Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning

Mireille Hildebrandt[*]

This Article takes the perspective of law and philosophy, integrating insights from computer science. First, I will argue that in the era of big data analytics we need an understanding of privacy that is capable of protecting what is uncountable, incalculable or incomputable about individual persons. To instigate this new dimension of the right to privacy, I expand previous work on the relational nature of privacy, and the productive indeterminacy of human identity it implies, into an ecological understanding of privacy, taking into account the technological environment that mediates the constitution of human identity. Second, I will investigate how machine learning actually works, detecting a series of design choices that inform the accuracy of the outcome, each entailing trade-offs that determine the relevance, validity and reliability of the algorithm's accuracy for real life problems. I argue that incomputability does not call for a rejection of machine learning per se but calls for a research design that enables those who

[*] Tenured Research Professor of "Interfacing Law and Technology," appointed by the Research Council of the Vrije Universiteit Brussels (VUB) at the research group of Law, Science, Technology and Society studies (LSTS) at the Faculty of Law and Criminology at VUB. She is also a part-time Full Professor of "Smart Environments, Data Protection and the Rule of Law," at the institute of Computing and Information Sciences (iCIS), Science Faculty of Radboud University, Nijmegen. This Article was presented at the International Conference "The Problem of Theorizing Privacy," co-organized by Michael Birnhack, Julie Cohen and myself at Tel Aviv University on 8th January 2018, and at the Privacy Law Scholars Conference (PLSC) Europe in Brussels on 27th January 2018. I thank Eran Fisher, Bart van der Sloot and Ben Wagner and the editors of TIL for their in-depth comments and many others for seriously engaging with the content during both conferences. Cite as: Mireille Hildebrandt, *Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning*, 20 Theoretical Inquiries L. 83 (2019).

will be affected by the algorithms to become involved and to learn how machines learn — resulting in a better understanding of their potential and limitations. A better understanding of the limitations that are inherent in machine learning will deflate some of the eschatological expectations, and provide for better decision-making about whether and if so how to implement machine learning in specific domains or contexts. I will highlight how a reliable research design aligns with purpose limitation as core to its methodological integrity. This Article, then, advocates a practice of "agonistic machine learning" that will contribute to responsible decisions about the integration of data-driven applications into our environments while simultaneously bringing them under the Rule of Law. This should also provide the best means to achieve effective protection against overdetermination of individuals by machine inferences.

Introduction

Privacy is an affordance. An environment either affords us privacy or it does not. This goes for our "material" environment (tents, houses, walls), for our face-to-face environment (family, friends, colleagues, shopkeepers, teachers), for our institutional environment (education, employment, religion, economics, and the law), and for our technological environment (social networks, search engines, ecommerce, smart energy grids, connected cars, robots and the rest of the upcoming internet of things). Without shelter, without fellow human beings who respect our unwillingness to share private thoughts, and without a legal system that gives us an effective right to ward off intrusions into our private life, we have no privacy. The technological environment, however, permeates and mediates much of our material, face-to-face, and institutional environments. In a world crowded with automated decision systems, our privacy increasingly depends on the design of these systems.

In this Article, I argue for the protection of the incomputable self as core to privacy and identity in the era of surreptitious data-driven decision-making. Increasingly, tasks formerly performed by human beings are outsourced to machines. This confronts humans with computational "others" that supposedly "outperform" human experts.[1] Though their decisions may be difficult to

[1] Katja Grace et al., *When Will AI Exceed Human Performance? Evidence from AI Experts*, J. Artificial Intelligence Res. 729 (2017); Richard Susskind & Daniel Susskind, The Future of the Professions: How Technology Will Transform the Work of Human Experts (2015). *See also* Rodney Brooks, *Machine Learning Explained*, RODNEY BROOKS (Aug. 28, 2017), http://rodneybrooks.com/forai-

explain, they are said to display "high accuracy" while "optimizing" their algorithms. Combined with spectacular news items about computing systems winning highly complex games, such as chess and Go,[2] a new type of magical thinking has taken hold of the public imagination, fed by popular science revelations about "general artificial intelligence."

Many of the advances in what is called artificial intelligence have been made in the field of machine learning, notably also in domains of cognitive assessment and decision-making, such as finance, law, medicine and accounting. Machine learning depends on the brute force of computing power, combined with sophisticated statistical operations and mathematics.[3] Machine learning basically seeks to detect the mathematical target function that properly describes a dataset, hoping the function will apply to new data. The idea is that the more data are available, the greater the chances that the target function will indeed apply to new (future) data. This, however, depends on many factors, and obviously the assumption that such a function even exists is more credible in the case of a rule-based game like chess or Go than in the case of human intercourse. Also, the assumption that real life can be translated adequately into machine-readable data is flawed, even though such translation can generate new insights and be applied in a variety of productive ways. My concern in this Article is how the behaviorist assumptions of machine learning (which necessarily reduces human interaction to behavioral data) impact human identity in relation to privacy.

machine-learning-explained/ (an antidote to counter over-the-top expectations). *See also, e.g.*, COMPAS, https://doc.wi.gov/Pages/AboutDOC/COMPAS.aspx (last visited Aug. 13, 2018) (COMPAS software is used to predict recidivism); Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 Sci. Advances (2018) (concluding that COMPAS does not — so far — outperform humans); State v. Loomis, 881 N.W.2d 749 (2016) (holding a similar conclusion by the Supreme Court of Wisconsin). On the — related — issue of bias, from a computer science perspective, see Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 Big Data 153 (2017). On bias, see also *infra* Part II.A.

[2] Jose Camacho Collados, *Is AlphaZero Really a Scientific Breakthrough in AI?*, Medium (Dec. 11, 2017), https://medium.com/@josecamachocollados/is-alphazero-really-a-scientific-breakthrough-in-ai-bf66ae1c84f2. *See infra* section II.A.

[3] Thomas M. Mitchell, Machine Learning (1997) [hereinafter Mitchell, Machine Learning]; Tom M. Mitchell, *Key Ideas in Machine Learning*, *in* Machine Learning (forthcoming 2018) [hereinafter Mitchell, *Key Ideas*] (confirming and extending his position of 1997).

In Part I, I will argue that our new data-driven environment requires us to recognize, defend and protect a dimension of privacy that may have been taken for granted before, but is now under attack. This dimension concerns the foundational incomputability of human identity. I will approach this dimension by developing a theoretical account of the relational nature of human identity, highlighting the indeterminacy and the ensuing incomputability this entails. To assess the impact of an environment that tends towards pseudo-religious devotion to machine learning systems, I will extend the relational conception of privacy to an ecological understanding, addressing the fact that the relationality of human identity is shaped by the technological environment that co-constitutes us as human beings (being human is being technological).

This will enable me to investigate, in Part II of this Article, how machine learning may affect both human identity and privacy. My point is not that we have a tiny incomputable essence, whereas the rest can indeed be calculated. Instead, I argue that our essence is *that we are incomputable*, meaning that any computation of our interactions can be performed in multiple ways — leading to a plurality of potential identities. The need to navigate this plurality is what shapes and nourishes our agency; to deny or reduce this plurality is to diminish our agency. In a text-driven world, such plurality comes "naturally" based on the semantic ambiguities of natural language. In a data-driven environment, this plurality must be reinvented and protected, as otherwise consistent overdetermination by means of data-driven choice architectures may diminish our agency, as manipulation does not depend on whether the inferences of machine learning are correct, but on how they reconfigure our environment — based on the belief that they are correct. I propose to develop ways and means to engage in agonistic machine learning — rejecting unhelpful objectivist accounts of machine learning as agnostic with regard to bias (thus demonstrating that bias is core to machine learning and in itself not problematic but productive).

In Part III of the Article, I will relate the idea of agonistic machine learning to the notion of legal protection by design, as agonistic machine learning refers to the design stage of machine learning, requiring that we build adversariality and democratic participation into *the makings* of our new world. This will include a discussion of how the EU General Data Protection Regulation has the potential to provide effective and practical rights to resist and contest problematic overdetermination by machine learning decision systems.

I. Privacy as the Protection of the Incomputable Self

A. A Relational Conception of Privacy and Identity

Privacy relates to the foundational indeterminacy of human identity.[4] Instead of taking this for granted, we can explain this by building on Paul Ricoeur's work on *Oneself as Another*, on Mead's thoughts about the "I" and the "me," Helmuth Plessner's work on the artificial nature of being human, and Hannah Arendt's concept of natality.[5] With Paul Ricoeur, we can distinguish between *idem*- and *ipse*-dimensions of personal identity, where *idem* refers to sameness in the sense of both similarity and continuity (though not to identicalness), while *ipse* refers to the first-person perspective that grounds the third-person (objectified) perspective. *Idem* and *ipse* should not be understood as separate parts of a single identity, or as substances that can be located somewhere in the brain, but rather as conceptual tools to distinguish between the irreducible first-person perspective (*ipse*) that enables the construction of a third-person perspective on the self (*idem*).

Both the *ipse* and the *idem* perspectives develop in interaction with other selves in a shared environment. This highlights the primordial role of the second-person perspective, which concerns how we are addressed as a first person by another. To understand the primacy of the grammatical position of the second- and first-person perspectives for third-person (objectified) perspectives on the self, we can resort to the seminal work of George Herbert Mead. Mead explains that objectification of the "I" does not replicate the "I" but produces a "me." The "me" is constituted by the third-person perspective taken by the "I," and thereby fails to capture the "I," as "I" turn(s) into "me" every time I try to capture my ephemeral "I" (which does the capturing instead of being captured). *Idem* dimensions of human identity concern the "me" that develops from the myriad of objectifications by individual or institutional

[4] Mireille Hildebrandt, *Privacy and Identity*, *in* Privacy and the Criminal Law 43 (Erik Claes, Antony Duff & Serge Gutwirth eds., 2006). For salient accounts of the riddles of human identity, see Amélie Oxenberg Rorty, The Identities of Persons (1976).

[5] Paul Ricoeur, Oneself as Another (1992); George Herbert Mead, Mind, Self, and Society from the Standpoint of a Social Behaviorist (1962); Plessner's Philosophical Anthropology: Perspectives and Prospects (Jos de Mul ed., 2015); Hannah Arendt, The Human Condition (1958). *See also* Mireille Hildebrandt, *Profiling and the Identity of the European Citizen*, *in* Profiling the European Citizen: Cross-Disciplinary Perspectives 303 (Mireille Hildebrandt & Serge Gutwirth eds., 2008).

others, whose objectifications shape my own reflection on my self. *Ipse* dimensions of human identity concern the position from which I construct and reconstruct my identity, based on how my personal and institutional environment addresses me. In that sense, the grammatical second-person precedes both the first- and the third-person perspectives.[6] The point is that such addressing does not, however, overdetermine my self-understanding, as language enables me to resist overdetermination due to its ambiguity and the fundamental challenges posed by the so-called double contingency.[7] The latter refers to the uncertainty that grounds my anticipation of how others will anticipate me, knowing that the same abyss grounds the other's anticipation of my expectations.

This is notably the case in a pluralist society that forces individuals to develop a "me" that copes with the contradictory demands and expectations of what Mead called "the generalized other." The latter stands for the dynamic set of expectations that orient my own conduct and Mead nicely demonstrated the complexity and temporality of this "generalized other" by giving the example of one who plays basketball: the individual player must have a fair idea of the rules that define the game, of how this affects the interactions of other players in the context of a concrete game. This is not a matter of developing a "theory of mind" about what other players might do, but requires an intuitive and ad hoc grasp of what opportunities this offers in the course of an actual game. Clearly, real life requires an "I" capable of coping with a variety of games, defined by a variety of rules. Otherwise than in basketball, in "real" life these rules are both ambiguous and shifting, requiring keen attention to the performative nature of the speech acts that codetermine the myriad language games we play.[8]

Both Ricoeur and Mead highlight the interplay of the first-, second- and third-person perspectives that shape self-identity, demonstrating the dynamic

[6] This is the core of Levinas' first philosophy. *See* Bettina Bergo, *Emmanuel Levinas*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2017). *See also* JUDITH BUTLER, GIVING AN ACCOUNT OF ONESELF 11 (2005).

[7] Raf Vanderstraeten, *Parsons, Luhmann and the Theorem of Double Contingency*, 2 J. CLASSICAL SOC. 77 (2007). *Cf.* MIREILLE HILDEBRANDT, SMART TECHNOLOGIES AND THE END(S) OF LAW: NOVEL ENTANGLEMENTS OF LAW AND TECHNOLOGY 51-57 (2015).

[8] Wittgenstein's "language games" come to mind, in combination with Austin's speech act theory. LUDWIG WITTGENSTEIN, PHILOSOPHICAL INVESTIGATIONS (Peter Hacker & Joachim Schulte eds., G. E. M. Anscombe et al. trans., Wiley-Blackwell rev. 4th ed. 2009) (1953) [hereinafter PHILOSOPHICAL INVESTIGATIONS]; JOHN L. AUSTIN, HOW TO DO THINGS WITH WORDS (2d ed. 1975). On their relationship, see CHARLES TAYLOR, PHILOSOPHICAL ARGUMENTS (1995).

and relational nature of individual identity construction. The recursive nature of human self-understanding can be further explained by what Plessner called the ex-centricity of self-perception, emphasizing that our notion of self depends on a constitutive de-centering: we look back upon our self via the gaze of others, not because we care so much about what others think of us, but because our self is constituted when imagining the view others have of us (another way of framing the double contingency mentioned above). Access to the *ipse* dimension of our self is always mediated by the *idem* dimensions that shape us; our self is born from the *friction caused b*y and the *resistance against* the way others address and define us.

This brings us, finally, to Arendt's salient work on *the human condition*, drawing on her concept of natality.[9] Arendt understands human freedom as a practice of speaking rather than calculating, of acting rather than behaving, and of facing the uncertainty of being (mis)understood in one way or another. Her concept of natality emphasizes the inexorable indeterminacy of human action, initiated by the new and uncertain beginning that defines newly born infants and their concomitant need to learn, to start from scratch, spilling over into adult life as the continuous need to learn and to review what is learnt.

To grasp the abysmal implications of our natality, which is obviously related to our mortality, we may turn to Herbert Simon in his highly relevant 1983 address on machine learning.[10] Though readers familiar with Arendt may not have expected to learn about natality from one of the founding fathers of artificial intelligence, Simon seems to "capture" some of its salient features.[11] He discusses what computer engineers would now call the "legacy problem" of software programs: their inoperability with new or other programs and the bugs that cannot always be detected or removed. This leads him to the

[9] Arendt, *supra* note 5, at 177-78 (highlighting the connection between the concepts of "action" and "beginning" (referring to the Greek *archein* and the Latin *agere*) and noting: "The new always happens against the overwhelming odds of statistical laws and their probability . . . . The fact that man is capable of action means that the unexpected can be expected from him . . . ."). Natality, however, cannot be taken for granted; it may be destroyed once the shift from thinking in terms of action to thinking in terms of behavior determines our capabilities. *Id.* at 322.

[10] Herbert A. Simon, *Why Should Machines Learn?*, *in* 1 Machine Learning 25 (Ryszard S. Michalski, Jaime G. Carbonell & Tom M. Mitchell eds., 1983).

[11] I am using the term "capture" as in Agre's seminal text on privacy, where capture refers to the way computers require the datafication of their environment, underlining that capturing data is an intervention, not merely a recording. Philip E. Agre, *Surveillance and Capture: Two Models of Privacy*, 10 Info. Soc'y. 101, 106-07 (1994).

conclusion that replacing a program is sometimes the only solution[12]: "Old programs do not learn, they simply fade away. So do human beings, their undebuggable programs replaced by younger, possibly less tangled, ones in other human heads." Simon, in this text, seems to qualify the human capability to learn as a potentially superior type of learning, as it does not require us to code our brain. He notes that we do not even have access to the way our brains operate and we don't need such access to actually learn: "But at least until the state of undebuggability is reached, human programs are modified adaptively and repeatedly by learning processes that don't require a knowledge of the internal representation."[13]

Of course, "capturing" natality in computational terms may reduce it to its representation and, as we should remind ourselves, a representation is not the same as what is represented. Though this also goes for natural language, human language can also *do what it describes*,[14] thus presenting rather than *re*presenting a shared world: "I declare you man and wife" is not a description but an act that institutes what it describes. Such "performative" speech acts, which can be oral as well as written, account for what lawyers call the legal effect of certain actions or occurrences, such as declaring two people to be married, accepting an offer to conclude a contract, or committing a tort, which result respectively in a legally valid marriage with all kinds of legal consequences, a binding contract, or an obligation to pay compensation. Such performative effects create the artificial — but very real — world that shapes our legitimate expectations and molds the institutional backbone of daily intercourse. Contract, ownership and tort depend on a performativity that cannot be understood in the computational terms of performance metrics or mathematical optimization. Indeed, our own learning processes hinge on a continuous iteration of the natality that defines us. Our institutional environment depends on the largely implicit alignment of individual consciousness with

[12] Simon, *supra* note 10, at 34. We must note that Simon wrote at a point in time when machine learning was unsuccessful, due to (1) a lack of training data and (2) a lack of the computing power that enables massive parallel processing and multilayered artificial neural networks. I am not sure whether this changes the point he makes here. See, for example, the scientist who "invented" backpropagation and deep learning, Geoffrey Hinton, who claims it is nowhere near "general intelligence" and does not in any way compare to human learning. Steve LeVine, *Artificial Intelligence Pioneer Says We Need to Start Over*, Axios (Sept. 15, 2017), https://www.axios.com/ai-pioneer-advocates-starting-over-2485537027.html.

[13] Simon, *supra* note 10, at 34.

[14] Philosophical Investigations, *supra* note 8; John R. Searle, Speech Acts: An Essay in the Philosophy of Language (1969).

societal norms that in turn depend on the individual consciousness of all those who form a society. This underscores the relational nature of individual consciousness; human beings are always in the process of — incrementally and/or radically — reinventing themselves and their shared world.

B. Incomputability

Having argued for a relational concept of privacy and the fundamental indeterminacy of human identity that it implies, I will now clarify what I mean by the incomputability of the self. In computer science, the term incomputability refers to a specific type of decidability, meaning that it is impossible to develop "a single algorithm that always leads to a correct yes-or-no answer."[15] This type of incomputability is related to Gödel's incompleteness theorem and keenly demonstrates that the formalization that is necessary to turn real life events into machine-readable data (including programs) necessarily results in uncertainty at the level of mathematical decidability (even if for all practical purposes many decidable problems can be framed, based on productive assumptions).[16] A similar point has been made more specifically for machine learning by David Wolpert, proving mathematically that no machine learning algorithm will necessarily provide optimized output on new data.[17]

My interest, however, concerns the preliminary question whether real life events can be formalized in the first place, which is a precondition for their computability (in the practical sense of making them available for processing by a digital computing system). This means that by computability I refer to the numerization, digitization or datafication (taken as synonyms) of objects, processes, states or events in the "real" world of atoms. The answer is as simple as it is pertinent: yes, we can translate "real" life events into machine-readable data and programs, but as with every translation, something

[15] Undecidable Problem, Wikipedia, https://en.wikipedia.org/w/index.php?title=Undecidable_problem&oldid=810547800 (last visited Nov. 18, 2017). For a more in-depth discussion, see Walter Dean, *Computational Complexity Theory*, *in* The Stanford Encyclopedia of Philosophy (Edward N. Zalta ed., 2015). *See also*, saliently, Giuseppe Dari Mattiacci, *Godel, Kaplow, Shavell: Consistency and Completeness in Social Decision-Making*, 79 Chic.-Kent L. Rev. 497, 514-15 (2004).

[16] Panu Raatikainen, *Gödel's Incompleteness Theorems*, *in* The Stanford Encyclopedia of Philosophy (Edward N. Zalta ed., 2018). For a constructivist approach, see Jean Paul Van Bendegem, *A Defense of Strict Finitism*, 7 Constructivist Found. 141 (2012).

[17] David H. Wolpert & William G. Macready, *No Free Lunch Theorems for Optimization*, 1 IEEE Transactions Evolutionary Computation 67 (1997).

gets lost. Above all, we must avoid mistaking the translation for what has been translated. In the end, atoms are not computable, though their datified representation is. We can move from atoms to bits, but at some point, we must return to atoms, because that is where we live, that is what matters and that is where the computational operations conducted on bits will make a difference. This clearly regards all aspects of the "real" world, and has two implications. First, only data are computable, meaning that physical reality is only computable after having been datafied; in itself it is not computable. Second, once datafied, the complexity of high dimensional hypotheses space introduces the undecidability and incomputability theorems that are inherent in mathematical complexity. Those who assume that n=all in machine learning,[18] are glancing over the inherent limitations of mathematics in the face of new data.[19] Though new data can be predicted based on historical data, new data cannot inform the prediction as no algorithm can be trained on future data. The temporality that grounds us and the natality it entails confronts machine learning with the fundamental uncertainty of the real world.

My concern is with the natality that is core to being human in an inherently uncertain world. Taking an ecological perspective, I am foremost interested in the institutional environment that shapes the self, as discussed above, and the technological backbone of institutions that protect the indeterminate dynamic self that emerges, grows, erupts and disrupts on the cusp of "me" and "I." This indeterminacy and the implied incomputability is not rooted in the translation from atoms to bits, or in the temporality that forms the abyss of unpredictability of the physical world.[20] It is rooted in the double contingency that erupts whenever I am addressed by another human being who addresses me as a grammatical first person, thus inviting me to change perspective — gazing back upon myself from the point of view of the other (thereby instituting both her grammatical second person as perceived by me and my grammatical first person as the person who is addressed by her). This also goes for being addressed by a group or an institution, for instance inviting me to identify with a people, a class, an employer, a government, a religion. My concern is that this particular first-person perspective cannot be formalized, or captured in terms of data or programs, because this would

[18] Viktor Mayer-Schönberger & Kenneth Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think (2013).

[19] For a salient presentation of the implications of Wolpert's NFL theorems, see No Free Lunch Theorems, http://www.no-free-lunch.org. (last visited Aug. 1, 2018).

[20] The Speculative Turn: Continental Materialism and Realism (Levi Bryant, Nick Srnicek & Graham Harman eds., 2011).

always result in a third-person (or *idem*) perspective: a "her," "him" or "it." Though the "me" is also born by taking a third-person perspective, it has a special status, as it always ties up with an "I." "Me" and "I" thus form the incomputable self (the *ipse*) that cannot be represented other than via the bypass of an objectified (third-person, *idem*) perspective. What matters is that this bypass is necessarily ephemeral, it requires hard work to stabilize and — in the end — remains underdetermined. This is core to our nonessentialist essence.[21]

C. From a Relational to an Ecological Understanding of Privacy

The incomputability of human identity is, however, not a bug but a feature. It does not entail that "I" and/or "me" form an immutable essence. On the contrary, "I" and "me" emerge from being addressed by "you" and "them," and *much depends on the information and communication infrastructure (ICI) that affords this address*.[22] Whereas this address is primarily shaped by natural language, which affords first-, second-, and third-person perspectives, the "technologies of the word" (writing, print and audio-visual renderings) reconfigure these initial affordances by broadening their scope, making it possible to reach a larger audience and a distantiation in time and space between author and reader, between author and text and, as a result thereof, between text and meaning. This has actually created the "interpretability problem" that is core to cultures that build on written and printed text.[23] Indeed, this

[21] Mattiacci saliently demonstrates why human decision-making thrives on antinomian criteria that defy logic-based decision-systems as either incomplete or inconsistent. However, in the first part of his article he — perhaps unintentionally — shows that defining policy choices in terms of the mutually exclusive criteria of welfare enhancement and fairness is a hazardous undertaking that rules out so many other relevant considerations and generates so many interpretation issues as to be both ridiculous and dangerous. Mattiacci, *supra* note 15. Antinomies must not be denied or overcome but appreciated and taken into account, see e.g. Radbruch's antinomian concept of law and Dworkin's integrity of law. Gustav Radbruch, *Legal Philosophy*, *in* The Legal Philosophies of Lask, Radbruch, and Dabin 44 (Harv. Univ. Press reprt. 2014 ed.) (1950); Ronald Dworkin, Law's Empire (1991).

[22] Previous information and communication infrastructures — those of orality, the script and the printing press — have other affordances then data-driven infrastructures. *See* Hildebrandt, *supra* note 7, at 159-85.

[23] Paul Ricoeur, Interpretation Theory: Discourse and the Surplus of *Meaning* (1976); Pierre Lévy, Les Technologies de L'intelligence. L'avenir de la Pensée à L'ère Informatique (1993). Ricoeur and Levy both describe the affordances of text in terms of the need for interpretation, which emerges when speech is

interpretability problem has extended both the double contingency that vouches for creative misunderstandings and — as discussed above — the inner world that develops while being addressed by conflicting voices and primed by "longreads" that require extensive sequential information processing (books are read from left to right, from top to bottom and from first to last page — or in any other sequence, but always sequentially).

To better understand what incomputability means in an era when anything — and potentially everything — is being datafied, numerized and rendered computable, we can draw on an important text by Philip Agre, on privacy and "capture."[24] In this text he observes that threats to privacy are often understood in terms of surveillance, embedded in visual metaphors, and rooted in the historical experience of state surveillance. Though he does not reject this surveillance model, he points out that the pervasive character of computational infrastructures requires another model to comprehend the novel threats to privacy. His capture model emphasizes the fact that data-driven systems reconfigure their environment to gain access to more data, turning both our environment *and ourselves* into data engines. This entails extensive tracking schemes to mine *e.g.* behavioral data, and it is tempting to frame this in terms of surveillance:

> Yet tracking schemes have another side: the practical arrangements through which the data are collected in the first place, including the arrangements that make human activities and physical processes trackable. As human activities become intertwined with the mechanisms of computerized tracking, the notion of human interactions with a "computer" — understood as a discrete, physically localized entity — begins to lose its force. In its place we encounter activity systems that are thoroughly integrated with distributed computational processes. It is this deeper implication of tracking that forms the central motivation for this paper.[25]

Agre thus highlights that the need of computational systems to "capture" data reorganizes their (and our) environment, thereby changing the affordances of their (and our) environment. The changes in the prevailing ICI are indeed

externalized, requiring sequential processing and iterant reinterpretation by e.g. new generations. Ronald Dworkin, *Law as Interpretation*, 60 Tex. Law Rev. 527 (1982) (emphasizing the role of legal text as requiring the integrity — rather than merely the consistency — of legal decision-making, which necessarily requires iterant interpretation).

[24] Agre, *supra* note 11.

[25] *Id.* at 105.

transforming their environment (which includes us): “Computationalists' discourse rarely brings to the surface the connotations of violence in the metaphor of capture; captured information is not spoken of as fleeing, escaping, or resenting its imprisonment.”[26] Agre notes a set of characteristics that distinguishes the capture model from the more familiar surveillance model. I believe these characteristics to be pertinent for a better understanding of how data-driven architectures transform the environment that we depend upon, while also transforming our selves in the process. Since we are an important asset within the environment of these ICIs, we must be reconfigured in ways that enable the capture of behavioral and other data “from” us.

The capture model entails the parsing and reconfiguration of human behavior in a way that fits the need for formalization, for instance by means of tracking of keystroke and clickstream behavior and so-called “like-behavior,” or sensor technologies that pick up on our physical states and behaviors. Agre highlights that the “[d]riving aims . . . . are not political but philosophical, as activity is reconstructed through assimilation to a transcendent (‘virtual’) order of mathematical formalism.”[27]

This brings me to McQuillan’s notion of “machinic neo-platonism” to better understand the impact that Agre refers to.[28] McQuillan starts by observing that at this point in time “data science” is not merely a method, but rather an “organizing idea.” He finds that:[29] “Data science does not only make possible a new way of knowing but acts directly on it; by converting predictions to pre-emptions, it becomes a machinic metaphysics.” In his pivotal article, McQuillan traces the trajectory of neo-platonism in the history of both mathematics and science, explaining it as “the belief in [a] hidden layer of reality which is ontologically superior, expressed mathematically and apprehended by going against direct experience.”[30] He emphasizes that other than in science, the cyber-physical infrastructures that are built on computational connectivity are not merely theoretical constructions, but physical machines that may have an “impact” in the most literal sense of that term (as in “trauma” resulting from “impact”). Due to the nature of computability, the upcoming Internet of Things (smart energy grids, smart cities), combined with robotics (connected cars), cloud-, fog- and edge-computing, may come to drive an overcomplete datafiction of anything and everything based on the idea that the mathematics

[26] *Id.* at 106.

[27] *Id.* at 107.

[28] Dan McQuillan, *Data Science as Machinic Neoplatonism*, 31 Phil. & Tech. 253 (2018).

[29] *Id.* at 253.

[30] *Id.* at 261.

that grounds all these machines reveals the ultimate layer of a hidden reality. Note the following statement made by Mark Zuckerberg: "I'm also curious about whether there is a fundamental mathematical law underlying human social relationships that governs the balance of who and what we all care about . . . . I bet there is."[31]

Though one may be inclined to "read" the actions of Mark Zuckerberg as focused on financial gain, (t)his statement may provide a deeper insight into his motivation to capture as much "social data" as possible. As indicated, my concern with the protection of the incomputable nature of the self, resides in the transformation of the choice architecture of our immediate environment that is based on a set of assumptions regarding the nature of reality, including the reality of human intercourse and the human self. There are elements of totalitarianism in some of these assumptions, notably where they prefer bits to atoms and mathematical theory to the reality we actually face. I believe it is urgent to develop the means to protect against such totalitarianism and the first step would be to understand privacy as the protection of the incomputable self — and the right to privacy as the effective and practical remedy to *protect what counts but cannot be counted*: the fragile but robust, indeterminate but sustainable, ecological and irreducibly subjective self. It would be, however, a mistake to believe that a relational understanding of privacy and identity suffices. Rather, we need to pay keen attention to the material, institutional and technological environment that enables and constrains the relationship between self and other. This requires an investigation into the affordances of different types of environments.[32] In the next Part we will investigate how machine learning is reconfiguring our environment, by inquiring into the key design choices it entails and the tradeoffs they imply.

II. Democracy and the Rule of Law: From Agnostic to Agonistic Machine Learning

A. Machine Learning, Bias and Purpose Limitation

In his handbook on machine learning, Tom Mitchell defines it as follows: "A computer program is said to learn from experience E with respect to some

[31] *Facebook's Zuckerberg Wants to Figure Out Social Equation*, Phys.org (July 1, 2015), https://phys.org/news/2015-07-facebook-zuckerberg-figure-social-equation.html.

[32] Mireille Hildebrandt, *Law As an Affordance: The Devil Is in the Vanishing Point(s)*, 4 Critical Analysis L. 116 (2017). Further references can be found there.

class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."[33]

He gives the example of learning to play checkers, defining the task as "playing checkers," the performance metric as "percentage of games won against other players" and the experience as "playing practice games against itself."[34] The research design for this "learner" aims to develop a so-called target function, a mathematical function that correctly determines the best next move to be made, choosing from the set of "legal" moves that are available. The best next move is defined as the move that — assuming a certain "board state" — has the highest probability of contributing to a sequence of moves that results in a win. There are different ways of developing such a target function. One could immediately target a function that selects the best next move. However, this may be very difficult due to the fact that there are many dependencies related to not knowing which move the other party will make and how this will reconfigure the best strategy for winning the game. It is easier to target a function that attributes a numerical value to all board states, thus ranking their relationship to winning the game. The "best move from any current board position" can then be selected "by generating the successor board state[s] produced by every legal move, then using V to choose the best successor state and therefore the best legal move."[35] As Mitchell explains, most often this ideal target function is not "efficiently computable." Instead, the system is trained to learn an "operational description of the ideal target function."[36] Even this may be very difficult in the end: "In fact, we often expect learning algorithms to acquire only some approximation to the target function, and for this reason the process of learning the target function is often called function approximation."[37]

The next step in the learning process is to find a way to represent the target function, in the form of a mathematical formula that contains both variables and numerical coefficients or weights that are adjusted by the system until they best approximate the ideal target function. We shall skip the algebra, but note that a number of choices have already been made to prepare the next step, that of choosing the "function approximation algorithm."[38] This

[33] Mitchell, Machine Learning, *supra* note 3; Mitchell, *Key Ideas, supra* note 3. *See also* Brooks, *supra* note 1.

[34] *See* Mitchell, *Key Ideas*, *supra* note 3, at 10 (discussing the latest advances in these types of games under the heading of "[d]istant rewards and reinforcement learning.").

[35] Mitchell, Machine Learning, *supra* note 3, at 7.

[36] *Id.* at 8.

[37] *Id.* at 8.

[38] *Id.* at 9.

involves selecting relevant training examples (in this case subsequent board states), taking into account that the only information the system has about whether a move is right or not (or better than other moves) is the outcome of the game. Assigning weights to board states is not obvious, because "even if the program loses the game, it may still be the case that board states occurring early in the game should be rated very highly and that the cause of the loss was a subsequent poor move."[39]

Mitchell describes the final step in designing the checkers learning system as consisting of the design of four distinct program modules that actually characterize many learning systems:[40] (1) *the performance system*, which tests the accuracy of different weights in terms of their contribution to winning the game; (2) *the critic*, which produces a set of training examples of the target function, showing which actual sequences of moves result in winning or losing the game; (3) *the generalizer*, which "generalizes from the specific training examples, hypothesizing a general function that covers these examples and other cases beyond the training examples,"[41] where the hypothesis takes the form of a mathematical function (aiming to approximate the ideal target function); and (4) *the experiment generalizer*, which "takes as input the current hypothesis (currently learned function) and outputs a new problem (*i.e.*, initial board state) for the Performance System to explore. Its role is to pick new practice problems that will maximize the learning rate of the overall system."[42] Together, these four modules should enable a series of iterant "runs" that form the learning process and result in an increasingly accurate prediction of which next move will contribute to winning the game.

The *generalizer* is in many ways the core module, as it hopes to approximate a function that predicts not only the outcome of historical games (the training and validation data), but rather the outcome of out-of-sample games (test data). This is made possible by means of an iterant combination of the *critic* that provides training examples and the *experiment generalizer* that tests how well the current hypothesis function predicts "best moves" that result in winning the game. We can identify three types of experience and thus three types of datasets here. First, the training set, which is used to train the algorithm towards attributing the best distribution of weights within the mathematical function that has been developed to approximate the (unknown) target function. Second, the validation set, which is used to check how well the hypothesis target function does in predicting moves that result in winning

[39] *Id.* at 10.

[40] *Id.* at 11-13.

[41] *Id.* at 12.

[42] *Id.*

a game. Usually, the available data are split into *e.g.* 80% training data and 20% that is kept apart for validation. Finally, we have the test set, which consists of data not yet available when the system was trained. The first and second datasets are historical (or streaming) data, while the test set refers to future data — which, however, turn into historical or streaming data once available. In the case of highly dynamic and continuous learning systems, the validation set and the test set seem to become conflated. We must, however, remember that no system can be trained on future data — as soon as it tests its hypothesis against new data, these data will be historical data.

The fact that systems cannot be trained on future data may sound trivial, but it is actually core to both the potential and the limitations of machine learning. This is related to the uncertainty that is inherent in anticipating the future, which is less of a problem in a game ruled by unambiguous rules and a limited set of options on both sides, than in a real life situation.[43] The goal of machine learning is training algorithms to uncover patterns that do not merely fit the training data, but also fit future data. It has been mathematically proven that whereas algorithms can be optimized to learn specified tasks, this never implies that the optimization works on new data or with a view to another task.[44] This is related to the fact that the target function (that would correctly describe the mathematical relationship between input and output data) is *necessarily* unknown, because we *necessarily* only have historical data. In real life situations (other than in games with a limited set of immutable rules), future data can always disrupt the predictive accuracy of the hypothesis target function. This need not be a problem, depending on the type of application we wish to develop. If, however, damage or harm results from the wrongful assumption that a particular hypothesis target function is accurate, or if harm, damage and benefits are wrongly (re)distributed when applying the algorithm, the opacity of the machine learning system does become a problem, as it may not be easy to redress — or even address — the problem. This is particularly alarming once we acknowledge that high accuracy with regard to the training and validation data does not imply that the algorithm is getting things right for the real world, as has been demonstrated repeatedly. For example, an algorithm trained on patient data to help physicians decide whether or not to send a patient with pneumonia to a hospital, found 3 indicators for low

[43] *Cf.* Collados, *supra* note 2.

[44] David H. Wolpert, *What the No Free Lunch Theorems Really Mean; How to Improve Search Algorithms* (Santa Fe Inst. Working Paper No. 2012-10-017, 2012), https://sfi-edu.s3.amazonaws.com/sfi-edu/production/uploads/sfi-com/dev/uploads/filer/33/44/33440e97-fe46-4827-a1eb-a27196e1c49a/12-10-017.pdf.; *Cf.* Mitchell, *Key Ideas*, *supra* note 3, at 4-5.

risk: chest pain, asthma, and recent heart problems. Though the algorithm was found to perform with high accuracy (on the data), medical practitioners immediately disqualified it as getting things completely wrong. In fact, these are 3 indicators of high risk. Because patients with such indicators are routinely hospitalized, their morbidity may actually be lower than that of other patients, who are not necessarily sent to the hospital. The point of this example is twofold. First, it is not always obvious that a system with high accuracy on the data may be getting things wrong nevertheless. Second, if a deep learning algorithm is used, the indicators will be hidden in the black hole of the learning system, which will probably provide a score to each patient, without providing an explanation in terms of identifiable indicators. Trusting systems such as these and coming to depend on them, will have consequences for critical infrastructure and could easily cause harm and damage. Another example concerns the COMPAS software, which supports courts in making decisions on parole and sentencing. COMPAS provides a risk score that refers to the risk that a person will recidivize. Because the training data was biased in the sense that black persons recidivized more often than white persons, the system ended up with a disparate error for those who did not recidivize. Of those who did not recidivize, black persons were more often wrongly classified as high risk, whereas white persons were more often wrongly classified as a low risk. Here, the point is that this output is the result of a choice of the research design, which was not tasked with a correction of the risk score for those who do not recidivize, to prevent unfair bias. Developing such a correction is not obvious, however, as it depends on expertise in machine learning and a willingness to detect and foresee the statistical implications of the fact that based on the training data black persons recidivized more often than white people. Again, trusting such a — proprietary — software system, and outsourcing decisions on deprivation of liberty to such systems, would be highly problematic for those whose libery is at stake.[45]

Mitchell explains that machine learning can be seen as involving "searching a very large space of possible hypotheses to determine one that best fits the observed data and any prior knowledge held by the learner."[46] This means

[45] Rich Caruana et al., *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission*, 21 Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery & Data Mining 1721 (2015); Julia Angwin et al., *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks*, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Chouldechova, *supra* note 1.

[46] Mitchell, Machine Learning, *supra* note 3, at 14.

seeing machine learning as a search problem, where success is determined by choices made as to search strategies, the structure of the search space, and the relationship between the size of the hypothesis space, the number of training data and the confidence we can have with regard to generalizing to new data. Mitchell sums up a series of methodological issues that are inherent in any machine learning research design:

> What algorithms exist for learning general target functions from specific training examples? In what settings will particular algorithms converge to the desired function, given sufficient training data? Which algorithms perform best for which types of problems and representations? How much training data is sufficient? What general bounds can be found to relate the confidence in learned hypotheses to the amount of training experience and the character of the learner's hypothesis space? When and how can prior knowledge held by the learner guide the process of generalizing from examples? Can prior knowledge be helpful even when it is only approximately correct? What is the best strategy for choosing a useful next training experience, and how does the choice of this strategy alter the complexity of the learning problem? What is the best way to reduce the learning task to one or more function approximation problems? Put another way, what specific functions should the system attempt to learn? Can this process itself be automated? How can the learner automatically alter its representation to improve its ability to represent and learn the target function?[47]

The issues targeted by Mitchell are not obvious to those outside the domain of machine learning. On the one hand, many people seem to believe that machine learning is agnostic, in the sense of being oblivious to human bias or independent of the design choices that determine its performance accuracy. In that sense, however, machine learning is not agnostic. On the other hand, many people seem to believe that undesirable bias in the training data can be remedied in a straightforward way, thus restoring some kind of neutral training set, resulting in agnostic machine learning. Much has been written on this front, both on the side of computer science (e.g., discrimination-aware data mining),[48] and on the side of law (following up on claims of racial bias in

[47] *Id.* at 15.

[48] *See, e.g.*, Dino Pedreshi, Salvatore Ruggieri & Franco Turini, *Discrimination-aware data mining*, 14 Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery & Data Mining 560 (2008); Salvatore Ruggieri, Dino Pedreschi & Franco Turini, *Data Mining for Discrimination Discovery*, 4 ACM Transactions Knowledge

software used to assess risks of recidivism).[49] With regard to discriminatory bias, we must note such bias cannot easily be remedied, because protected attributes often correlate with other — quasi-innocent — attributes that will operate as proxies.[50] In that case the solution does not result in "objective" data but in making discrimination less visible. Another caveat concerns the fact that translating legal or ethical notions of fairness into machine learning research design is not at all obvious, also because different notions of fairness may be incompatible.[51]

In this Article I will not move into these issues. Instead, I will probe the assumptions and implications that inform any machine learning research design, clarifying that issues of bias are inherent in machine learning and must be understood at the level of its methodological integrity. The issues summed up by Mitchell above inform such methodological integrity of the solutions presented in real-life applications. They also crucially help to understand that machine learning research designs involve a number of tradeoffs between *e.g.* speed, predictive accuracy, overfitting (low utility) or overgeneralizing (blind spots), confirming that each choice amongst competing strategies has a cost: there is no free lunch as to the research design for machine learning. These tradeoffs relate to what Mitchell calls the inductive bias that is inherent in any machine learning research design. Under the heading of "[t]he futility of bias free learning," he explains bias as a "fundamental property of inductive inference: *a learner that makes no a priori assumptions regarding the identity of the target concept has no rational basis for classifying any unseen instances* (my emphasis)."[52] The inductive bias refers to the fact that in order to come up with an approximation of the ideal target function, we have to make a number of design choices as well as assumptions — without which we cannot even begin to train an algorithm: "Thus, we define the inductive bias of a learner as the set of additional assumptions B sufficient to justify its inductive inferences as deductive inferences."[53]

Discovery from Data 1 (2010); Mitchell, Machine Learning, *supra* note 3, at 39-45.

[49] *See, e.g.*, Angwin et al., *supra* note 45; Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Calif. Law Rev. 671 (2016).

[50] Moritz Hardt, Eric Price & Nathan Srebro, *Equality of Opportunity in Supervised Learning*, 16 Proc. Int'l Conf. Neural Info. Processing Systems 3323 (2016).

[51] Chouldechova, *supra* note 1.

[52] Mitchell, Machine Learning, *supra* note 3, at 42.

[53] *Id.* at 43. *See also* Mitchell, *Key Ideas*, *supra* note 3, at 5 (framing bias as a potential error):

> This can occur when the learner's hypothesis space H is insufficient to represent every function that can be labeled over X, or alternatively even

Interestingly, this accords with key insights from the tradition of philosophical hermeneutics. In his seminal work on *Truth and Method*,[54] Gadamer basically said the same thing when explaining that we need some form of prejudice to even begin to understand whatever it is we face. His point is that prejudice (or bias) is not necessarily a bad thing (even if Enlightenment thinking introduced a negative connotation for prejudice),[55] depending on the extent to which one is willing to question one's own prejudices. Without the acknowledgment that no understanding is possible without an initial bias, it becomes very difficult to distinguish between a bias that fits the object of understanding and a bias that does not. In the case of machine learning, the issue is further aggravated by the fact that the bias is not merely in the design of the hypothesis space, but also in the choice of the training data, which may contain an undesirable bias with regard to the issue that is at stake. Note that training data is necessarily biased, as it is the bias that allows algorithms to detect and confirm patterns. The task of a proper research design is to uncover whether the bias is a computational artefact (bug) in the dataset, or a pattern in the world of atoms and meaning (that is, the world outside the dataset, about which the data supposedly provides nontrivial and relevant information). At the same time, the research design should contribute to detecting alternative explanations of the same bias, by exploring (1) the complexities of the underlying causalities,[56]

> if H is sufficiently expressive but the learner has some preference (bias) for choosing between two hypotheses that perform equally over the training data (e.g., a preference for short decision trees).

Basically, this is another use of the term bias, which highlights the critical impact of how the hypothesis space is designed — confirming that as design choices are made, one of the tradeoffs may be an incorrect bias. Note that this is not about the bias that may be inherent in the training data, which is the bias most commonly referred to in the literature on automated bias. *See* Harry Surden, *Machine Learning and Law*, 89 Wash. Law Rev. 87, 106 (2014); Barocas & Selbst, *supra* note 49.

[54] Hans-Georg Gadamer, Truth and Method (2004).

[55] *Id.* at 273.

[56] In machine learning the relationship between correlations and causality is an important subdomain. Let's not get carried away by the superficial dreams of Chris Anderson in his *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, Wired Mag., June 23, 2008, at 16. *See, e.g.*, Judea Pearl, Causality: Models, Reasoning and Inference (2d ed. 2009); Judea Pearl & Dana Mackenzie, The Book of Why: The New Science of Cause and Effect (2018) (Pearl's recent adaptation for a broader audience). *See also* Hildebrandt, *supra* note 7, at 37-40; Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 Harv. J.L. & Tech. 841, 845 (2018) (though not concerned

and (2) dependencies between the concepts that define the variables in the target function.[57]

Before developing the notion of agonistic machine learning in the next section, I will sum up the main conclusions from this brief discussion of machine learning. Let us note, first, that the underlying assumption of any machine learning exercise is the existence of an ideal target function that determines the relationship between input data (*e.g.*, the board state) and output data (winning or losing the game). In respect to a game of checkers, this assumption may hold, but once we move from checkers to human behaviors that are not constrained by a set of unambiguous rules this assumption is simply wrong. This is directly related to the idea of an incomputable self and the productive uncertainty generated by our double contingency. Let us note, second, that even for a relatively simple game such as chess, machine learning has to accept an operational approximation of an assumed target function, rather than the target function itself. The many choices made in the course of developing the research design also highlight that one can develop and test various alternative hypothesis target functions, when trying to approximate an ideal target function (that may not really "exist" anyway). This is actually good news, as it highlights the pragmatic, rather than the essentialist or neo-Platonic nature of machine learning as a discipline. Let us note, third, that any machine learning operation requires the determination of a task, a performance metric and a type of experience. Without defining a purpose for them, machines cannot learn. This is not to suggest that the concept of a purpose in data protection law is equivalent with the concept of a task in machine learning. But they are connected because the methodological integrity of machine learning requires keen attention to the research design in terms of the purpose of the exercise, as this should inform the machine-readable articulation of task, performance metric and training data, including the design of the hypothesis space and the choice of the test data. Indeed, Van der Lei made this point several years ago in the realm of medical informatics, underlining the importance of avoiding "low hanging fruit" when working with patient data. He formulated the first law of informatics: "Data shall be used only for the purpose for which they

with causality, distinguishing between explanation as "an attempt to convey the internal state or logic of an algorithm that leads to a decision" and counterfactual explanations that "describe a dependency on the external facts that led to that decision").

[57] To calculate, we must first qualify specific events or states as the same events. In that sense qualification always — even if not explicitly — precedes quantification, ranking and computation. Michel Callon & John Law, *On Qualculation, Agency and Otherness*, 23 Env't. & Plan. D: Soc'y & Space 717 (2005).

were collected. And the collateral: If no purpose was defined prior to the collection of data, then the data should not be used."[58]

Van der Lei was not concerned with data protection law when developing this law of informatics. His interest was the validity, relevance and accuracy of inferences made from medical data. Instead of getting rid of the principle of "purpose binding" in data protection law, we should acknowledge that to a large extent the methodological integrity of machine learning requires advance specification of the purpose, as this will inform the solidity and productivity of the relevant research design. Machines cannot learn if we do not define for them what qualifies as an improvement of their performance; this is only possible if we are clear about the purpose of processing.

B. From Agnostic to Agonistic Machine Learning

In data-driven environments, the choice architecture we face is determined by inferences made from behavioral data. Neither the data nor the inferences need to be personal data to have a major impact on the choice architecture we confront. As indicated above, the capture of these data requires a profound reconfiguration of the environment of data-driven systems, so as to keep a steady flow of data available for training, validating and testing their algorithms. Admittedly, to a large extent *we are* the environment that is being reconfigured, and this entails that our behaviors will be reconfigured as well — to fit the computational modules that inform data-driven applications. This creates a tension with the need to protect the incomputable nature of the human self, its foundational indeterminacy and the natality it expresses, precisely because the self develops in relation to the world it inhabits. Overdependence on computational decision-systems may result in a shrinking of the inner self, as we learn to internalize the logic of computational feedback to better adapt to our new environment. The elasticity, ex-centricity and ecological nature of the inner mind are what makes us human, but thereby also vulnerable to being hacked by an environment that is conducive to cognitive automation. This shrinkage of the inner self is highly problematic, not merely from the

[58] J. van der Lei, *Use and Abuse of Computer-Stored Medical Records*, 30 Methods Info. Med. 79 (1991). *See also* Simon de Lusignan & Chris Mimnagh, *Breaking the First Law of Informatics: The Quality and Outcomes Framework (QOF) in the Dock*, 14 J. Innovation Health Informatics 153 (2006); Federico Cabitza, Davide Ciucci & Raffaele Rasoini, *A Giant with Feet of Clay: On the Validity of the Data that Feed Machine Learning in Medicine, in* 28 Lecture Notes in Information Systems and Organisation, Organizing for the Digital World 113 (Federico Cabitza, Carlo Batini & Massimo Magni eds., 2017).

perspective of privacy as a private interest, but also for privacy as a public good, and notably for the substance of intellectual privacy that is closely related to the freedom of information, and to the capability to develop a mind of one's own regarding matters of personal and public interest.

However, instead of rejecting or obstructing machine learning per se, I propose to acknowledge that though we can be "made" computable, (1) this does not imply that such computability entirely defines us, and (2) there are always many — and sometimes radically different — ways of computing the same person.[59] In line with the latter, one way of protecting our privacy is to require what I call "agonistic machine learning," *i.e.*, demanding that companies or governments that base decisions on machine learning must explore and enable alternative ways of datafying and modelling the same event, person or action.[60] This should ward off monopolistic claims about the "true" or the "real" representation of human beings, their actions and the rest of the universe in terms of data and their inferences. It requires us to move from agnostic to agonistic machine learning, from assuming that machine learning will get it right because of these systems' aura of numerical objectivity to testing whether the bias they detect in their training set makes good sense or is incorrect, unfair or spurious. Many applications of machine learning actually work with a so-called "ground truth" to anchor the performance metric; to test whether the system gets it right, machine learning will often require a machine-readable indication of what is "right."[61] The ground truth is, for instance, based on surveys or interviews where people are asked to assess their own position, emotions, or preferences or, alternatively, based on expert opinion such as medical diagnoses made by medical doctors. This implies that some parameter is introduced as the "real" truth, it being taken for granted that whenever the system aligns with this ground truth it is getting things right. This has three implications.

[59] *Cf.* Mireille Hildebrandt, *Profiles and Correlatable Humans*, *in* Who Owns Knowledge? Knowledge and the Law 265 (Nico Stehr & Bernd Weiler eds., 2008).

[60] Kate Crawford asks the preliminary question in her *Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics*, 41 Sci. Tech. & Hum. Values 77 (2016). My answer is that algorithms can be designed in agonistic ways, but it is not obvious that machinic decisions and the algorithms that "make" them can be agonistic. This raises another preliminary question as to whether we should want to employ machine learning in the first place. I believe that agonistic machine learning will contribute to informed answers to this question — where concrete applications are considered.

[61] This is notably the case for reinforcement learning. *See* Brooks, *supra* note 1.

First, this type of machine learning or cognitive computing is parasitizing on human domain expertise or simply on human experience. These systems are not developing a proper understanding of law or medicine or accounting, but more or less excellent *simulations* of such expertise or experience. This has unprecedented consequences, both positive (offloading tasks to cognitive machines) and negative (deskilling of the experts whose knowledge has been offloaded).[62] Second, the better the simulation, the higher the risk that it will follow the bias that is hidden in the ground truth (which may not be so true after all).[63] Third, the "ground truth" itself is often contestable and indeed contested, as medical doctors disagree about diagnoses and individuals provide incorrect answers or change their mind.[64]

Agonistic machine learning would bring the adversarial core of the Rule of Law into the heart of the design of data-driven environments, thus also aligning with the methodological core of reliable machine learning. Taking democracy, the Rule of Law and scientific method seriously, we should require that the research design of our supposedly smart architectures be based on *agonistic debate*, *built-in falsifiability* and a *robust constructive distrust*. This should result in *testable* and *contestable* decision-systems whose human overlords can be called to account, squarely facing the legal interpretability problem and its relationship with the computer science interpretability problem.[65]

[62] Federico Cabitza, Breeding Electric Zebras in the Fields of Medicine (Jan. 27, 2017) (unpublished manuscript), http://arxiv.org/abs/1701.04077; NICHOLAS CARR, THE GLASS CAGE: AUTOMATION AND US (2014).

[63] Cathy O'Neil, *False Negatives Can Be a Matter of Life and Death*, BLOOMBERG (Nov. 30, 2017), https://www.bloomberg.com/view/articles/2017-11-30/false-negatives-can-be-a-matter-of-life-and-death.

[64] Cabitza, Ciucci & Rasoini, *supra* note 58. Jana Diesner, *Small Decisions with Big Impact on Data Analytics*, 2 BIG DATA & SOC'Y 1 (2015).

[65] Mireille Hildebrandt, *The Dawn of a Critical Transparency Right for the Profiling Era*, *in* DIGITAL ENLIGHTENMENT YEARBOOK 2012, at 41 (Jacques Bus ed., 2012); Bryce Goodman & Seth Flaxman, *European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation,"* 38 AI MAGAZINE 50 (2017); Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76 (2017); Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a "Right to Explanation" is Probably Not the Remedy You are Looking for*, 16 DUKE L. & TECH. REV. 18 (2017); Gianclaudio Malgieri & Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 243 (2017); Andrew Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 INT'L DATA PRIVACY L. 233 (2017); Margot E. Kaminski, *The Right to Explanation, Explained* (Univ. Colo. Law Legal Studies Research Paper No.

Whereas the Rule of Law is aligned with the notion of adversarial procedure,[66] I prefer the notion of agonistic for two reasons. The first is that the concept of "adversarial machine learning" has already been "taken" by security researchers, referring to the use of machine learning to defend against attacks against information systems.[67] The second reason is that the concept of agonism has been developed both within democratic theory,[68] and in the context of constructive technology assessment,[69] providing salient arguments for what DiSalvo calls "a condition of forever looping contestation."[70] DiSalvo actually developed a concept of "adversarial design" to designate the integration of agonistic pluralism into design practices in the broad sense of that term, including engineering, architecture, institutionalization, art and other forms of devising "courses of action aimed at changing existing situations into preferred ones" (quoting Herbert Simon, one of the founding fathers of artificial intelligence).[71] Disalvo argues that

> The ongoing disagreement and confrontation are not detrimental to the endeavour of democracy but are productive of the democratic condition . . . . From an agonistic perspective, democracy is a situation in

18-24, 2018), https://papers.ssrn.com/abstract=3196985; Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, 3 Big Data & Soc'y 1 (2016); Wachter, Mittelstadt & Russell, *supra* note 56; Frank Pasquale, *Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society*, 78 Ohio St. L.J. 1243 (2017).

[66] Jeremy Waldron, *The Rule of Law and the Importance of Procedure* (N.Y.U. Pub. Law & Legal Theory Research Paper Series, Working Paper No. 10-73, 2010), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1688491.

[67] J. D. Tygar, *Adversarial Machine Learning*, 15 IEEE Internet Computing 4 (2011). In that context, "adversarial perturbations" have been developed as a means of obfuscation; such perturbations can also be used to detect "counterfactual explanations," as a means to improve actionable explainability of automated decisions. *Cf.* Wachter, Mittelstadt & Russell, *supra* note 56, at 11-12.

[68] Chantal Mouffe, *Deliberative Democracy or Agonistic Pluralism?*, 66 Soc. Res. 745 (1999).

[69] Arie Rip, *Constructing Expertise: In a Third Wave of Science Studies?*, 33 Soc. Stud. Sci. 419 (2003). On the aggregative, deliberative and participatory dimensions of democratic theory, see Mireille Hildebrandt & Serge Gutwirth, *(Re)presentation: pTA Citizens' Juries and the Jury Trial*, 3 Utrecht L. Rev. 24 (2007).

[70] Carl DiSalvo, Adversarial Design 5 (2d ed. 2015).

[71] *Id.* at x (quoting Herbert A. Simon, The Sciences of the Artificial 111 (3d ed. 1996)).

> which the facts, beliefs, and practices of a society are forever examined and challenged . . . . Perhaps the most basic purpose of adversarial design is to make . . . . spaces of confrontation and provide resources and opportunities for others to participate in contestation.[72]

Such a loop of contestation should not be confused with post-truth postmodernist relativism, but aligned with the idea of an open society that is willing to face the potential falsification of mainstream assumptions,[73] and is capable of doubting *anything* if good reason does arise, even though one cannot doubt *everything*.[74] This is not a matter of antagonism (being against anything, whatever the reason) but a matter of co-designing our material, face-to-face and institutional environment in a robust, sustainable way that affords equal respect and concern. Taking democracy seriously means that whenever technologies that could reconfigure our environment, are developed, marketed and employed, we must make sure that those who will suffer or enjoy the consequences are heard and their points of view taken into account.[75] Not merely to be nice, but because they will bring specific expertise to the table and contribute to achieving "robust" societal architectures. Discussing the preconditions for "constructive technology assessment," Rip qualifies agonism in terms of learning processes: "These are processes of agonistic, collective learning, which hopefully lead to robust outcomes."[76] A productive arrangement, then, is one that is conducive to agonistic learning and robust outcomes.

Agonistic learning affords robust as well as fair outcomes, based on serious consideration of potential objections and alternative designs, while steering free of untested assumptions that are prone to generating vulnerabilities. Agonism will also protect against the surge of self-driving technologies that feed on overdetermination of human action in terms of machine-readable behaviors, combining machine learning with hyper-nudging,[77] reducing human individuals to pawns in a game of chess played by the overlords of seemingly omnipotent platforms.[78] The idea is not to reject or denounce

[72] *Id.* at 5.

[73] Karl R. Popper, The Open Society and Its Enemies 111 (Princeton University Press 2013) (1945).

[74] Hilary Putnam, Pragmatism: An Open Question 21 (1995).

[75] John Dewey, The Public and Its Problems (1927).

[76] Rip, *supra* note 69, at 425. *See also* Dan McQuillan, *People's Councils for Ethical Machine Learning*, 4 Soc. Media + Soc'y 1 (2018).

[77] Karen Yeung, *'Hypernudge': Big Data as a Mode of Regulation by Design*, 20 Info. Comm. & Soc'y 118 (2017).

[78] Nick Srnicek, Platform Capitalism (2016).

machine learning but to contribute to reliable and testable research designs, e.g., such as proposed by Hofman, Sharma and Watts in their succinct but seminal article on "prediction and interpretation," where they discriminate between "exploratory" and "confirmatory" analysis:

> In exploratory analyses, researchers are free to study different tasks, fit multiple models, try various exclusion rules, and test on multiple performance metrics. When reporting their findings, however, they should transparently declare their full sequence of design choices to avoid creating a false impression of having confirmed a hypothesis rather than simply having generated one. Relatedly, they should report performance in terms of multiple metrics to avoid creating a false appearance of accuracy.
>
> To qualify research as confirmatory, however, researchers should be required to preregister their research designs, including data preprocessing choices, model specifications, evaluation metrics, and out-of-sample predictions, in a public forum such as the Open Science Framework (https://osf.io). Although strict adherence to these guidelines may not always be possible, following them would dramatically improve the reliability and robustness of results, as well as facilitating comparisons across studies.[79]

One is reminded of the rules on prior disclosure in medical science, which should prevent the hiding of trials with unfavorable results, the tweaking of performance metrics to upgrade the findings, or the use of unreliable, incomplete or irrelevant data.[80] Hofman, Sharma and Watts describe the methodological heart of what I mean by agonistic machine learning, though I would like to emphasize the need to highlight the tradeoffs — discussed above — that are implicated in any machine learning research design. Also, agonistic machine learning agonistic machine learning responds to the need to call out the ethical and political implications of *who decides task T*, *performance metric P* and *experience E*, and to investigate *how this is done*, taking into account *which (and whose) concerns are at stake*.

[79] Jake M. Hofman, Amit Sharma & Duncan J. Watts, *Prediction and Explanation in Social Systems*, 355 Sci. 486 (2017).

[80] Chris Chambers, *Clinical Trials Revolution Could Change the Future of Medical Research*, The Guardian (Aug. 24, 2017), https://www.theguardian.com/science/head-quarters/2017/aug/24/clinical-trials-revolution-could-change-the-future-of-medical-research.

III. Legal Protection and Agonistic Machine Learning

This, finally, brings me to the legal correlate of the call for agonistic machine learning. Should data scientists be forced to adhere to legal standards that safeguard the integrity of scientific method or should this be left in the hands of the scientists? I certainly believe the latter: science and scientists thrive when independent. However, two caveats apply. First, law should contribute to safeguarding this independence, which may be diluted when external funding of scientific research gains too much nudging power over those hoping for follow-up funding.[81] Second, if we are speaking of machine learning precepts that are applied outside the laboratory of academic computer science, a different regime should apply. If applications are put on the market and/or employed in the context of government, commercial or nonprofit organizations, the real-world consequences must be faced. Machine learning applications in the real world of atoms, people and institutions require a legal framework to ensure the testability and contestability of cyber-physical systems that reconfigure both us and our world.

Which domain of law might best contribute to adversarial design and agonistic construction of smart cities, smart energy grids, smart policing, connected cars, data-driven insurance, tax fraud detection, and the pervasive cyber-physical infrastructure that is being developed as we speak? As indicated previously,[82] I do not believe that law will solve the problems generated by data-driven agency all by itself; nor do I believe that data protection law is a panacea. I do believe, nevertheless, that we need to make sure that the rules of the game create a level playing field and afford effective protection of *privacy as the protection of the incomputable self.* This necessitates a smart legal architecture, consisting of private law (tort law, consumer protection, competition law), public law (fundamental rights, data protection law) and criminal law (enforcement of gross violations of individual human dignity

[81] Adam Rogers, *Google's Academic Influence Campaign: It's Complicated*, Wired (July 14, 2017), https://www.wired.com/story/googles-academic-influence-campaign-its-complicated/; Brody Mullins & Jack Nicas, *Paying Professors: Inside Google's Academic Influence Campaign*, Wall Street J. (July 14, 2017), https://www.wsj.com/articles/paying-professors-inside-googles-academic-influence-campaign-1499785286. External funding that raises issues concerning independence is not limited to commercial enterprise, as university departments that are dependent on government assignments raise similar questions. *Cf.* Chris Jones, Market Forces: The Development of the EU Security-Industrial Complex (2017).

[82] Hildebrandt, *supra* note 7, at 17.

and other human rights). I will end with two contested elements from data protection law (purpose limitation and profile transparency), briefly arguing that they align with robust machine learning practices, and ending with a call to properly distinguish between explanation and justification in the case of automated decision-making based on profiling.

As to the legal principle of purpose limitation,[83] as indicated above, defining the purpose of processing is part and parcel of any machine learning research design. It is defined in the delineation of task T, in the choice and curation of the training and validation data (experience E), and more precisely in the selection of the performance metric P. In light of Hofman, Sharma and Watts' distinction between exploratory and confirmatory research designs and Van der Lei's first law of informatics, we must conclude that data collected without a purpose or for another purpose is not fit for a proper research design. Using such data may result in shoddy output and insofar as the research design has not been registered the results cannot be validated, leading to untrustworthy machine learning practices. Nevertheless, the purpose that must be defined by the data controller is not the same as the task that is defined by the data scientists; the first defines the purpose *of the controller*, the second defines a purpose *for the learning algorithm*. The purpose of the controller must be specified in a way that makes sense considering her relationship with the data subject, while the task for the learning algorithm must be specified by way of formalization. The difference is, however, not a bug but a feature. By requiring the specification of one or more legitimate purposes by the controller, data

[83] *Council Regulation 2016/679,* 2016 O.J. (L 119) 1 (EU) [hereinafter GDPR] (stipulating that personal data may only be processed for specified, explicit and legitimate purposes); *id.* at art. 5.1(c) (adding that they shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed); *id.* at art. 6.1(a) (stipulating that consent can only be provided for one or more specific purposes); *id.* at arts. 5.1(b), 89 (stipulating that further processing for another purpose that is incompatible with the original purpose is only allowed if consent is provided for the new purpose, or if based on law which is necessary and proportionate); *id.* at recital 33 (presenting crucial exceptions for further processing for purposes of scientific research, which is considered to be compatible with the original purpose by default, though safeguards apply). Note that purpose plays a crucial and central role in the GDPR: it does not merely determine whether data may be processed (based on a necessity criterion), but simultaneously determines who is responsible and liable for compliance (see, for example, *id.* at art. 5.2). *See Opinion of the Article 29 Data Protection Working Party on Purpose Limitation*, WP 203, 00569/13 (2013). *See, e.g.*, Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos, 2014 EUR-Lex CELEX 62012CJ0131 (May 13, 2014).

protection law unwittingly contributes to sustainable research designs that have a better chance of making good sense of the data than sloppy exploratory research that covers its tracks in the name of experimentation and the freedom to tinker. The latter may be fine, but not with someone else's personal data and not if the unreliable results have an impact on people's lives in the real world (think of invisible discriminatory targeting based on credit rating, sentencing or employability algorithms; disproportional monitoring based on questionable fraud detection algorithms; or critical infrastructure that breaks down or wastes resources). Purpose specification is not *equivalent to* but certainly *aligns with* Van der Lei's First Law of Informatics, rejecting overstated claims of a tradeoff between predictive accuracy and interpretability in machine learning applications. Obviously, purpose limitation in constitutional, administrative and data protection law was not invented to serve the methodological integrity of data science — it derives from the legality principle that applies to the exercise of government competences and is meant to provide legitimacy and transparency to decision-making by powerful actors. Purpose limitation relates to the justification of such decision-making rather than its explanation in the sense of its heuristics.[84]

This brings me to the second element of EU data protection law that will contribute to a more robust application of machine learning: the extension of profile-transparency rights. These rights do not aim to justify machine decisions but to clarify how they came about (their heuristics), and thus raise the issue of their computational and human interpretability. Though tradeoffs may exist between predictive accuracy of the output and the interpretability of the underlying process, these tradeoffs necessarily depend on access to "ground truth" which should then be uncontroversial (otherwise the accuracy cannot be determined).[85] The more important applications should be based on confirmatory research that includes inquiry into causality, so as to prevent delusional inferences that are wrongly taken for granted precisely because there is no understanding of the causal dependencies on potentially unknown

[84] On the history of purpose limitation and its relationship with the legality principle, see for example, Maximilian von Grafenstein, The Principle of Purpose Limitation in Data Protection Laws (2018); E. R. Brouwer, *Legality and Data Protection Law: The Forgotten Purpose of Purpose Limitation*, *in* The Eclipse of Legality Principle in the European Union 273 (Leonard Besselink, Frans Pennings, & Sacha Prechal eds., 2011). Note that in data protection law, the principle applies equally to nongovernmental data controllers, thus protecting the opacity of individual persons against any big players. *See also* Hildebrandt, *supra* note 7, at 203-06.

[85] Cabitza, Ciucci & Rasoini, *supra* note 58.

parameters.[86] This relates to the fact that the bigger the data, the more spurious patterns will be generated.[87] It therefore seems unwise to assume that predictive accuracy goes up when interpretability goes down, or vice versa — it all depends, and major caveats apply. On top of that, we now have various types of software that provide insight into the specific parameters that determined an individual decision,[88] even if the machine learning application actually took into account thousands of variables that both overlap and interact with each other in ways that cannot be grasped by an individual human mind.[89]

Let me remind the reader of which profile transparency rights and obligations have been attributed by the GDPR. Note that recitals are not binding in the way that legislation or case law is, but having been articulated by the legislator, recitals are considered crucial indications for correct interpretation. In articles 13.2(f), 14.2(g) and 15.1(h) we find the following transparency obligations for data controllers and rights for the data subject: "the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."[90] Recital (71) adds:

> In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the

[86] Pearl, *supra* note 56.

[87] On the surge in spurious correlations in big datasets, see Cristian S. Calude & Giuseppe Longo, *The Deluge of Spurious Correlations in Big Data*, 22 Found. Sci. 595 (2017).

[88] Marco Tulio Ribeiro, *LIME - Local Interpretable Model-Agnostic Explanations*, Marco Tulio Ribeiro (Apr. 2, 2016), https://homes.cs.washington.edu/~marcotcr/blog/lime/. *See also* the underlying paper, Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, 22 Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery & Data Mining 1135 (2016); Transparent Data Mining for Big and Small Data (Tania Cerquitelli, Daniele Quercia, & Frank Pasquale eds., 2017).

[89] At least not in the sense of being able to explain the output or to reason how it came about. This type of high-dimensional learning algorithms may be closer to our unconscious intuitions than our ability to reason about things. Yann LeCun, Yoshua Bengio & Geoffrey Hinton, *Deep learning*, 521 Nature 436, 441 (2015). In that sense, we can speak of a digital unconscious. *Cf.* Hildebrandt, *supra* note 7, at 65-77.

[90] GDPR, *supra* note 83, at art. 13 (concerning the processing of data obtained from the data subject, in which case the information must be provided "at the time when personal data are obtained."); *id.* at art. 14 (concerning the processing where the data has not been obtained from the data subject).

> right to obtain human intervention, to express his or her point of view, *to obtain an explanation of the decision reached after such assessment and to challenge the decision* (emphasis added).[91]

In a series of articles, these rights have been coined as rights to "profile transparency,"[92] as a "right to explanation,"[93] and/or as a "right to information about."[94] Discussions have emerged about what "meaningful information" could mean, whether the explanation should take an ex ante perspective and concern only the logic of the decision-making process (which is not personal data and may be protected by trade secret or intellectual property rights), or also (or only) take an ex post perspective and concern the precise parameters that determined the individual decision (which will be personal data as it targets specific data points of the individual that is the object of the decision).[95]

Other discussions focus on the conditions that define the automated decision of art. 22.1: "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."[96] Recital 71 adds:

> (…) such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. Such processing includes "profiling" that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal

[91] Profile transparency should not disproportionally adversely affect trade secret and intellectual property rights. *See id.* at recital 71 ("[t]hat right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software. However, the result of those considerations should not be a refusal to provide all information to the data subject.").

[92] Hildebrandt, *supra* note 65.

[93] Goodman & Flaxman, *supra* note 65.

[94] Wachter, Mittelstadt & Floridi, *supra* note 65.

[95] *See, e.g.*, Edwards & Veale, *supra* note 65.

[96] GDPR, *supra* note 83, at art. 22 (whether a prohibition, a right to object or something else, providing for three exceptions that justify automated decisions — under specific conditions such as relevant safeguards and a right to human intervention, while also providing for a clear-cut prohibition of automated decisions based on special categories of data — basically data that would enable discrimination on the basis of race, etc. — with very narrowly defined exceptions).

> preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her.

Here the discussion circles around whether this must be read as a prohibition,[97] as a right to object, or as something else (but what?);[98] whether "based solely on automated processing" excludes decisions where human intervention is restricted to routine endorsement of the output of profiling;[99] what could be meant by legal effect concerning her (in the strict sense any contract has legal effect, also when buying bread); and whether "similarly significantly affects her" implies that only a significant legal effect should be taken into account, or whether any legal effect is considered significant and therefore all other effects that somehow resemble a legal effect must be taken into account.[100]

I could continue this inventory, since many other points have been and will be made that are all highly relevant and/or interesting. In the context of this article, however, I want to make two points. First, in the case of the IoT, automated decisions will abound and it will be increasingly important that humans become aware of whether they are dealing with human or machine interlocutors. The transparency rights around profiling, at least in the case of automated decisions, will help to remind us whenever we are confronted with machine-made decisions. The transparency obligations demand that (1) the existence of such decisions is communicated, including (2) meaningful

[97] Art. 29 Working Party reads art. 22.1 as a straightforward prohibition. *Guidelines of the Article 29 Data Protection Working party on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679*, at 9, WP 251, 17 (Oct. 3, 2017).

[98] Lee Bygrave, *Minding the Machine: Art. 15 of the EC Data Protection Directive and Automated Profiling*, 17 COMPUTER L. & SECURITY REP. 17 (2001). Wim Schreurs et al., *Cogitas Ergo Sum. The Role of Data Protection Law and Non-discrimination Law in Group Profiling in the Private Sector*, *in* PROFILING THE EUROPEAN CITIZEN: CROSS-DISCIPLINARY PERSPECTIVES, *supra* note 5, at 241; Edwards & Veale, *supra* note 65.

[99] Art. 29 Working Party considers that "if someone routinely applies automatically generated profiles to individuals without any actual influence on the result, this would still be a decision based solely on automated processing," calling this "fabrication of human involvement." *Article 29 Data Protection Working Party Guidelines*, *supra* note 97, at 10.

[100] *See, e.g.*, *id.* at 10:

> A legal effect suggests a processing activity that has an impact on someone's legal rights, such as the freedom to associate with others, vote in an election, or take legal action. A legal effect may also be something that affects a person's legal status or their rights under a contract.

information about the underlying logic and (3) the significance and the envisaged consequences. Though it makes a difference whether this must be understood as relating to the decision-making system or as relating to the individual decision itself, *those employing machine learning applications to replace human decision-makers will have to provide information that is meaningful for those confronted with such decisions, including information that enables them to anticipate and contest the consequences*. In that sense, I would agree with Wachter, Mittelstadt and Russell, where they argue for "counterfactual explanations":

> Explanations of automated decisions need not hinge on the general public understanding how algorithmic systems function. *Even though such interpretability is of great importance and should be pursued*, explanations can, in principle, be offered without opening the black box. Looking at explanations as a means to help a data subject act rather than merely understand, one could gauge the scope and content of explanations according to the specific goal or action they are intended to support (my emphasis).[101]

The salience of their concept of "counterfactual explanations" resides in attempts to clarify *for individuals targeted by automated decisions*, amongst others, "what would need to change in order to receive a desired result in the future, based on the current decision-making model."[102] More precisely: "Counterfactuals describe a dependency on the external facts that lead to that decision without the need to convey the internal state or logic of an algorithm."[103] The "need to change" in "what would need to change" can refer to the machine learning application, to those who train the algorithms, but — obviously — also to the individual who may decide to change her behaviors. My argument would be that agonistic machine learning may enable such counterfactual explanations, while highlighting the potential of the "need to change" for all sides (not just the data subject), including a redistribution of actionability and

[101] Wachter, Mittelstadt & Russell, *supra* note 56. In many ways, the salience of their concept of "counterfactual explanations" aligns with the idea of agonistic machine learning and with previous calls for "counter profiling," which refers to using machine learning on the side of end-users that should help to infer how their behaviors are captured and what inferences can be made based on such behavioral user data. *See, e.g.*, Adrian Popescu et al., *Increasing Transparency and Privacy for Online Social Network Users - USEMP Value Model, Scoring Framework and Legal*, 9484 Privacy Tech. & Pol'y 38 (2016); Hildebrandt, *supra* note 7, at 222-24.

[102] Wachter, Mittelstadt & Russell, *supra* note 56, at 4.

[103] *Id.* at 5.

responsibility amongst developers, profilers and those profiled. On top of that, agonistic machine learning highlights the contestability at the level of the inner workings of machine learning systems — which goes further and is aligned with the methodological integrity of machine learning practice. Articles 13-15 and 22 of the GDPR seem to me a critical contribution to restoring checks and balances in the relationship between us humans and the machines that reconfigure our environment, as it will allow contestation of the accuracy, relevance and reliability of these systems — hopefully resulting in instances of agonistic machine learning, with people pointing out that the output of the algorithm may be inaccurate, spurious, irrelevant or otherwise debatable. Crucially, this need not necessarily be done by individuals, as Article 80.1 stipulates that individual data subjects can mandate their rights to an effective remedy to a dedicated NGO (apart from the right to compensation, which depends on national law). If agonistic machine learning makes any sense, it will be that *the learning process is not merely one of machines, but becomes an individual, social and institutional learning process*.

The second point I want to make is that we should not mistake the legal obligation to justify actions or decisions for the right to explanation and/or information, even though they are clearly related. *Explanation in itself does not imply justification, and justification does not always require an explanation of the underlying logic of the decision system*. If the decisions of an automated machine learning application indirectly discriminate on the basis of gender or race they may qualify as prohibited discrimination; explaining why the system so decided may be interesting but will not legally justify the decision. A decision of an automated system should be justifiable independently of how the system came to its conclusion. When a court decides a case, it cannot justify its decision by spelling out the heuristics of the judge(s) involved, such as their political preferences, what they had for breakfast or how they prepared the case.[104] Though this may be of interest for legal sociologists, the law requires that they motivate their decisions in reference to a set of available legal reasons which thus restrict their ability to decide one way or another. And this is precisely the role of an independent court.

[104] Though some may believe that machine learning in law will reveal the truth of what they think is "legal realism" (showing that the judge's breakfast is the "real" explanation of the court's decision), serious legal realism does not trade formal positivism for sociological positivism. *Cf.* William Twining, Karl Llewellyn and the Realist Movement (2d ed. 2012). For a balanced understanding of pragmatist and normative accounts of legal practice, see, for example, Sanne Taekema, *Theoretical and Normative Frameworks for Legal Research: Putting Theory into Practice*, 2018 Law & Method 1.

As to commercial enterprise, merely providing the inner workings of learning algorithms will not do either — when it come to the lawfulness of a decision or action. For instance, if credit for an online sale is refused based on the freedom to contract and there is no reason to believe that the refusal constitutes prohibited discrimination, the decision may be legally justified by the freedom to contract and thereby lawful. The fact that the system may be biased towards a few trivial data points, such as *e.g.* birthdate, timing of the purchase (day or night) and payment method, or towards specific clickstream behaviors does not necessarily turn it into an unlawful decision, even though in the EU data subjects may have a specific profile transparency right if the decision has a legal or similarly significant impact. This transparency right is nevertheless crucial to initiating agonistic machine learning practices that involve not only domain experts and data scientists, but also citizens and those that wish to mandate their right to enable collective action (based on Article 79 of the GDPR).

Be that as it may, we must not allow the discourse of explainability to stand in the way of the question whether a decision is legally justified, which requires a specific type of *legal* reasons. For instance, the conviction of a defendant cannot be based on the predictive accuracy of an algorithm, even if we "understand" how it works, but only on the legal grounds that justify conviction. Refusing credit, flexible pricing or raising an insurance premium may be based on the freedom to contract, but that freedom is not unlimited and consumer law, competition law, financial services law and insurance law may stipulate further restrictions that must be met, potentially requiring a motivation for a refusal or specific types of price differentiation.[105] Such motivation does not concern the inner workings of a machine learning application, but reasons as provided by law. It may be part of the capture and reconfiguration of our environment that we have become so focused on the black box of machine algorithms,[106] instead of demanding legal justification. We should resist attempts to lure us into accepting the drawbacks of "computer says no" based on a flawed belief in computers that supposedly "outperform" human decision-makers.[107] It is time to recapture our environment, move back from

[105] Mireille Hildebrandt, *Primitives of Legal Protection in the Era of Data-Driven Platforms*, 2 Geo. L. Tech. Rev. 252, 270-73 (2018).

[106] *See* Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (2015).

[107] Some would blame automation bias as a fallacy that is inherent in human cognition, while others may point out that such automation bias may actually be exploited or even instigated by the choice architectures offered by data driven environments. *See, e.g.*, Danielle K. Citron, *Technological Due Process*, 85 Wash. U. L. Rev. 1249, 1271-72 (2008).

bits to atoms and celebrate our incomputable self. Not by rejecting machine learning, but rather by rejecting the assumption that its output defines us, and by getting involved in the politics of defining task T, performance metric P and experience E.

Conclusion

This Article has probed the notion of privacy as the protection of the incomputable self in the era of global, local and virtual data-driven infrastructures. To understand the idea of an incomputable self, I have mobilized the philosophy of the self, starting from the enigma of the grammatical first-person perspective that enables objectified third-person perspectives (a third-person perspective can only be taken by a first person). Highlighting the philosophical work of four key philosophers who worked on the process of self-constitution, I have approached the "self as another" (Ricoeur), as a curious amalgam of an ephemeral "I" and a dynamic "me" (Mead), underlining our crucial capability to take an "ex-centric position" as core to our nonessentialist essence (Plessner), and finally calling attention to our "natality" as the heart of the matter (Arendt). Natality roots our ability to learn new things, to recognize and establish new patterns, and to respond with new ways of navigating our shared world. It is not about behavior but concerns action and interaction. Natality is rooted in the plasticity of our brains and the generative nature of human language, reinforced by the technologies of the script and the printing press. Based on this, the first part of this Article argued that a relational conception of privacy misses out on the ecological perspective that is needed to frame the technological mediation of our relationality.

Part II of the Article examined the affordances of machine learning as critical to the reconfiguration of the self as a computable entity, whose machine-readable behaviors can be used to frame, target and manipulate its consumer, political, religious and other preferences. The firestorm that erupted in March 2018 around the 50.000.000 Facebook profiles that were "leaked" to Cambridge Analytica and used in the 2016 U.S. elections shows outrage at the breach of trust between Facebook and its users. More to the point, it demonstrates that a platform built to induce clickstream behaviors to serve advertising profits can also be used to induce voting and other political behavior*s*.[108] Political

[108] Zeynep Tufekci, *Facebook's Surveillance Machine*, N.Y. Times (Mar. 19, 2018). Cass Sunstein, Republic.com (2001); Eli Pariser, The Filter Bubble: What the Internet is Hiding from You (2011). Robert Epstein & Ronald E. Robertson, *The Search Engine Manipulation Effect (SEME) and its Possible Impact on the Outcomes of Elections*, 112 Proceedings Nat'l. Acad. Sci. 4512 (2015). See also

opinion becomes a political preference, as manipulable as any other preference — based on a computational version of old-school behaviorism.[109] Such behaviorism may reduce our agency and diminish our natality. In Arendt's own words, "[t]he trouble with modern theories of behaviourism is not that they are wrong but that they could become true."[110] In this Article, I have taken the position that our foundational incomputability does not rule out machine learning as a means to enhance both our agency and our natality, depending on whether we take the time and make the effort to understand how it can be designed in ways that set us free instead of chaining us to patterns mined from historical data. I have proposed the notion of agonistic machine learning, to highlight how the new manipulability that comes with computational inferencing requires rethinking democratic theory as well as the practice of technology assessment.

In Part III, I discussed how agonistic machine learning relates to legal protection, and more specifically, how the EU General Data Protection Regulation will help to enhance the methodological integrity of machine learning, while also bringing adversariality into the research design of machine learning. I have described how, contrary to received opinion, the core principles of data minimization and purpose limitation do fit the requirements for robust machine learning, steering free of both data obesitas and pattern obesitas. Agonistic machine learning should enable us, the people, to make informed choices about whether, when and how machine learning applications can best be integrated in human society, instead of taking for granted that they will solve major problems without creating more complex and costly problems. In the end this means that our incomputability is in part protected by a practical and actionable right to reject computation and/or to be computed in alternative ways, underlining the indeterminate nature of each and every individual person and the "equal respect and concern" that our governments owe each of them.[111]

three recent reports by the UK Information Commissioner's Office (ICO) about their investigation into data analytics for political purposes. *Investigation Into Data Analytics for Political Purposes*, Info. Comm'r Off. (July 11, 2018), https://ico.org.uk/action-weve-taken/investigation-into-data-analytics-for-political-purposes/.

[109] Mireille Hildebrandt, *Learning as a Machine: Crossovers Between Humans and Machines*, 4 J. Learning Analytics 6 (2017).

[110] Arendt, *supra* note 5, at 322.

[111] The idea that taking rights seriously follows from the imperative that people have a right to equal concern and respect from those who govern them was developed by Ronald Dworkin. Ronald Dworkin, Taking Rights Seriously (1978).