 Schreurs and others (n 5) 246–47.

102 Council of Europe (n 6) para 57.

103 And which echoes the Art 29 WP's view on identification through purpose, Art 29 WP (n 16) 16.

104 Mohammadi and Al-Fuqaha (n 70); Ranesh Kumar Naha, Saurabh Garg and Andrew Chan, 'Fog-Computing Architecture: Survey and Challenges' (2018) arXiv:1811 arXiv 1; Hany Atlam, Robert Walters and Gary Wills, 'Fog Computing and the Internet of Things:A Review' (2018) 2 Big Data and Cognitive Computing 10.

developed and are far from constituting the reality of smart environments nowadays.[105]

Technical solutions to anonymization

As indicated, this contribution takes the perspective that anonymization is a socio-technical phenomenon. This means that advances in technology have a role to play, but only among other factors. Recent literature on anonymization has developed a re-identification algorithm that addresses these concerns, by showing that even properly anonymized datasets have nonetheless a 99.8 per cent chance of being re-identified.[106] However, this should not be equated to the end of anonymization. From a purely technical perspective, the authors acknowledge a number of constraints such as a sufficiently high number of features per concerned individual.[107] From a social perspective, the study by Rocher and others is based on the assumption of a third party de-anonymizing a dataset that has been anonymized before being shared at large, a usual practice in scientific research,[108] and that such third-party based de-anonymization is legal, which is far from being always the case.[109] This shows again that if anonymization should ever end, it would not be solely caused by advances in technology.

Long-term issues with AI and machine learning

Beyond these narrower technological issues, one can also wonder whether there aren't more general issues associated to machine learning technology and which still prevent the advent of the onlife scenario.

Even though AI and machine learning currently enjoy a 'hype', which is characterized among others by broad institutional and financial support,[110] one should keep in mind that the history of AI is one characterized by a succession of hypes and delusions, the so-called 'AI winters'.[111] As a matter of fact, Jordan argues that machine learning has not yet reached maturity, and is yet to become a 'proper' engineering science.[112] This owes to a number of factors such as the absence of commonly agreed upon levels of desired quality and associated error bars.[113] Such shortcomings might end up having dramatic short-to-mid-term consequences in the form of an epidemic of false positives and negatives.[114] More particularly, developments in big data are currently growing faster than the statistical strength of the data (ie the capacity of machine learning algorithms to make sense of it), so that many of the resulting machine learning processes are simply likely to be wrong.[115]

As a matter of fact, a number of authors have started predicting the advent of a new AI winter, not least because of the existing gap between the expectations created and the remaining major hurdles.[115] This emphasizes again that the linear account of the development of data-driven technology which undergirds the onlife scenario and Purtova's hypothesis might not give due account of the intricacies associated with the social uptake of a technology.[117]

Conclusions: data protection as the law of everything: expanding scope of personal data and the regulation of emerging digital technologies

This contribution has endeavoured to take seriously Purtova's hypothesis of data protection law as 'the law of everything'. The latter is predicated upon advances in machine learning technology following which 'everything is information', and all this information qualifies as personal data.

In discussing her hypothesis, this contribution has so far put forth two key points. First, her account of current technological capabilities and future technological development is sometimes too linear, and one might argue slighlty over-optimistic about what can be expected—now and later—from technology as far as the

105 Mohammadi and Al-Fuqaha (n 70); Paul F Christiano and others, 'Deep Reinforcement Learning from Human Preferences', *31st Conference on Neural Information Processing Systems (NIPS, 2017)* (2017); Chelsea Finn and others, 'Generalizing Skills with Semi-Supervised Reinforcement Learning' *ICLR 2017* (2017); Sangdoo Yun and others, 'Action-Decision Networks for Visual Tracking with Deep Reinforcement Learning' *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society 2017).

106 Rocher, Hendrickx and Montjoye (n 4).

107 Ibid 1.

108 See, eg ibid 2.

109 This is often illegal in the case of high-dimensional data such as genomic data, see, eg Mark Phillips, Edward S Dove and Bartha M Knoppers, 'Criminal Prohibition of Wrongful Re-identification: Legal Solution or Minefield for Big Data?' (2017) 14 Bioethical Inquiry 527. See also, *Breyer*, paras 48–49.

110 Jordan (n 48).

111 See, Dominique Cardon, Jean-Philippe Cointet and Antoine Mazières, 'La Revanche Des Neurones' (2018) 211 Réseaux 173.

112 Jordan (n 48) 3.

113 Jordan in Lee Gomes, 'Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts' [2014] *IEEE Spectrum* 1, 6. For more of the issues associated with machine learning, see eg Mireille Hildebrandt, 'Privacy As Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning' (2019) 20 Theoretical Inquiries in Law 83, 102.

114 Jordan in Gomes (n 118) 5.

115 Jordan in ibid.

116 Filip Piekniewski, 'AI Winter Is Well on Its Way' (*Piekniewski's blog*, 2018) 8 <https://blog.piekniewski.info/2018/05/28/ai-winter-is-well-on-its-way/> last accessed 1 August 2019; Jordan in Gomes (n 118) 7.

117 For a seminal account, see Bruno Latour, *Aramis or the Love of Technology* (English, Harvard University Press Cambridge, Mass; London, England 1996).

transformation of the physical world into information is concerned. Secondly, she puts forth a technological view of anonymization that might overlook anonymization as a complex socio-technical phenomenon (which enables her to argue that anonymization is on the verge of disappearance). Such a socio-technical dimension means that advances in technology alone cannot put an end to anonymization, the existence of which remains a contextual assessment.[118] In other words, precisely because the socio-technical dimension of identifiability requires a case-by-case assessment, it prevents from making general claims such as 'all data is identifiable', or 'anonymization does not exist anymore' (even though some cases are obviously clearer than others).[119]

Does this mean that Purtova's hypothesis should be infirmed? Far from it. Even though it is possible to disagree with her on the points raised herein above, it remains hard to disagree with the broader and overall image of the steadily expanding scope of the notion of personal data given the advances in information and data-driven technology as well as their increasing and pervasive uptake at every level of the societal fabric. In particular, in the case that she refers to (the Stratumseind Living Lab—SLL), available empirical description has evidenced a wide ranging sharing between actors and databases, so that in this very case the personal nature of the data at stake is very likely indeed (but again this is a social factor pertaining to the identifiable nature of the data).[120] In other words, data protection might not be the law of everything in the hyperbolic sense that she puts forth, but its scope of application is nonetheless steadily increasing.

So, where does this leaves us? One can argue that these discussions epitomize well some of the issues surrounding the use of the notion of personal data as the key material scope of application for data protection law. Following the work of Black, one can argue that personal data is a clear example of a legal concept that is both over and under-inclusive. As she puts it:

> 'Rules are never perfectly congruent with their purpose – they are always over-inclusive and under-inclusive. Rules are inevitably either under-inclusive, failing to catch things that the rule-maker might want to catch, and/or over-inclusive, catching things that the rule-maker might not want to catch when applied to particular sets of circumstances.'[121]

In other words, Purtova's scenario is one that stretches the over-inclusivity of personal data to its limits, and which sketches the potentially lethal consequences for data protection law that go with it. This is why she foresees a situation of 'system overload' threatening the whole sustainability of EU data protection law.[122] This is so because of the potentially unlimited broad scope of data protection law coupled with its new resource intensive and non-scalable compliance regime, and with its heavy non-compliance fines regime.[123]

It is precisely these instances of over-inclusiveness that motivate Purtova to look for an alternative scope of application that would be more congruent with the pervasive datification of our societies (ie not susceptible to the type of over-inclusiveness that she points to), and which in so doing would also adequately address the data-driven or 'information-induced harms' stemming from these technologies.[124]

Thinking along with Purtova, can we think of an alternative material scope? Such an endeavour goes beyond the remits of this contribution, but it is possible to at least point to an overall direction, and this is the third key point of this contribution. Going back to Black's point about rules being both over and under-inclusive, one can argue that Purtova exclusively puts the focus on the over-inclusive dimension of the notion of personal data. But what about its under-inclusive dimension? In the present case, one can argue that personal data is under-inclusive as far as knowledge (or information) inference is concerned. The latter is key to profiling technologies such as machine learning (and is at play in tasks 1-3 underpinning smart environments),[125] yet it remains out of the scope of Purtova's discussion. In a nutshell, information inference refers to the capacity of algorithms to learn from data and information, that is, to deduce additional information from the information at their disposal. Profiling algorithms have the potential to transform trivial data (personal or not) into sensitive insights.[126] From this perspective, the main challenge in terms of personal data is not whether a piece of inferred information will qualify as personal data (ie Purtova's

118 Relying upon a number of subjective value choices such as how much weight to the purpose of the processing, what exactly is considered only a hypothetical risk of identification, etc., see Elliot and others (n 12) 216.

119 See, ibid 215–17.

120 See, Galič (n 43).

121 Julia Black, Martyn Hopper and Christa Band, 'Making a Success of Principles-Based Regulation' [2007] Law and Financial Market Review 191, 194. See also, Julia Black, *Rules and Regulators* (OUP Oxford 1997) 6–10.

122 Purtova (n 10) 75, 77.

123 Ibid.

124 Ibid 80.

125 See, Mireille Hildebrandt, 'Defining Profiling: A New Type of Knowledge?' in Mireille Hildebrandt and Serge Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer Dordrecht 2008).

126 See, Tal Z Zarsky, 'Incompatible: The GDPR in the Age of Big Data' (2017) 47 Seton Hall Law Review 995.

issue, namely everything turning into personal data), since inferred content about an individual will always relate to them in content.

The challenge here resides in the quality of the knowledge produced, that is, in the massive creation and inference of information about individuals, the veracity of which remains subject to caution, not to mention the hyper granular level of insight it can produce whether true or false. Of course, this challenge is not new and has already been underscored by the literature.[127] However, in the context of Purtova's analysis it takes on a renewed importance since knowledge inference is only sub-optimally regulated by data protection rules. As a matter of fact, nearly 15 years ago already, Hildebrandt pointed to the limits of data protection law for addressing machine learning by arguing that it focuses on data, whereas machine learning is about knowledge and the inference/creation thereof.[128] As a result, data protection is unable to, for instance, adequately protect us against 'the unwarranted application of profiles'.[129] These criticisms have been re-stated more recently in discussions surrounding Article 22 GDPR (the main provision for algorithmic regulation),[130] and in general discussions surrounding the adequacy of data protection law in the context of machine learning.[131]

This leads to the following paradoxical situation. On the one hand, advances in machine learning lead to a radical expansion of data protection's scope. This creates a number of 'internal' challenges, which Purtova has adequately underscored (cf her point on 'system overload'). However, these advances also highlight the shortcomings of data protection law as a means to adequately regulate these novel technologies in the first place (which one can refer to as 'external challenges'). In other words, this leads to a situation whereby the scope of data protection is radically expanded. Yet, and simultaneously, the quality of the protection afforded by data protection when it applies to these technologies is increasingly challenged.

Therefore, in trying to think with Purtova about another starting point for the regulation of data-driven technologies, the lack of adequate data protection-based regulation of the knowledge inference process seems like an interesting place to begin devising a material scope criteria more congruent with current advances in information technologies.

doi:10.1093/idpl/ipaa023
Advance Access Publication 8 January 2021

127 See, Paul Ohm and Scott Peppet, 'What If Everything Reveals Everything?' in Cassidy R Sugimoto, Hamid R Ekbia and Michael Mattioli (eds), *Big Data is Not a Monolith* (MIT Press Cambridge, Mass; London, England 2016).

128 Hildebrandt, 'Profiling: From Data to Knowledge. The Challenges of a Crucial Technology' (n 104) 550.

129 Serge Gutwirth and Mireille Hildebrandt, 'Some Caveats on Profiling' in Serge Gutwirth, Yves Poullet and Paul De Hert (eds), *Data Protection in a Profiled World* (Springer Netherlands 2010) 37.

130 Wachter and Mittelstadt (n 11). See also, Sandra Wachter, 'Data Protection in the Age of Big Data' (2019) 2 Nature Electronics 6.

131 Seda Gürses and Joris van Hoboken, 'Privacy After the Agile Turn' in Evan Selinger, Jules Polontesky and Omer Tene (eds), *Cambridge Handbook of Consumer Privacy* (CUP Cambridge 2017).