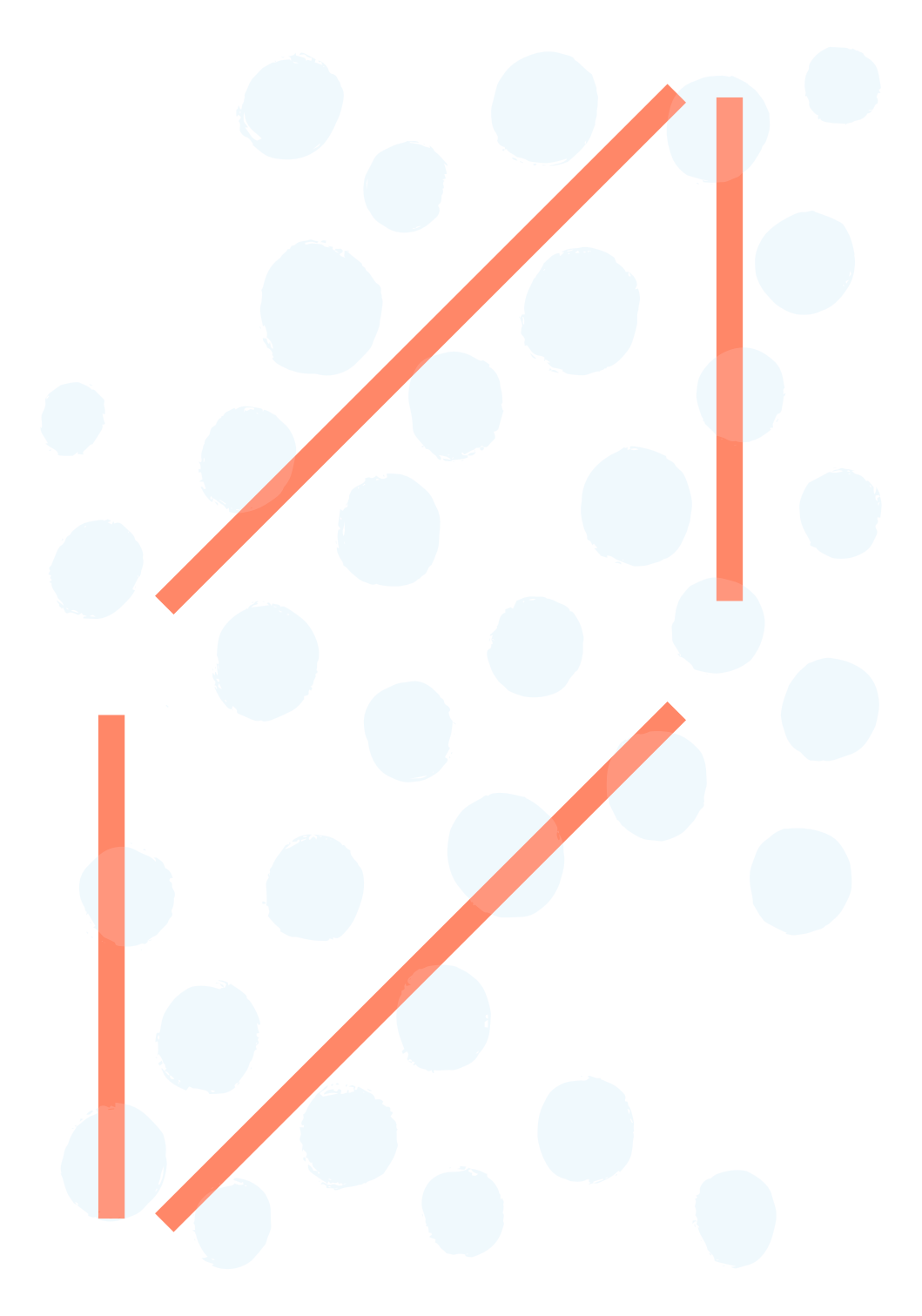
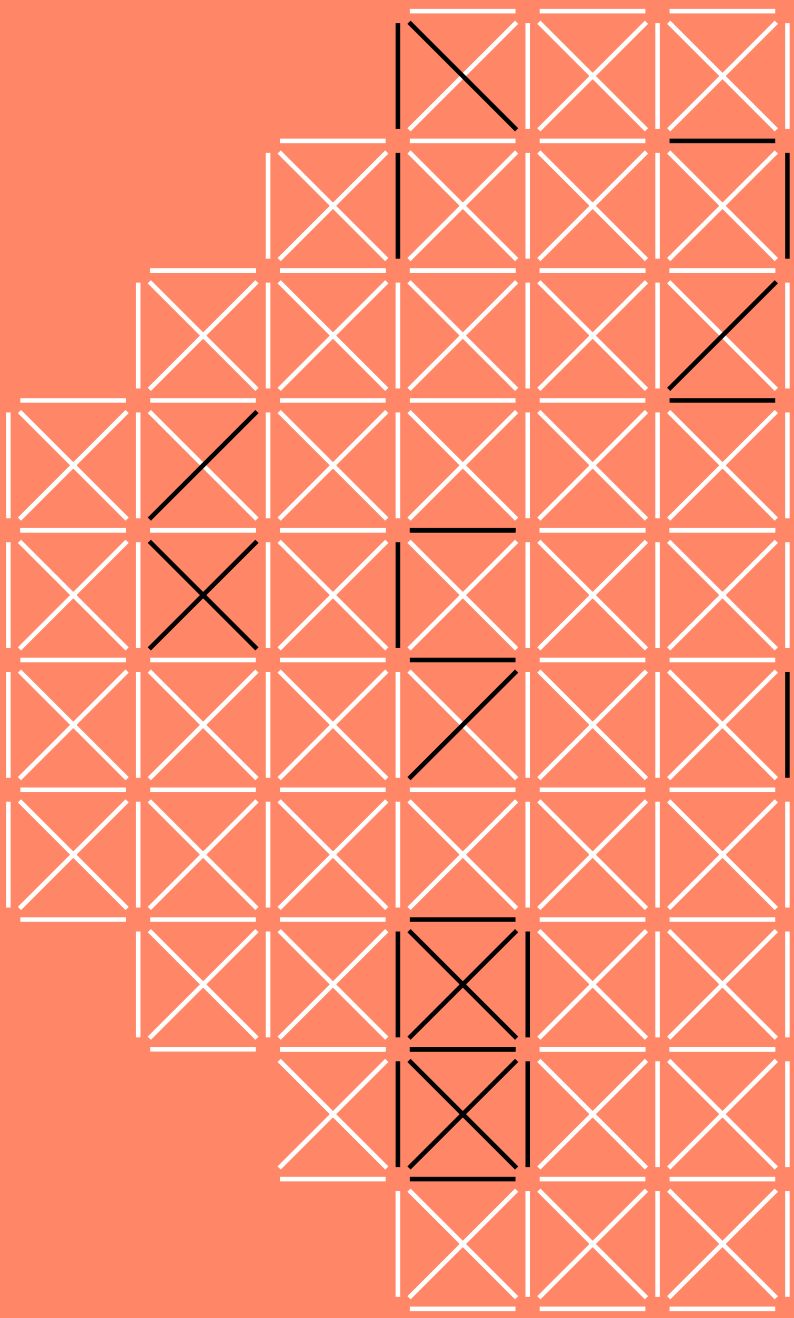
BEING PROFILED:COGITAS ERGO SUM





ISBN 978 94 6372 212 4
e-ISBN 978 90 4855 018 0
DOI 10.5117/9789463722124
NUR 740

Design: Bob van Dijk Studio, Amsterdam
Print: Akxifo, Poeldijk

BEING PROFILED: COGITAS ERGO SUM 10 YEARS OF 'PROFILING THE EUROPEAN CITIZEN'

EMRE BAYAMLIOĞLU,
IRINA BARALIUC,
LIISA JANSSENS,
MIREILLE HILDEBRANDT
(EDS)

Amsterdam University Press B.V., Amsterdam 2018

TABLE OF CONTENTS

PART II
Transparency theory for data-driven decision making

PART III
Presumption of innocence in data-driven government

PROFILING THE EUROPEAN CITIZEN: WHY TODAY'S DEMOCRACY NEEDS TO LOOK HARDER AT THE NEGATIVE POTENTIAL OF NEW TECHNOLOGY THAN AT ITS POSITIVE POTENTIAL

PAUL NEMITZ

This book contains detailed and nuanced contributions on the technologies, the ethics and law of machine learning and profiling, mostly avoiding the term AI. There is no doubt that these technologies have an important positive potential, and a token reference to such positive potential, required in all debates between innovation and precaution, hereby precedes what follows.

The Law neither can nor should aim to be an exact replica of technology, neither in its normative endeavour nor indeed in its use of terminology. Law and ethics need to be technology neutral wherever possible, to maintain meaning in relation to fast evolving technologies, and so should be the writing about the law and ethics.

The technological colonisation of our living space raises fundamental questions of how we want to live, both as individuals and as a collective. This applies irrespective of whether technologies with potentially important negative effects also have an important positive potential, even if such negative effect are unintended side effects.

While the technological capabilities for perfect surveillance, profiling, predicting, and influencing of human behaviour are constantly evolving, the basic questions they raise are not new.

It was Hans Jonas (1985), who in his 1979 bestseller *The imperative of responsibility* criticized the disregard, which the combined power of capitalism and technology shows for any other human concern. He laid the ground for the principle of precaution, today a general principle of EU law,[1] relating to any technology, which fulfils two conditions: Long-term unforeseeable impacts, and a possibility that these long term impacts can substantially negatively affect the existence of humanity. While his motivator at the time were the risks of nuclear power, he already in the 'Principle of responsibility' mentioned other examples of trends, which needed a precautionary approach, such as increasing longevity. Nuclear power at the time contained the great promise of clean and cheap, never-ending energy, alongside the risks of the technology which were known early on. And today again, the great promises of the internet and artificial intelligence are accompanied by risks which are already largely identified.

Large-scale construction of nuclear power plants proceeded, in part because the risk of radiation was invisible to the public. Only after the risks became visible to the general public through not only one, but a number of successive catastrophic incidents, did the tide change on this high risk technology.

And again today, digital technologies proceed largely unregulated and with numerous known risks, which however are largely invisible to the general public.

The politics of invisible risks, whether relating to nuclear power, smoking or digital surveillance, artificial intelligence, profiling and manipulative nudging, consists of a discourse of downplaying risks and overstating benefits, combined with the neo liberal rejection of laws that constrain enterprises, in order to maintain the space for profit as long as possible.

The question thus could be, following the example of nuclear power: How many catastrophes of surveillance, profiling and artificial intelligence going wild do we have to go through before the tide changes, before the risks are properly addressed?

With the technologies of the internet and artificial intelligence, we cannot afford to learn only by catastrophe, as we did relating to nuclear power. The reason is that once these technologies have reached their potential to win every game, from the stock markets to democratic decision-making, their impacts will be irreversible. There is no return from a democracy lost in total surveillance and profiling, which makes it impossible to organise opposition. And there is no return to the status quo ante due to a stolen election or popular vote, as we are now witnessing with the Brexit. The British people will go through a decade-long valley of tears because their vote on Brexit was stolen by the capabilities of modern digital technological manipulation of the vote. A whole generation of British youth pays the price for a lack of precaution as regards these technologies.

Like in relation to nuclear power, it is vital that those who develop and understand the technology step forward and work with rigour to minimise the risks arising from the internet, surveillance, profiling and artificial intelligence. We need the technical intelligentsia to join hands with social science, law and democracy. Technological solutions to achieve risk mitigation must go hand in hand with democratic institutions taking their responsibility, through a law, which can be enforced against those actors who put profit and power before democracy, freedom and the rule of law.

Constitutional democracy must defend itself again against absolutist ambitions and erosions from within and from the outside.[2] In the times of German Chancellor Willy Brandt, a drive to convince the technical intelligentsia to engage for a just society, for democracy and environmental sustainability took off, spurred by both his principles of 'Mehr Demokratie wagen' ('Dare more Democracy')[3] and 'Wehrhafte Demokratie' ('A democracy which defends itself') and its critical reception.[4] It is this spirit of post-1968, which we need to bring back into the digital global debate.

From the Chinese dream of perfecting communism through surveillance technology and social scoring to the Silicon Valley and Wall Street dream of perfect predictability of market related behaviour of individuals: The dystopian visions of total surveillance and profiling and thus total control over people are on the way of being put in practice today. We are surrounded by regressive dreams of almightiness based on new technology (Nida-Rümelin and Weidenfeld 2018).

Individuals in this way become the objects of other purposes – they are being nudged and manipulated for profit or party line behaviour, disrobed of their freedom and individuality, their humanity as defined by Kant and many world religions.

Finding ways of developing and deploying new technologies with a purpose restricted to supporting individual freedom and dignity as well as the basic constitutional settlements of constitutional democracies, namely democracy, rule of law and fundamental rights is the challenge of our time.

And continuing to have the courage to lay down the law guiding the necessary innovation through tools such as obligatory technology impact assessments and an obligation to incorporate principles of democracy, rule of law and fundamental rights in technology, is the challenge for democracy today: Let us dare more democracy by using the law as a tool of democracy for this purpose. And let us defend democracy through law and engagement. Europe has shown that this is possible, the GDPR being one example of law guiding innovation through 'by design' principles and effective, enforceable legal obligations regarding the use of technology.

Paul Nemitz[5]
Brussels, November 2018

Notes

[1] See to that effect the blue box on page 3 of the Strategic Note of the European Political Strategy Centre of the European Commission (2016), available at https://ec.europa.eu/epsc/sites/epsc/files/strategic_note_issue_14.pdf; see also ECJ C- 157/96, Para 62 ff, C-180/96, Para 98 ff. and C-77/09, Rn. 72.

[2] Nemitz (2018), see also Chadwick (2018).

[3] See Brandt (1969).

[4] A key action of 'Wehrhafte Demokratie' under Willy Brandt was the much contested order against radicals from the left and the right in public service of 28 January 1972, available at https://www.1000dokumente.de/index.html?c=dokument_de&dokument=0113_ade&object=translation&st=&l=de. On this, see also Wissenschaftlicher Dienst des Deutschen Bundestages (2017), and more recently the translation of 'Wehrhafte Demokratie' as 'militant democracy' in the press release of the German Constitutional Court (2018) on an order rejecting constitutional complaints against prohibitions of associations. This order recounts in part the history of 'Wehrhafte Demokratie' and the lack of it in the Weimar Republic.

[5] The author is Principal Advisor in DG JUSTICE at the European Commission and writes here in his personal capacity, not necessarily representing positions of the Commission. He is also a Member of the German Data Ethics Commission, a Visiting Professor of Law at the College of Europe in Bruges and a Fellow of the VUB, Brussels.

References

Brandt, Willy. 1969. Regierungserklärung (Government declaration) of 28 October 1969, available at https://www.willy-brandt.de/fileadmin/brandt/Downloads/Regierungserklaerung_Willy_Brandt_1969.pdf.

Chadwick, Paul. 2018. 'To Regulate AI We Need New Laws, Not Just a Code of Ethics | Paul Chadwick'. The Guardian, 28 October 2018, sec. Opinion. https://www.theguardian.com/commentisfree/2018/oct/28/regulate-ai-new-laws-code-of-ethics-technology-power.

European Political Strategy Centre of the European Commission. 2016. "Towards an Innovation Principle Endorsed by Better Regulation", Strategic Note 14 of June 2016, available at https://ec.europa.eu/epsc/sites/epsc/files/strategic_note_issue_14.pdf.

German Constitutional Court, Press Release No. 69/2018 of 21 August 2018 on the Order of 13 July 2018 1 BvR 1474/12, 1 BvR 57/14, 1 BvR 57/14, 1 BvR 670/13, available at https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/EN/2018/bvg18-069.html.

Jonas, Hans. 1985. The Imperative of Responsibility: In Search of an Ethics for the Technological Age. Chicago (Ill.): University of Chicago Press.

Nemitz, Paul. 2018. 'Constitutional Democracy and Technology in the Age of Artificial Intelligence'. Phil. Trans. R. Soc. A 376 (2133): 20180089. https://doi.org/10.1098/rsta.2018.0089.

Nida-Rümelin, Julian, and Nathalie Weidenfeld. 2018. Digitaler Humanismus: Eine Ethik für das Zeitalter der Künstlichen Intelligenz. München: Piper.

Wissenschaftlicher Dienst des Deutschen Bundestages. 2017. Parlamentarische und zivilgesellschaftliche Initiativen zur Aufarbeitung des sogenannten Radikalenerlasses vom 28. Januar 1972, Ausarbeitung WD 1 - 3000 - 012/17, available at https://www.bundestag.de/blob/531136/a0a150d89d4db6c2b-dae0dd5b300246d/wd-1-012-17-pdf-data.pdf.

INTROITUS: WHAT DESCARTES DID NOT GET

MIREILLE HILDEBRANDT

Entering the hardcopy of this book is a tactile experience, a rush on the senses of touch, vision and possibly smell. Colour, graphics and the brush of unusual paper against one's digits (Latin for fingers) may disrupt the expectations of the academic reader. This is what information does, according to Shannon and Wiener, two of the founding fathers of information theory (Hildebrandt 2016,16-18). Information surprises by providing input we did not anticipate, it forces us to reconfigure the maps we made to navigate our world(s). The unexpected is also what draws our attention, that scarce good, so in vogue amongst ad tech companies. Maybe, this is where hardcopy books will keep their edge over the flux of online temptations.

Computing systems have redistributed the playing field of everyday life, politics, business and art. They are game changers and we know it. We now have machines that learn from experience; inductive engines that adapt to minuscule perturbations in the data we feed them. They are far better at many things than previous machines, that could only apply the rules we gave them, stuck in the treadmill of a deductive engine.

We should, however, not be fooled by our digital companions and their masters, the new prestidigitators. As John Dewey (2008, 87) reported in his *Freedom and Culture* in the ominous year 1939, we should remember that:

> *the patter of the prestidigitator enables him to do things that are not noticed by those whom he is engaged in fooling.*

A prestidigitator is a magician, paid to fool those who enjoy being tricked into expected surprises. A successful magician knows how to anticipate their audience, how to hold the attention of those seated in front of them and how to lure their public into awe and addiction. A good audience knows it is fooled and goes back to work in awe but without illusions.

What Descartes and the previous masters of artificial intelligence did not get was how others shape who and what we are. How anticipation, experience and feedback rule whatever is alive. We are not because we think (*cogito ergo sum*); we are because we are being addressed by others who 'think us' – one way or another (*cogitas ergo sum*) (Schreurs et al. 2008). Being profiled by machines means being addressed by machines, one way or another. This will impact who we are, as we are forced to anticipate how we are being profiled, with what consequences.

In 2008, *Profiling the European Citizen* brought together computer scientists, lawyers, philosophers and social scientists. They contributed with text and replies, sharing insights across disciplinary borders. On what profiling does, how it works and how we may need to protect against its assumptions, misreadings and manipulative potential. Today, in 2018, *BEING PROFILED* does the same thing, differently. Based on 10 years of incredibly rapid developments in machine learning, now applied in numerous real-world applications. We hope the reader will be inspired, informed and invigorated on the cusp of science, technology, law and philosophy – ready to enjoy magic without succumbing to it.

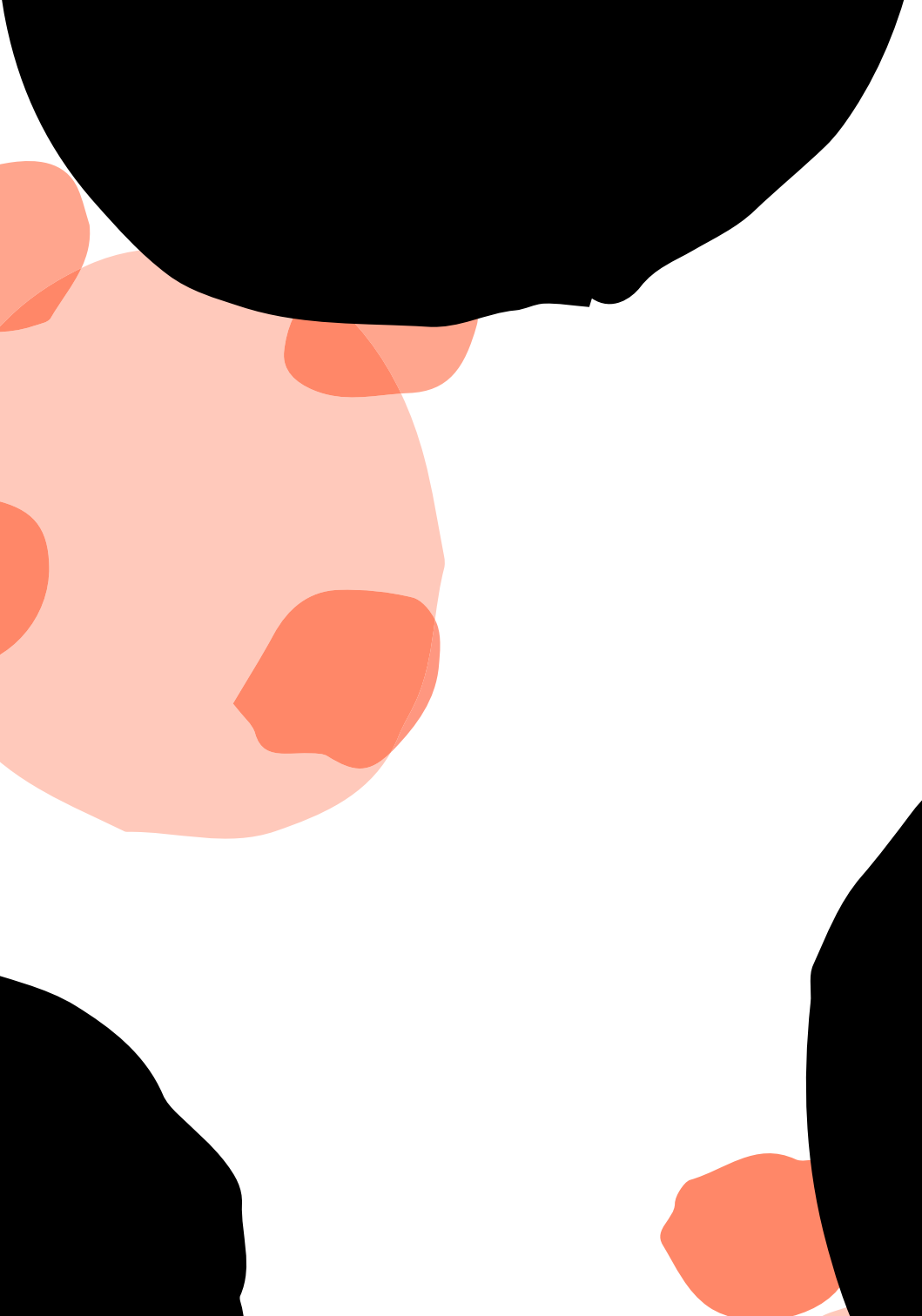
Mireille Hildebrandt
December 2018, Brussels

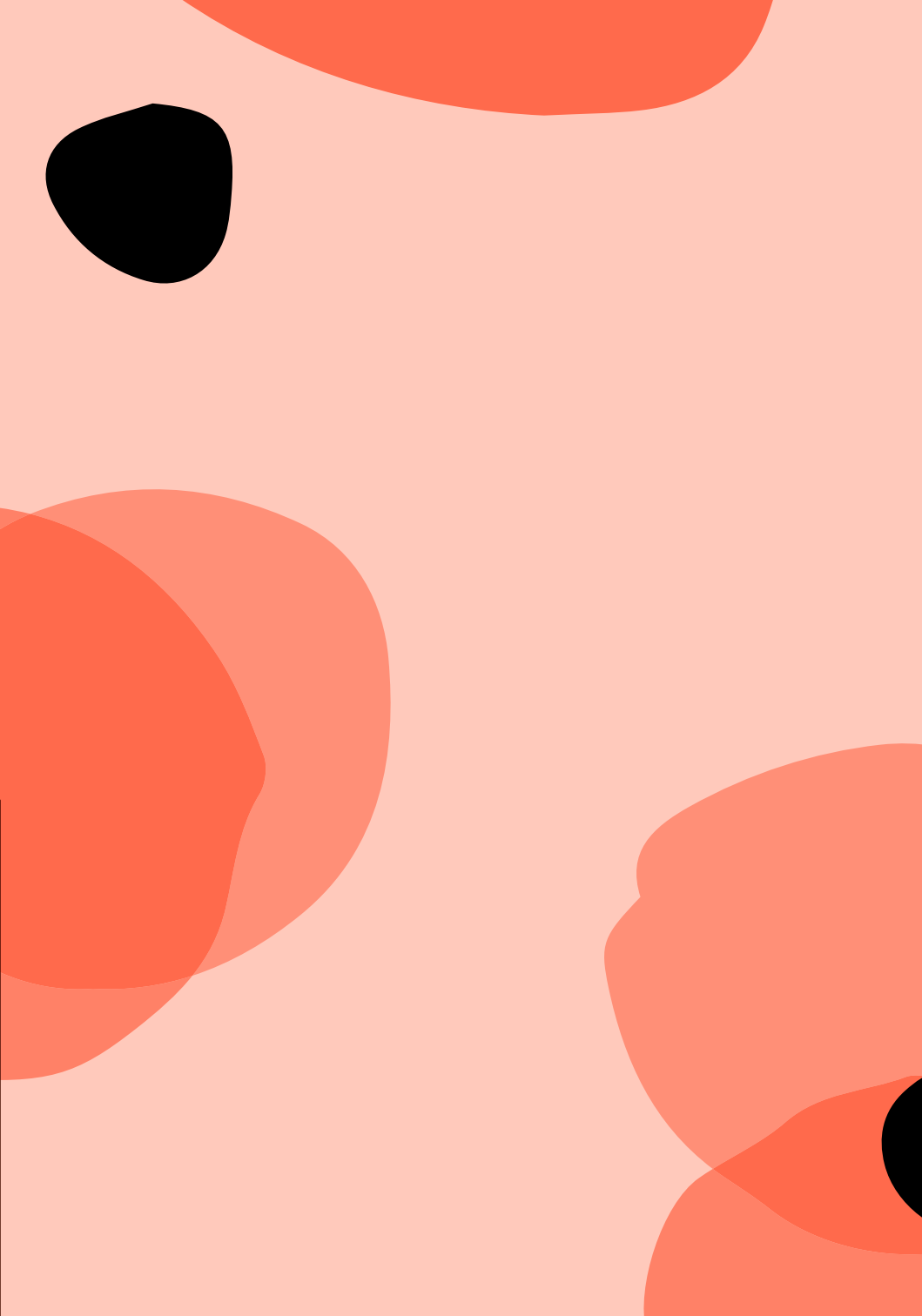
References

Dewey, John. 2008. 'Freedom and Culture'. In The Later Works of John Dewey, 1925 – 1953, Vol. 13: 1938-1939, Experience and Education, Freedom and Culture, Theory of Valuation, and Essays edited by Jo Ann Boydston, 63-188. Carbondale: Southern Illinois University Press.

Hildebrandt, Mireille. 2016. 'Law as Information in the Era of Data-Driven Agency'. The Modern Law Review 79 (1): 1–30. doi:10.1111/1468-2230.12165.

Schreurs, Wim, Mireille Hildebrandt, Els Kindt, and Michaël Vanfleteren. 2008. "Cogitas Ergo Sum: The Role of Data Protection Law and Non-Discrimination Law in Group Profiling in the Private Sphere." In Profiling the European Citizen: Cross-Disciplinary Perspectives, edited by M. Hildebrandt and S. Gutwirth, 242-270. Dordrecht: Springer.





FROM AGENCY-ENHANCEMENT TO INSTINCT-SUBSTITUTION: WHAT IS LOST IN PROFILE-BASED OPTIMISATION TOOLS?

SYLVIE DELACROIX

Whether it be by increasing the accuracy of Web searches, educational interventions or policing, the level of personalisation that is made possible by increasingly sophisticated profiles promises to make our lives better. Why 'wander in the dark', making choices as important as that of our lifetime partner, based on the limited amount of information we humans may plausibly gather? The data collection technologies empowered by wearables and apps mean that machines can now 'read' many aspects of our quotidian lives. Combined with fast evolving data mining techniques, these expanding datasets facilitate the discovery of statistically robust correlations between particular human traits and behaviours, which in turn allow for increasingly accurate profile-based optimisation tools. Most of these tools proceed from a silent assumption: our imperfect grasp of data is at the root of most of what goes wrong in the decisions we make. Today, this grasp of data can be perfected in ways not necessarily foreseeable even 10 years ago, when *Profiling the European Citizen* defined most of the issues discussed in this volume. If data-perfected, precise algorithmic recommendations can replace the flawed heuristics that preside over most of our decisions, why think twice? This line of argument informs the widely-shared assumption that today's profile-based technologies are agency-enhancing, supposedly facilitating a fuller, richer realisation of the selves we aspire to be. This 'provocation' questions this assumption.

Fallibility's inherent value

Neither humans nor machines are infallible. Yet our unprecedented ability to collect and process vast amounts of data is transforming our relationship to both fallibility and certainty. This manifests itself not just in terms of the epistemic confidence sometimes wrongly generated by such methods. This changed relationship also translates in an important shift in attitude, both in the extent to which we strive for control and 'objective' certainty and in the extent to which we retain a critical, questioning stance.

The data boon described above has reinforced an appetite for 'objective' certainty that is far from new. Indeed one may read a large chunk of the history of philosophy as expressing our longing to overcome the limitations inherent in the fact that our perception of reality is necessarily imperfect, constrained by the imprecision of our senses (de Montaigne 1993). The rationalist tradition which the above longing has given rise to is balanced by an equally significant branch of philosophy, which emphasizes the futility of our trying to jump over our own shoulders, striving to build knowledge and certainty on the basis of an overly restrictive understanding of objectivity, according to which a claim is objectively true only if it accurately 'tracks' some object (Putnam 2004) that is maximally detached from our own perspective. Such aspiring for a Cartesian form of objectivity (Fink 2006) is futile, on this account, because by necessity the only reality we have access to is always already inhabited by us, suffused with our aspirations.

To denigrate this biased, 'subjective' perspective as 'irrational' risks depriving us of an array of insights. Some of these simply stem from an ability for wonder, capturing the rich diversity of human experience, in all its frailty and imperfection. Others are best described as 'skilled intuitions' (Kahneman and Klein 2009) gained through extensive

experience in an environment that provides opportunity for constructive feedback, the insights provided by such skilled intuitions are likely to be dismissed when building systems bent on optimizing evidence-based outcomes. Instead of considering the role played by an array of non-cognitive factors in decisions 'gone wrong', the focus will be on identifying what machine-readable data has been misinterpreted or ignored. If factors such as habits and intuitions are known to play a role, they are merely seen as malleable targets that can be manipulated through adequate environment architecture, rather than as valuable sources of insights that may call into question an 'irrationality verdict'.

Similarly, the possibility of measuring the likely impact of different types of social intervention by reference to sophisticated group profiles is all too often seen as dispensing policy-makers from the need to take into account considerations that are not machine-readable (such as the importance of a landscape). Indeed the latter considerations may not have anything to do with 'data' per se, stemming instead from age-old ethical questions related to the kind of persons we aspire to be. Some believe those ethical questions lend themselves to 'objectively certain' answers just as well as the practical problems tackled through predictive profiling. On this view, perduring ethical disagreements only reflect our cognitive limitations, which could in principle be overcome, were we to design an all-knowing, benevolent superintelligence. From that perspective, the prospect of being able to rely on a system's superior cognitive prowess to answer the 'how should we [I] live' question with certainty, once and for all, is a boon that ought to be met with enthusiasm. From an 'ethics as a work in progress' by contrast, such a prospect can only be met with scepticism at best or alarm at worst (Delacroix 2019b): on this view, the advent of AI-enabled moral perfectionism would not only threaten our democratic practices, but also the very possibility of civic responsibility.

Civic responsibility and our readiness to question existing practices

Ethical agency has always been tied to the fact that human judgment is imperfect: we keep getting things wrong, both when it comes to the way the world is and when it comes to the way it ought to be. The extent to which we are prepared to acknowledge the latter, moral fallibility—and our proposed strategies to address it—have significant, concrete consequences. The latter can be felt at a personal and at an institutional, political level. A commitment to acknowledging our moral fallibility is indeed widely deemed to be a key organising principle underlying the discursive practices at the heart of our liberal democracies (Habermas 1999). This section considers the extent to which the data-fed striving for 'objective certainty' is all too likely to compromise the above commitment.

Now you may ask: why is such a questioning stance important? Why muddle the waters if significant, 'data-enabled' advances in the way we understand ourselves (and our relationship to our environment) mean that some fragile state of socio-political equilibrium has been reached? First, one has to emphasise that it is unlikely that any of the answers given below will move those whose metaphysical or ideological beliefs already lead them to deem the worldview informing such equilibrium to be 'true',

rather than 'reasonable' (Habermas 1995). The below is of value only to those who are impressed enough by newly generated, data-backed knowledge to be tempted to upgrade their beliefs from 'reasonable' to 'true'. A poor understanding of the limitations inherent in both the delineation of the data that feeds predictive models and the models themselves is indeed contributing to a shift in what Jasanoff aptly described as the culturally informed 'practices of objectivity'. In her astute analysis of the extent to which the ideal of policy objectivity is differently articulated in disparate political cultures, Jasanoff highlights the United States' marked preference for quantitative analysis (Jasanoff 2011). Today the recognition of the potential inherent in a variety of data mining techniques within the public sector (Veale, Van Kleek, and Binns 2018) is spreading this appetite for quantification well beyond the United States.

So why does the above matter at all? While a commitment to acknowledging the fallibility of our practices is widely deemed a cornerstone of liberal democracies, the psychological obstacles to such acknowledgment -including the role of habit- are too rarely considered. All of the most influential theorists of democratic legitimacy take the continued possibility of critical reflective agency as a presupposition that is key to their articulation of the relationship between autonomy and authority. To take but one example: in Raz's account, political authority is legitimate to the extent that it successfully enables us to comply with the demands of 'right reason'(Raz 1986). This legitimacy cannot be established once and for all: respect for autonomy entails that we keep checking that a given authority still has a positive 'normal justification score' (Raz 1990). If the 'reflective individual' finds that abiding by that authority's precepts takes her away from the path of 'right reason', she has a duty to challenge those precepts, thereby renewing the fabric from which those normative precepts arise. In the case of a legal system, that fabric will be pervaded by both instrumental concerns and moral aspirations. These other, pre-existing norms provide the material from which the 'reflective individual' is meant to draw the resources necessary to assessing an authority's legitimacy. Much work has gone into analysing the interdependence between those different forms of normativity; not nearly enough consideration has been given to the factors that may warrant tempering political and legal theory's naive optimism—including that of Delacroix (2006)—when it comes to our enduring capacity for reflective agency.

Conclusion

To live up to the ideal of reflectivity that is presupposed by most theories of liberal democracy entails an ability to step back from the habitual and question widely accepted practices (Delacroix 2019a). Challenging as it is to maintain such critical distance in an 'offline world', it becomes particularly arduous when surrounded by some habit-reinforcing, optimised environment at the service of 'algorithmic government'. The statistical knowledge relied on by such form of government does not lend itself to contestation through argumentative practices, hence the temptation to conclude that such algorithmic government can only be assessed by reference to its 'operational contribution to our socio-economic life' (Rouvroy 2016). That contribution will, in many cases, consist in streamlining even the most personal choices and decisions thanks to a 'networked environment that monitors its users and adapts its

services in real time' (Hildebrandt 2008). Could it be the case that a hereto poorly acknowledged side effect of our profile-based systems (and the algorithmic forms of government they empower) consists in its leaving us sheep-like, unable to mobilise a normative muscle that has gone limp through lack of exercise? The more efficient those systems are, the more we are content to offload normative decisions to their 'optimised' algorithms, the more atrophied our 'normative muscles' would become. Considered at scale, the (endless) normative holidays that would result from such 'offloading' would spell the end of agency, and hence the end of legal normativity. All that in spite of the noble, agency-enhancing intentions that prompted the creation of such systems in the first place.

References

Bostrom, Nick. 2014. Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

de Montaigne, Michel. 1993. The complete essays. Translated and edited with an Introduction and Notes by M. A. Screech. London: Penguin Books.

Delacroix, Sylvie. 2006. Legal norms and normativity: an essay in genealogy. Oxford: Hart Publishing.

Delacroix, Sylvie. 2019a. Habitual Ethics? Oxford: Hart Publishing.

Delacroix, Sylvie. 2019b. "Taking Turing by surprise? Designing autonomous systems for morally-loaded contexts." arXiv:1803.04548.

Fink, Hans. 2006. "Three Sorts of Naturalism." European Journal of Philosophy 14(2): 202-21.

Habermas, Jurgen. 1995. "Reconciliation Through the Public use of Reason: Remarks on John Rawls's Political Liberalism." The Journal of Philosophy 92(3): 109-31.

Hildebrandt, Mireille. 2008. "Defining Profiling: A New Type of Knowledge?" In Profiling the European Citizen: Cross-Disciplinary Perspectives, edited by Mireille Hildebrandt and Serge Gutwirth, 17-45. Dordrecht: Springer.

Jasanoff, Sheila. 2011. "The Practices of Objectivity in Regulatory Science" In Social Knowledge in the Making, edited by Charle Camic, Neil Gross, and Michèle Lamont, 307-37. Chicago: University of Chicago Press.

Kahneman, Daniel, and Gary Klein. 2009. "Conditions for intuitive expertise: a failure to disagree." Am Psychol 64(6): 515-26. doi: 10.1037/a0016755.

Putnam, Hilary. 2004. Ethics without ontology. Cambridge, MA: Harvard University Press.

Raz, Joseph. 1986. The morality of freedom. Oxford: Clarendon Press.

Raz, Joseph. 1990. Practical Reason and Norms. Princeton, NJ: Princeton University Press.

Rouvroy, Antoinette. 2016. "La gouvernementalité algorithmique: radicalisation et stratégie immunitaire du capitalisme et du néolibéralisme?" La Deleuziana, (3): 30-36.

Veale, Michael, Max Van Kleek, and Reuben Binns. 2018. "Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making." Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. doi: 0.1145/3173574.317401.

MATHEMATICAL VALUES AND THE EPISTEMOLOGY OF DATA PRACTICES

PATRICK ALLO

Contemporary data practices, whether we call them data science or AI, statistical learning or machine learning, are widely perceived to be game changers. They change what is at stake epistemologically as well as ethically. This especially applies to decision-making processes that infer new insights from data, use these insights to decide on the most beneficial action, and refer to data and to an inference process to justify the chosen course of action. The profiling of citizens is now only one of many such processes.

One of the original goals of 'Profiling the European Citizen' was to understand the nature of the knowledge that data-mining creates and that profiles encode, and to critically assess the epistemic power that is exerted when a profile is applied to an individual. When we develop a critical epistemology for contemporary data practices, we still seek answers to the same questions. We want to know what kind of knowledge is being created, how we may evaluate it, and how it acquires its epistemic authority.

Developing a critical epistemology that does not merely restate the promises of data-driven inquiry, but instead allows us to understand the threats it may pose is a non-trivial task. There is a lack of clarity regarding the epistemological norms we should adhere to. Purely formal evaluations of decisions under uncertainty can, for instance, be hard to assess outside of the formalism they rely on. In addition, there is substantial uncertainty with regard to the applicable norms because scientific norms may appear to be in flux (new paradigms, new epistemologies, etc.) Finally, dealing with this uncertainty and lack of clarity is further complicated by promises of unprecedented progress and opportunities that invite us to imagine a data-revolution with many guaranteed benefits, but few risks.

My goal in this provocation is to focus on a small, easily disregarded, fragment of this broader epistemological project. The inquiry I would like to propose questions the role of mathematics and the role of our beliefs about the nature of mathematical knowledge within contemporary data-practices. What I contend is that, first, there are few reasons to leave the role of mathematics unexamined, and, second, that a conscious reflection on how mathematical thought shapes contemporary data-practices is a fruitful new line of inquiry. It forces us to look beyond data and code (the usual suspects of the critical research agenda on data) and can help us grasp how the epistemic authority of data science is construed.

The role of mathematics

Mathematics does not only contribute to the theoretical foundations of many existing data practices (from sheer counting to learning, categorising, and predicting), but it also contributes to the scientific respectability and trustworthiness of data science. Reliance on mathematics does not only enable (no calculation without mathematics) and certify (no correct calculation without mathematics) data science, but it also makes it credible. Following one of the central motivations of the Strong Programme in the Sociology of Science, I take the task of 'explain[ing] the credibility of a given body of knowledge in given context' (Barnes 1982, xi) to be essential for understanding the epistemology of data science. We should direct our attention to the epistemic authority

of mathematics, the epistemic authority granted by mathematics to its applications, and the view that relying on mathematics is epistemically as well as ethically commendable.

An analysis of the role of mathematics in data science that seeks to account for the credibility and authority of data science can be fruitfully developed with an explicit reference to mathematical values. This can help us understand the epistemological contribution of mathematics to data science. It reveals how mathematics, for many the one source of absolute certainty we have, could have any substantial influence on the epistemology of fallible or merely probable predictions. Certainty and truth, of course, are mathematical values, but so is the importance that is accorded to abstract reasoning, or the requirement that the only acceptable proofs and calculations are those that can independently be verified. By attending to such values, we can discern more clearly the influence of mathematical thought within the realm of uncertain reasoning. This is a first advantage of conceiving of the role of mathematics in terms of the values it promotes and the values it appeals to. In addition, when we shift our attention to values we are no longer restricted to a strict accuracy-centric assessment of probabilistic procedures. The latter perspective is traditionally associated with a consequentialist understanding of good decisions. Instead, we can follow a more flexible assessment that lets us to address additional socio-epistemic requirements like trust, responsibility, or accountability.

The critical evaluation of the role of mathematics in data science should not be reduced to the uncovering of the crushing power of the authority of mathematics, or the dismissal of the mathematically warranted neutrality of algorithmic processes. Instead, we should strive to re-think the ambivalent role of mathematics and of beliefs about mathematics in data science. The interaction between mathematics and data science is bi-directional. Data science appeals to mathematical values—such as objectivity, neutrality, and universality—to legitimate itself, but mathematics also promotes certain values—such as the openness of mathematical justification through proof and calculation—in the knowledge practices that rely on mathematics. I contend that data science seeks to associate itself to mathematical values it fails to live up to, but also that some of some mathematical values are not necessarily virtuous when deployed outside the realm of pure mathematics. Mathematical values can be used critically, for instance by underscoring the epistemic value of practically verifiable calculations, but they can also be used in less critical ways, for instance when mathematical techniques are presented as value-free technological artefacts.

A detailed overview of mathematical values is beyond the scope of the present contribution (I refer the interested reader to the seminal contributions of Alan Bishop and Paul Ernest on whose work I draw, e.g. Bishop 1991; Ernest 2016). I will now just focus on one value to illustrate the ambivalent influence of mathematical values on the epistemology of data science. I propose to focus on the importance that mathematical practices accord to 'closed texture' and will argue that as a property of concepts that is closely associated with the demands of abstraction, precision, and explicitness in mathematical reasoning, it is a perfect example of a janus-faced value that can

have beneficial as well as detrimental consequences in contexts where mathematical techniques are used to derive actionable knowledge from messy data.

Open and closed texture

The notion of 'open texture' was first coined by Friedrich Waismann (1945) to refer to the fact that many concepts or words we use to describe the world are such that the linguistic rules that govern their use do not determinately settle all their possible uses. Some use-cases appear to be open or unlegislated:

> *The fact that in many cases there is no such thing as a conclusive verification is connected to the fact that most of our empirical concepts are not delimited in all possible directions. (…) Open texture, then, is something like possibility of vagueness. Vagueness can be remedied by giving more accurate rules, open texture cannot (Waissman quoted in Shapiro 2006, 210–1).*[1]

Closed texture, then, is the absence of open texture. Mathematics and computing crucially depend on the absence of open texture, where the absence of unlegislated cases is associated with such values as clarity, explicitness, and univocality. The relevance of the contrast between open and closed texture is based on the paradoxical situation that, on the one hand, the semi-technical notion of an algorithm, understood as a procedure that can be executed without having to rely on the ingenuity or informed judgement of the executor of that procedure, is built on the assumption of closed texture, whereas, on the other hand, the concepts we use to deal with the world (so-called empirical concepts) exhibit open texture. Colours in the world exhibit open texture, but the values of a pixel do not; similarly, the properties of a data-subject may be underdetermined, but the values we find in each field of a data-base are, again, a determinate manner. It is because data, or 'capta' (Kitchin and Dodge 2011), especially when understood as simple syntactical objects, do not exhibit open texture that they are fit for algorithmic processing. This rudimentary insight is easily forgotten when learning-algorithms are deployed for tasks, like image-recognition, for which our human ability to interpret and use concepts that exhibit open texture or are imprecise cannot be captured in precise rules. Whether a given image shows a cat is arguably not something that can be mechanically decided, but whether a collection of pixels does or does not match a given pattern can be so decided. The goal of a learning algorithm is precisely to find a good enough replacement of problems of the former type with problems of the latter type.

This much should be uncontroversial but does not yet explain why 'closed texture' is a janus-faced requirement of mathematical reasoning and of algorithmic processing. This requires us to see that while (as I have just argued) closed texture is a technical requirement of any computational process, its epistemological import is not unequivocally positive. This is because, whereas aspiring to clarify as well as one can the concepts one uses is naturally perceived as an intellectual virtue and as a way to avoid

fallacies of equivocation, the closed texture of our concepts is often no more than a convenient (but false) assumption.

Proxies and their target

Let me, to conclude this provocation, briefly describe the risks that are associated with the assumption that all our algorithms operate in the absence of open texture. The risk in question is that the technical need to avoid open texture is easily turned into what van Deemter (2010) calls 'false clarity': our tendency to use imprecise concepts as if they were crisp. Because we replace a question of interest ('is this a cat?') that may not have a determinate answer with a proxy-problem that does have a determinate answer ('is this pattern present?') and can therefore be algorithmically resolved, it is tempting to confuse our ability to correctly solve the proxy-problem with our ability to provide a correct answer to the actual problem. This is especially problematic when the (mathematically supported) trust we place in the former is directly transferred to the latter. It is even more so when a question on which we can reasonably disagree (or whose resolution is context-dependent) is replaced by a question that can be resolved in a controlled environment that does not admit disagreement. In such cases, adherence to the demands of algorithmic processes may spill over into the unwarranted dismissal of critical objections because we confuse the impossibility of disagreeing about the (mathematically represented) proxy-problem with the possibility of disagreeing about the (real-world) target-problem.

Notes

[1] My exposition builds on Shapiro (2006), which focuses more directly on the role of the open and closed texture of concepts within the formal sciences than Hart's seminal work on the open texture of legal rules (Hart & Green 2012; Schauer 2013).

References

Barnes, Barry. 1982. T. S. Kuhn and Social Science. London and Basingstoke: MackMillan.

Bishop, Alan J. 1991. Mathematical Enculturation: A Cultural Perspective on Mathematics Education. Dordrecht, Boston: Kluwer Academic Publishers.

Ernest, Paul. 2016. "Mathematics and Values." In Mathematical Cultures. The London Meetings 2012-2014, edited by Brendan Larvor, 189–214. Cham: Springer International Publishing.

Hart, Herbert Lionel Adolphus, and Leslie Green. 2012. The Concept of Law, edited by Joseph Raz and Penelope A. Bulloch. 3rd edition. Oxford: Oxford University Press.

Kitchin, Rob, and Martin Dodge. 2011. Code/Space: Software and Everyday Life. Cambridge, MA: MIT Press.

Shapiro, Stewart. 2006. Vagueness in Context. Oxford: Oxford University Press.

Schauer, Frederick. 2013. "On the Open Texture of Law." Grazer Philosophische Studien 87(1): 197–215.

van Deemter, Kees. 2010. Not Exactly: In Praise of Vagueness. Oxford: Oxford University Press.

Waismann, Friedrich. 1945. "Verifiability." Proceedings of the Aristotelian Society, Supplementary Volume XIX: 119–150.

In the 90s, software engineering shifted from packaged software and PCs to services and clouds, enabling distributed architectures that incorporate real-time feedback from users (Gürses and Van Hoboken 2018). In the process, digital systems became layers of technologies metricized under the authority of optimization functions. These functions drive the selection of software features, service integration, cloud usage, user interaction and growth, customer service, and environmental capture, among others. Whereas information systems focused on storage, processing and transport of information, and organizing knowledge—with associated risks of surveillance—contemporary systems leverage the knowledge they gather to not only understand the world, but also to optimize it, seeking maximum extraction of economic value through the capture and manipulation of people's activities and environments.

The optimization problem

The ability of these optimization systems to treat the world not as a static place to be known, but as one to sense and co-create, poses social risks and harms such as social sorting, mass manipulation, asymmetrical concentration of resources, majority dominance and minority erasure.

In mathematical vocabulary, optimization is about finding the best values for an `objective function'. The externalities of optimization occur due to the way that these objective functions are specified (Amodei et al. 2016). These externalities include:

1 Aspiring for asocial behavior or negative environmental ordering (Madrigal 2018, Cabannes et al. 2018),
2 Having adverse side effects (Lopez 2018),
3 Being built to only benefit a subset of users (Lopez 2018),
4 Pushing risks associated with environmental unknowns and exploration onto users and their surroundings (Bird et al. 2016),[2]
5 Being vulnerable to distributional shift, wherein a system that is built on data from a particular area is deployed in another environment that it is not optimized for (Angwin et al. 2016),
6 Spawning systems that exploit states that can lead to fulfillment of the objective function short of fulfilling the intended effect (Harris 2018),
7 Distributing errors unfairly (Hardt 2014), and
8 Incentivizing mass data collection.

Common to information and optimization systems is their concentration of both data and processing resources, network effects, and ability to scale services that externalize risks to populations and environments. Consequently, today a handful of companies are able to amass enormous power.

In the rest of this provocation we focus on location based services (LBS). LBS have moved beyond tracking and profiling individuals for generating spatial intelligence to leveraging this information to manipulate users' behavior and create "ideal" geographies that optimize space and time to customers' or investors' interests (Phillips et al. 2003). Population experiments drive iterative designs that ensure

sufficient gain for a percentage of users while minimizing costs and maximizing profits.

For example, LBS like Waze provide optimal driving routes that promote individual gain at the cost of generating more congestion. Waze often redirects users off major highways through suburban neighbourhoods that cannot sustain heavy traffic. While useful for drivers, neighbourhoods are made busier, noisier and less safe, and towns need to fix and police roads more often. Even when users benefit, non-users may bear the ill effects of optimization.

Users within a system may also be at a disadvantage. Pokémon Go users in urban areas see more Pokémon, Pokéstops, and gyms than users in rural areas. Uber manipulates prices, constituting geographies around supply and demand that both drivers and riders are unable to control while being negatively impacted by price falls and surges, respectively. Studies report that Uber drivers (who work on commission) make less than minimum wage in many jurisdictions.

Disadvantaged users have developed techniques to tame optimization in their favour, e.g., by strategically feeding extra information to the system in order to change its behaviour. Neighbourhood dwellers negatively affected by Waze's traffic redirection have fought back by reporting road closures and heavy traffic on their streets - to have Waze redirect users out of their neighbourhoods. Some Pokémon users in rural areas spoof their locations to urban areas. Other users report to OpenStreetMaps—used by Pokémon Go—false footpaths, swimming pools and parks, resulting in higher rates of Pokémon spawn in their vicinity. Uber drivers have colluded to temporarily increase their revenue by simultaneously turning off their apps, inducing a local price surge, and turning the app back on to take advantage of the increased pricing.

While the effectiveness of these techniques is unclear, they inspire the type of responses that a more formal approach may provide. In fact, these responses essentially constitute adversarial machine learning, seeking to bias system responses in favour of the "adversary". The idea of turning adversarial machine learning around for the benefit of the user is already prevalent in Privacy Enhancing Technologies (PETs) literature, e.g., McDonald 2012. It is in the spirit of PETs that we attend to the optimization problem, i.e., we explore ideas for technologies that enable people to recognize and respond to the negative effects of optimization systems.

Introducing POTs

Protective optimization technologies (POTs) respond to optimization systems' effects on a group of people or local environment by reconfiguring these systems from the outside. POTs analyse how capture of events (or lack thereof) affect users and environments, then manipulate these events to influence system outcomes, e.g., altering optimization models and constraints or poisoning system inputs.

To design a POT, we first need to understand the optimization system. What are its user and environmental inputs (U,E)? How do they affect the capture of events? Which

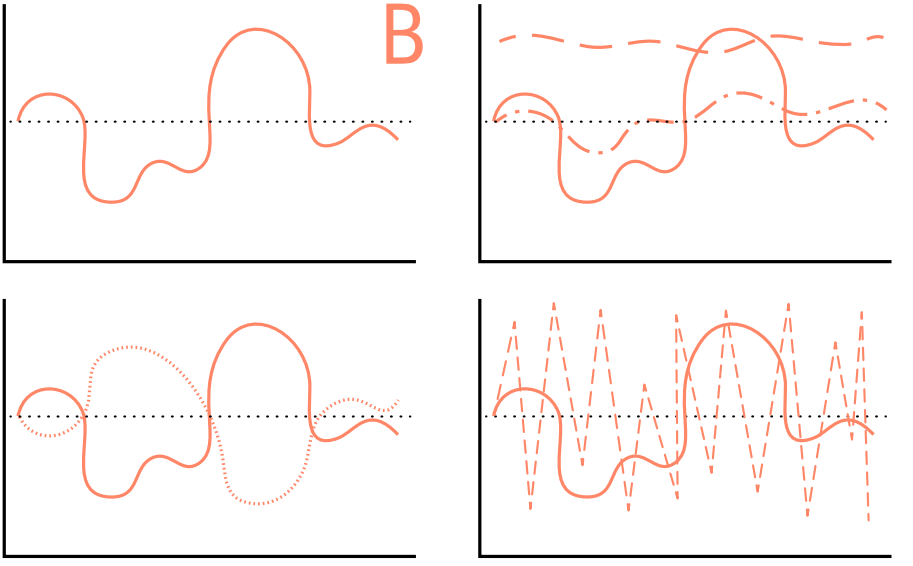


Figure 1. Benefit function (top left). POT strategies include redistribution (top right), protest (bottom left), sabotage (bottom right).

outcomes O = F(U,E) are undesirable for subpopulations or environments? With a characterization of the system, as given by F(U,E), we identify those who benefit from the system and those placed at a disadvantage by defining a benefit function, B(X, E'):(x,e', Value) → value that includes both users and non users (U ⊂ X) and affected environments(E ⊆ E'). The disadvantaged are those people and environments that reside in local minima of B and are gravely impacted by the system. We then set an alternative output B(X, E', Value'):(x,e) → value' the POT aims to achieve.

A POT's benefit function may attend to different goals (Figure 1). It may attempt to "correct" imbalances optimization systems create, i.e., by improving systems' outcome for populations put at an --often historically continuous-- disadvantage. Conversely, it may also strategically attempt to reverse system outcomes as a form of protest, highlighting the inequalities these systems engender. This further hints at the subversive potential of POTs. POT designers may concoct a strategy to produce an alternative to B to contest the authority of optimization systems, challenging the underlying objective functions these systems optimize to and their very raison d'être. To do that, a POT may attempt to sabotage or boycott the system, either for everyone or for an impactful minority that are more likely to effect change, leveraging the power asymmetries the POT precisely intends to erode.

Once we select a strategy, we must choose the techniques that implement it. These techniques involve changes to the inputs that users have control over and alterations to constraints over the objective function to reconfigure event capture (i.e., the system's

mechanism of detection, prediction, and response to events). Lastly, we deploy and assess the impact of the POT both in terms of local and global effects on users and environments and tweak it as necessary.

We note that POTs may elicit a counter response from the optimization systems they target, with service providers either neutralizing their effect or expelling POT users. Anticipating these responses may require POT designers to aim for stealth or undetectability, e.g., by identifying minimum alterations to inputs or optimizing constraints to prevent detection.

Discussion

POTs come with moral dilemmas. Some of these compare to concerns raised by obfuscation-based PETs, although these focus on protecting privacy and not protecting populations and environments from optimization. In their work on obfuscation, Brunton and Nissenbaum (2015) highlight four ethical issues: dishonesty, polluted databases, wasted resources and free riding.

Since optimization systems are not about knowledge, we may argue using POTs cannot be judged as dishonesty but as introducing feedback into the cybernetic loop to get optimization systems to recognize and respond to their externalities. POTs are likely to come at a greater cost to service providers and give rise to negative externalities that impact different subpopulations and environments. In fact, all of the harmful effects of optimization systems may be replicated: POTs may have asocial objective functions, negative side effects, etc. One may argue that if optimization is the problem, then more optimization may even come to exacerbate it. Moreover, POTs users may be seen as free riders. These are serious concerns, especially since whichever benefit function B we choose, there will be users who do not agree with or are harmed by the POT. Yet, this problem is inherent to optimization systems' externalities, especially when users are free-riding on non-users or on existing infrastructure.

Banging on POTs: A digital cacerolazo

Optimization history is also one of counter-optimization as evident in the case of search engine optimization or spammers. As optimization systems spread, POTs ensure that counter-optimization is not only available to a privileged few. One could insist that we should work within the system to design better optimization systems. Given service providers' track record in not responding to or recognizing their externalities, POTs aim to explore and provide technical avenues for people to intervene from outside these systems. In fact, POTs may often be the only way users and non-users can protect themselves and secure better outcomes. While short of a revolution, POTs bring people into the negotiations of how their environments are organized. They also help to provoke a popular response to optimization systems and their many impacts on society.

Notes

[1] We are indebted to Martha Poon for her original framing of the optimization problem and to Jillian Stone for her empirical insights into Pokémon Go. This work was supported in part by the Research Council KU Leuven: C16/15/058; the European Commission through KU Leuven BOF OT/13/070 and H2020-DS-2014-653497 PANORAMIX; and, generously supported by a Research Foundation - Flanders (FWO) Fellowship.

[2] We disagree with this paper's premise that optimization systems will lead to 'optimal' outcomes, with experimentation as its only potential externality – we appreciate their highlight of the latter.

References

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mane. 2016. "Concrete problems in AI safety." arXiv:1606.06565.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchne,. 2016. "Machine bias." ProPublica, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bird, Sarah, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. 2016. "Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI." Workshop on Fairness, Accountability, and Transparency in Machine Learning.

Brunton, Finn and Helen Nissenbaum. 2015. Obfuscation: A user's guide for privacy and protest. Cambridge, MA: MIT Press.

Cabannes, Théophile, Marco Antonio Sangiovanni Vincentelli, Alexander Sundt, Hippolyte Signargout, Emily Porter, Vincent Fighiera, Juliette Ugirumurera, and Alexandre M. Bayen. 2018. "The impact of GPS-enabled shortest path routing on mobility: a game theoretic approach". Presented at Transportation Research Board 97th Annual Meeting. Washington DC, USA. 7–11 January 2018. Issue 18-02304.

Gürses, Seda and Joris van Hoboken. 2018. "Privacy after the Agile Turn." In Cambridge Handbook of Consumer Privacy, edited by Evan Selinger, Jules Polonetsky, and Omer Tene, 579-601. Cambridge: Cambridge University Press.

Hardt, Moritz. 2014. "How big data is unfair?" Medium, https://medium.com/@mrtz/how- big-data-is-unfair-9aa544d739de.

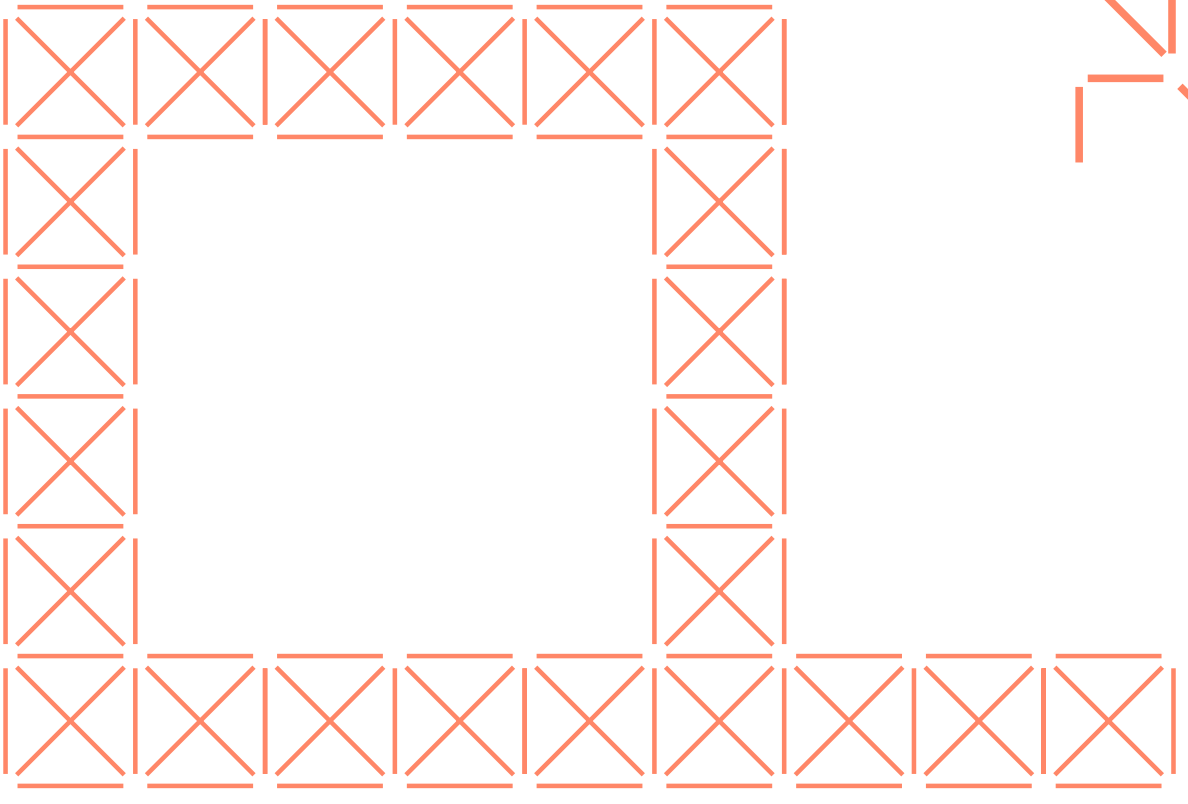
Harris, Malcolm. 2018. "Glitch capitalism." New York Magazine, April, 2018. https://nymag.com/selectall/2018/04/malcolm-harris-on-glitch- capitalism-and-ai-logic.html.

Lopez, Steve. 2018. "On one of L.A.'s steepest streets, an app-driven frenzy of spinouts, confusion and crashes." Los Angeles Times, April 2018. https://www.latimes.com/local/california/la- me-lopez-echo-park-traffic-20180404- story.html.

Madrigal, Alexis C. 2018. "The perfect selfishness of mapping apps." The Atlantic, March 2018. https://www.theatlantic.com/technology/archive/ 2018/03/mapping-apps-and-the-price- of-anarchy/555551/.

McDonald, Andrew W. E., Sadia Afroz, Aylin Caliskan, Ariel Stolerman, and Rachel Greenstadt. 2012. "Use fewer in- stances of the lettr "i": Toward writing style anonymization". In Privacy Enhancing Technologies PETS 2012, Vigo, Spain, July 11-13, 2012. Proceedings, edited by Simone Fischer-Hübner and Matthew Wright, 299-318. Berlin Heidelberg: Springer.

Phillips, David and Michael Curry. 2003. "Privacy and the phenetic urge." In Surveillance as Social Sorting: Privacy, Risk, and Digital Discrimination, edited by David Lyon, 137-153. New York: Routledge.



ON THE POSSIBILITY OF NORMATIVE CONTESTATION OF AUTOMATED DATA-DRIVEN DECISIONS

EMRE BAYAMLIOĞLU

Theorising transparency to see automated decision-making systems "at work" is a territory ever expanding as we attempt to map it (Leese 2014; Burrell 2016). The opacities and informational asymmetries inherent in machine learning (ML) result in a "mental invisibility" on the side of individuals that may only be counteracted through a visibility of different type. For the purposes of normative contestation, e.g. the one provided under Article 22 of the GDPR, this visibility should be an 'actionable transparency', an instrument to an effective and practical enforcement of rights (Hildebrandt 2017). Based on this, the provocation in hand proposes a follow-up on Ruben Binns's premise that 'algorithmic decision-making necessarily embodies contestable epistemic and normative assumptions' (2017, 4). The aim is to provide a systemisation of transparency requirements that enables the contestation of automated decisions, based on a 'reconstruction' of the system as a regulatory process containing different types of 'normativity'.

Normativity as a key to understand automated decisions

Regulatory systems are goal-oriented. Their behaviour may eventually be attributed to the values and assumptions that are implied in the rules and standards which guide the systems' response to a given input. This allows us to expect a related 'normativity' in the system's output. Since, by themselves, facts (input data) cannot provide 'reasons for action' (Raz 1979), looking through the lens of normativity informs us about the decisional criteria (norms) underlying the system, and thus opens the way to a rule-based (normative) evaluation of the observed behaviour/action.

Accordingly, challenging the truth claim or the accuracy of a decision, thus contesting 'what ought to be' in a given situation, will initially require a conceptualisation of the outcome as the result of a 'rule-based' process where certain input is rightfully matched with certain results— akin to a legal system where rules (norms) are applied to facts (input data) to make decisions (output data). In the context of automated decisions based on personal data processing, this would refer to how and why a person is classified in a certain way, and what consequences follow from that. As Leenes noted in *Profiling the European Citizen*: '[…]in the case of automated decision making about individuals on the basis of profiles, transparency is required with respect to the relevant data and the rules (heuristics) used to draw the inferences. This allows the validity of the inferences to be checked by the individual concerned, in order to notice and possibly remedy unjust judgements' (Leenes 2008, 299).

Rule-based modelling (RbM): reverse engineering the 'normativity' in machine learning

A 'rule-based explanation' of a decision means that given certain decisional ("factual") input data, the decision (output data) should be verifiable, interpretable, and thereby contestable with reference to the rules (normative framework) that are operational in the system. Following from above, the concrete transparency requirements of such a model entail an "explanation" about the following aspects of the system, to redefine it as a regulatory process:

Features as decisional cues: Any normative contestation will start with the knowledge

of what the system relies upon about the world in order to make decisions. This requires a perspective which treats the concept of "data" not as a tool of insight, but simply as certain representational or constructed input for decisional purposes.

In a ML process, data instances exist as variables of descriptive features where each feature such as age, height and weight is a dimension of the problem to be modelled (Sorelle A. Friedler et. al 2016). Depending on the nature of the analysis and the type of data available, features may also contain more constructed and computed representations such as one's habit of eating deep-fried food, educational level, speaking a dialect, or the level of intimacy between parties of a phone conversation. Features as decisional cues refers to the totality of the relevant data representations extracted from a set of variables. In case of personal data processing, a feature space maps how people will be represented as inputs to the algorithm. The objective of a ML model is the identification of statistically reliable relationships between the feature variables and some target variable (e.g. healthy or not, or at least 70% healthy). The features that a system infers to be significant and their relevant weightings help us understand which inputs (inferences) factored into a decision to get to the final result.

Normativity: Normative contestation of automated decisions can be based on two grounds, scrutinising two different types of 'normativity'. First, decisions may be contested on the basis of the selection and construction of the relevant features that the decision relies upon. What is questioned here is whether inferences made by way of selected features are sufficiently informative and causally reliable for the given purpose, e.g. whether one's search for deep-fryers suffices for the inference of one's eating deep fried food, and consequently being classified as risky. The normativity of decisional cues (features) lies in their being formal constructions by way of if-then rules. Both the accuracy and suitability of the features together with the methodology used for their selection and construction could be subject to normative scrutiny.

Second, normativity operates as a set of rules (decisional norms) for the determination of the ensuing effects. Decisional norms describe how a certain ML outcome (target value) is translated into concrete consequences in a wider decision-making framework, e.g. a certain health risk resulting in an increased insurance premium in an automated health insurance system. The question is: what is the meaning of the target variable(s) obtained? For instance, what score (in numeric or other quantified form) would suffice for a successful loan, and most importantly why? This type of scrutiny eventually reaches back to the goals and values encoded in the system, together with the underlying assumptions and justifications (ratiocinations).

The 'context' and further consequences: To fully evaluate the automated decisions for the purposes of contestation, the context of the decision—the particular situation, environment or domain in which the decision is to be made—is a key piece of information. This primarily involves informing of the data subject about where the decision starts and ends, and whether the system interoperates with other data processing operations. Accordingly, which other entities and authorities are informed of the decision; and for what other purposes or in which other contexts the results

could be used, are all crucial for a normative assessment. More importantly, the implementation of a transparency model, with contestation in mind, requires not only the knowledge of why a decision was made but also 'why a different decision was not made' (Miller 2017; Lipton 2004).

Responsible actors: This is an essential component of an actionable transparency model, meaning that the implications of automated decisions must be situated and analysed in an institutional framework, revealing the parties and the interests behind the decisions. The 'agency' behind automated decisions is not necessarily monolithic but often related to a plethora of conflicting, competing and partially overlapping interests and objectives which are linked to multifarious commercial frameworks and stately functions. This highly fragmented and obscure landscape requires a purposeful mapping of the institutional structures and the intricate web of relations among those who may be responsible for different parts or aspects of a decision, i.e. the data brokers, public and private clients, service providers, regulators, operators, code writers and system designers. Lacking this particular dimension, the transparency model remains incomplete.

Impediments and pitfalls

Both the determination of the decisional cues and the ensuing results are normative undertakings which, in theory, may be reconstructed in the if-then form (if condition1∧condition2∧condition3, then outcome). Thus, theoretically every decision that is claimed to be "rational" can be decomposed to infer which rules have been followed in what order. However, in case of automated decisions, neither the input inferred nor the rules that produce the outcome reveal themselves easily. Problems are not always as straightforward or easily verifiable as is the relation between eating habits and increased health risk—a plausible assumption based on common sense or past data.

In many cases, decisional cues do not exist as readily available features as they need to be constructed from a multi-dimensional data set. This increased dimensionality of the feature space (meaning that a great many variables are repeatedly correlated), entails that features are further selected and extracted to reduce the complexity of the data and consequently the model. In this process, physical meanings of features may not be retained, and thus it may not be possible to clarify how the final output of the system relates to any specific feature (Li 2017). The result is a set of overly constructed and computed features where correlations between feature variables and the target variable do not depend on the conventional understanding of 'cause and effect'—introducing seemingly irrelevant input. Think of e.g. using spelling mistakes for predicting overweight in a health insurance scheme, or the length of the screen name of a social media account for credit scoring. This implies that the assumed link between the input and the actual behaviour may not only turn out to be intrusive, incorrect, or invisible, but may even be non-existent due to spurious correlations. Especially in case of deep learning models, normative scrutiny of these overly constructed features may not be possible primarily because these systems have not been designed with such an assessment in mind.

A viable scheme

Based on the transparency model developed above, we propose the following set of questions as the basics of a viable contestation scheme, that may contribute to the contestability of automated data-driven decisions.

- Is the training data that was used to develop the decisional cues (input) representative of the data subject? If not, to what extent do the discrepancies matter, considering the purposes and the further impact of the decision as well as the regulatory context?
- Based on the decisional cues (selected and weighted features), are the consequences 'explainable' by providing legally, ethically and socially acceptable reasons?
- Are the results interpreted and implemented in line with the declared purposes of the system (purpose limitation principle)?
- Are data subjects made aware of how they can contest the decisions and who is liable for insufficient transparency?

Where those responsible fail to respond to these contestability requirements, their automated decisions may be regarded as *per se* unlawful (Hildebrandt 2016, 58), or as ethically questionable, depending on whether or not they violate legal norms.

References

Binns, Ruben. 2017. "Algorithmic Accountability and Public Reason" Philosophy & Technology:1-14. doi: 10.1007/s13347-017-0263-5.

Burrell, Jenna. 2016. "How the machine 'thinks': Understanding opacity in machine learning algorithms" Big Data & Society 3(1): 1-12.

Friedler, A. Sorelle, Carlos Scheiddegger, and Suresh Venkatasubramanian. 2016. "On the (im)possibility of fairness". arXiv:1609.07236v1.

Hildebrandt, Mireille. 2019 (forthcoming). "Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning". Theoretical Inquiries in Law 19(1). doi: 10.2139/ssrn.3081776.

Hildebrandt, Mireille. 2016. "The New Imbroglio. Living with Machine Algorithms" In The Art of Ethics in the Information Society: Mind You, edited by Liisa Janssens, 55–60. Amsterdam: Amsterdam University Press.

Koops, Bert-Jaap. 2013. "On Decision Transparency, or How to Enhance Data Protection after the Computational Turn." In Privacy, Due Process and the Computational Turn, edited by M. Hildebrandt & K. De Vries, 196-220. Abingdon: Routledge.

Leenes, Ronald. 2008. "Reply: Addressing the Obscurity of Data Clouds." In Profiling the European Citizen: Cross-disciplinary Perspectives edited by Mireille Hildebrandt and Serge Gutwirth, 293-300. Dordrecht: Springer.

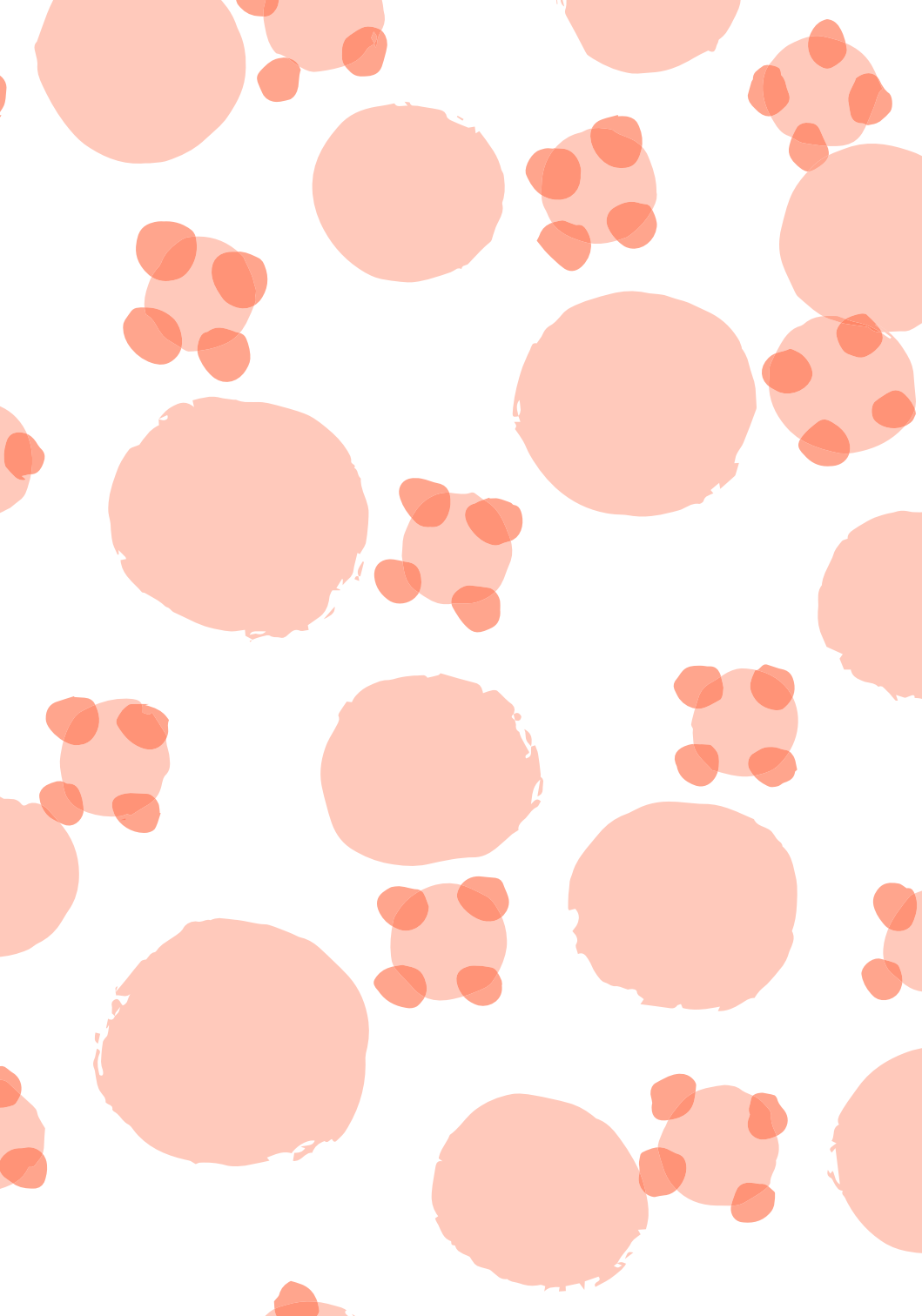
Leese, Matthias. 2014. "The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union" Security Dialogue 45(5): 494-511.

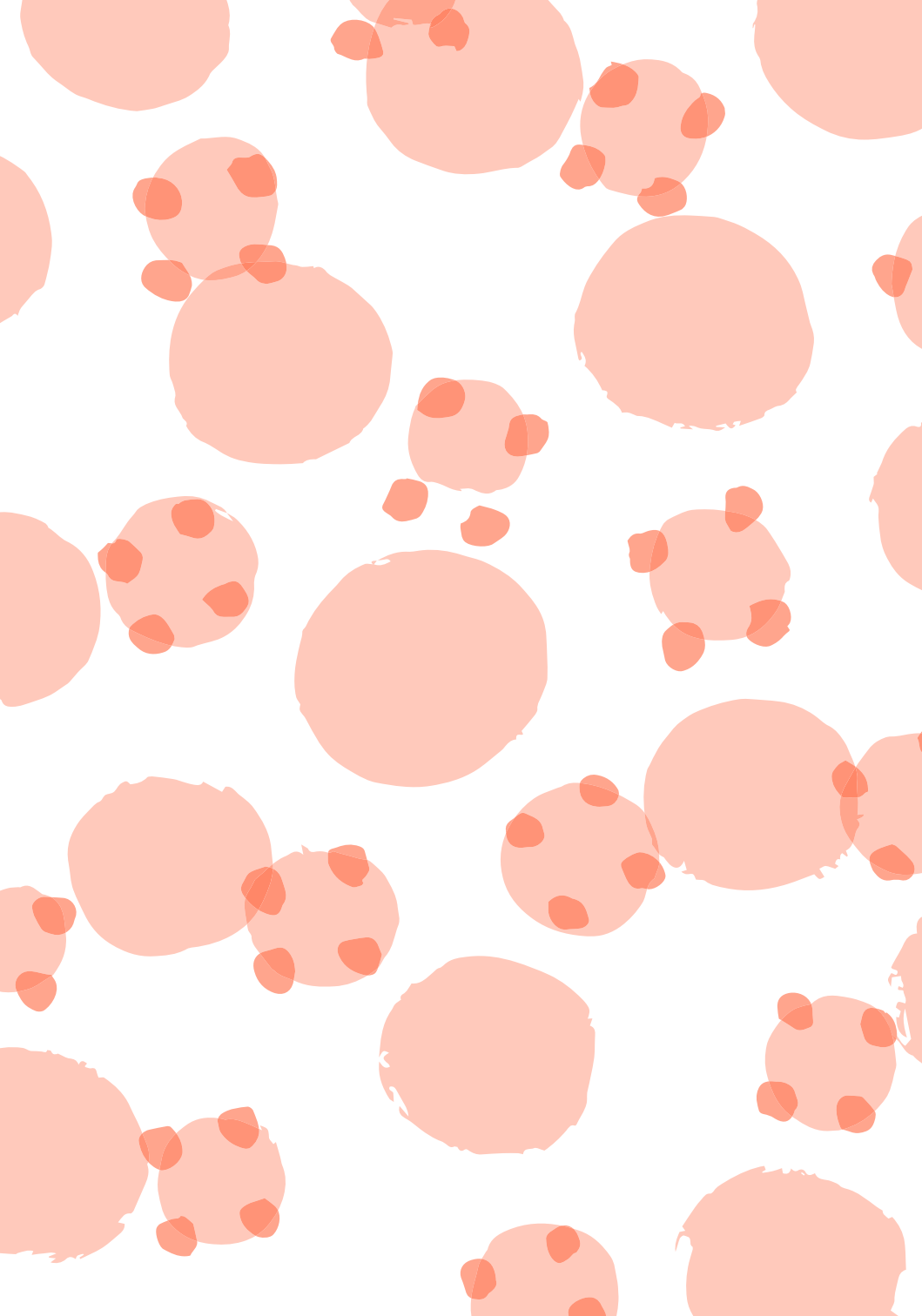
Li, Jundong, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. 2017. "Feature Selection: A Data Perspective." ACM Computer Surveys 50, 6, Article 94. doi: 10.1145/3136625.

Lipton, Peter. 2004. Inference to the Best Explanation, London: Routledge.

Miller, Tim. 2017. "Explanation in Artificial Intelligence: Insights from the Social Sciences". arXiv:1706.07269.

Raz, Joseph. 1979. The Authority of Law. Oxford: Clarendon Press.





HOW IS 'TRANSPARENCY' UNDERSTOOD BY LEGAL SCHOLARS AND THE MACHINE LEARNING COMMUNITY?

KAREN YEUNG AND ADRIAN WELLER

Algorithmic decision-making systems are increasingly in use yet often lack transparency. The opacity of these 'black boxes' leads to decisions that can be hard to understand and contest, creating substantial risks of injustice. ML researchers are developing methods for improving the transparency of these systems ('explainable AI'). Unless this 'challenge of transparency' (Weller 2017) can be addressed appropriately, alongside concerns including reliability, 'fairness' and 'algorithmic accountability', the public is unlikely to trust these systems (RSA 2018), despite their many benefits.

Integrating legal scholarship with ML research

For lawyers, the challenge of transparency is familiar for *human* decision-makers, particularly for decisions by public authorities. Within contemporary constitutional democratic orders, governmental decision-makers must exercise their authority in accordance with law. Contemporary equality legislation is also concerned with preventing and remedying decision-making that is unfairly discriminatory in relation to 'protected' grounds (gender, race, etc.). The law imposes various constraints to address and prevent particular kinds of flaws in human decision-making. These constraints are ultimately grounded in recognition that decision-making authority is vulnerable to corruption and abuse. Transparency is critical for ensuring that decision-making is lawful and accountable. Since the advent of computerised decision-making systems, various jurisdictions have introduced legally enforceable duties, entitling those directly and significantly affected by certain kinds of fully automated decisions to receive an explanation for that decision, although the precise nature of this duty is uncertain.

Within debates about what transparency in machine decision-making requires, many terms are employed by different disciplines, leading to significant potential confusion. Accordingly, we seek to clarify various concepts and terms used in discussions about transparency in decision-making, focusing on the legal and ML communities. We consider why transparency matters to these two communities, aiming to improve cross-disciplinary insight. Because this entails sweeping generalisations, our reflections are offered as heuristics, seeking to capture the kinds of concerns that are frequently raised, thereby facilitating enhanced interdisciplinary understanding and dialogue.

Why transparency matters

For both the legal and ML communities, the needs for transparency are highly context-dependent. In ML, transparency is typically desirable for understanding both specific algorithmic behaviour and the broader socio-technical environment in order to consider how the system will be used in practice. For systems that rely upon data processing to generate decision outputs, transparency is also desirable for the datasets themselves: identifying which data is used, who decides this, and other questions about the data's provenance such as source, volume, quality and pre-processing (Gebru et al 2018). In relation to the computational component of the system, identifying what transparency requires is a function of its context and the character, capacities and motivations of the intended audience (Weller 2017). For example, developers typically want to understand how their overall system works, thereby enabling them to identify and rectify any problems and undertake improvements to system performance.

In contrast, individuals directly affected by a machine decision may be concerned with how and why a particular decision was arrived at (a 'local explanation'), in order to evaluate its accuracy and fairness and to identify potential grounds to contest it. Different types of explanation might be appropriate for the affected individual, or for an expert or trusted fiduciary agent.

For human and organisational decision-making, lawyers also recognise the importance of context in identifying what transparency requires. Transparency concerns can be understood as grounded in the requirements of the contemporary concept of the rule of law, which captures a set of normative ideas about the nature and operation of law in society (Craig 1997). One of the rule of law's core requirements is that the laws themselves should be transparent: laws should be publicly promulgated (Fuller 1964, 49) so that all legal subjects can know the law's demands in advance and thus alter their behaviour accordingly. The existence of 'secret' laws of which legal subjects are unaware and could not reasonably have discovered is the antithesis of the rule of law ideal, its tyrannical consequences vividly depicted in Kafka's The Trial (Kafka 1998). The argument made by Schreurs et al. (2008), that data subjects should have access to the knowledge and potential secrets implied in the profiles that are applied to them when they match the criteria of a profile (including in private settings) 'in order to anticipate the actions and decisions that may impact our later life' is a specific application of this general principle applied to automated data-profiling.

The rule of law also requires that the exercise of power by public authorities has a lawful basis. Transparency is necessary to evaluate whether a decision is lawful, and therefore legally justified. Legal justifications typically require explanations. An explanation is typically comprised of the provision of reasons in response to the question: why did you decide that? These reasons, including the factors that were taken into account by the decision-maker, how much weight they were given, and how the totality of relevant factors was evaluated to arrive at a decision, would constitute such an explanation. Justification and explanation are different – an explanation may not, in itself, establish that a decision is legally justified. To justify a decision, the explanation must meet the criteria laid down by law, thereby establishing that the decision-maker had legal authority to make the decision, that no legally impermissible factors were taken into account, and, at least in relation to decisions made by public officials, that the legal conditions that constrain how the decision-making process is conducted were complied with, and whether the substantive decision itself falls within the bounds of legal acceptability (the terminological touchstone for which will vary between jurisdictions – in English administrative law, for example, this requires that the decision must not be 'so unreasonable that no reasonable decision-maker would have arrived at it' – the test established in the famous Wednesbury case). In short, an explanation is necessary but not sufficient for establishing that a decision is legally justified.

Even if a decision cannot be legally justified, it might nonetheless be lawfully excused. This distinction is significant: a justified decision entails no wrongdoing (Hart 1968); a decision or action that is not legally justified might still be lawfully excused, thereby

reducing the seriousness of the wrong when considered in the law's response. For example, consider the case of 95-year old Denver Beddows. He repeatedly hammered his wife's head and struck her with a saucepan, despite his lifelong devotion to her, intending to respect her continual requests that he end her life following the deterioration of her health. He was convicted for attempted murder but, owing to the circumstances of the case, was given a suspended sentence in recognition of the moral context and significance of his actions (The Independent 2018). This example points to the crux of why explanations matter: as moral agents, we want not only to understand ourselves as rational actors who can explain our actions by reference to reasons (Gardner 2006) but we also want to understand why we have been treated in a particular way by reference to reasons, in terms that we can comprehend. Only then can we can evaluate, both legally and morally, whether that treatment was justified or otherwise excused. Accordingly, if computational systems make decisions that significantly affect us, we rightly expect – as a community of moral agents in a liberal democratic society – that those decisions can be explained by reference to reasons that are intelligible to us, thereby enabling us to evaluate whether the decisions were legally and morally justified.

Terminology

Transparency intersects with many related concepts, which are sometimes used interchangeably. To help avoid confusion within and across disciplines, we consider terms and their relationship to each other.

a Interpretability, intelligibility and transparency: Within the ML community, a distinction is increasingly made between (i) 'transparency', understood as the ability to inspect the inner details of a system, for example by seeing the entire code, and (ii) 'interpretability', in the sense of intelligibility to an individual so she can understand why a particular output was generated, in terms that she can comprehend.

b Information, reasons and explanations: Rendering any decision-making system intelligible to those directly affected by the decisions which it generates will typically require the provision of the underlying reasons why it was reached. For lawyers and legal scholars, providing reasons is distinct from providing information. As legal philosopher, Joseph Raz puts it:

> *Whatever provides a (correct) answer to questions about the reasons why things are as they are, become what they become, or to any other reason-why question, is a Reason….What is important is the distinction between providing (or purporting to provide) information ('It is 4 pm', 'She is in Sydney') and providing (or purporting to provide) explanations. Reasons provide explanations. (Raz 2011, 16)*

In short, explanations require reasons. Raz explains that explanations may be relative to the person(s) for whom they are intended. For him, an explanation is a good one if it explains what it sets out to explain in a way that is accessible to its addressees, i.e. in a way that the addressees could understand were they minded

to do so, given who they are and what they could reasonably be expected to do in order to understand it (Raz 2011, 16). Yet it is also necessary to specify what it explains in order to convey any useful information. But whether an explanation is a good one does not affect its character as an explanation. For Raz, an explanation of the nature of laser radiation suitable for university students is an explanation of laser radiation, even when addressed to primary school children (Raz 2011, 16).

c Normative reasons and justifications: Explanations are the subject of a huge body of philosophical reflection, especially for philosophers interested in 'normative reasons'. Raz argues that normative reasons are those which count in favour of that for which they are reasons: they potentially justify and require what they favour (Raz 2011, 18) although they do not always do so. For both lawyers and philosophers, justifications are particularly important, because they serve to establish that a particular action was not morally wrongful and therefore not worthy of blame or punishment (Gardner 2006). Accordingly, if a decision generated by an algorithmic decision-making system can be regarded as justified, this means that that the decision entailed no wrongdoing. For the individual who is unhappy with the decision in question, then that individual would have no basis for challenging the outcome of the decision on the basis that the wrong outcome was arrived at.

Conclusion

Further work to clarify the needs for appropriate transparency is urgently needed for legitimate and effective deployment of algorithmic systems across society. For both communities, work to improve transparency may have a cost in terms of other values such as privacy. We shall explore these themes in a longer article to come.

References

Craig, Paul. 1997. "Formal and Substantive Conceptions of the Rule of Law: An Analytical Framework." Public Law 33: 467–87.

Fuller, Lon L. 1964. The Morality of Law. New Haven: Yale University Press.

Gardner, John. 2006. "The Mark of Responsibility (With A Postscript on Accountability.)" In Public Accountability, Designs, Dilemmas and Experiences, edited by Michael W. Dowdle, 220-42. Cambridge: Cambridge University Press.

Gebru, Timmit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III and Kate Crawford. 2018. "Datasheets for Datasets." Proceedings of the Accountability, and PMLR 80, 2018. arxiv.org/abs/1803.09010.

Hart, H. L. A. 1968. Punishment and Responsibility. Oxford: Clarendon Press.

Osborne, Samuel. 2018. "Elderly man who tried to kill wife with hammer so she could avoid care home spared jail after she forgives him." The Independent, April 25, 2017. https://www.independent.co.uk/news/uk/crime/man-hammered-wife-to-kill-spared-jail-after-she-forgives-him-denver-beddows-olive-a7702076.html.

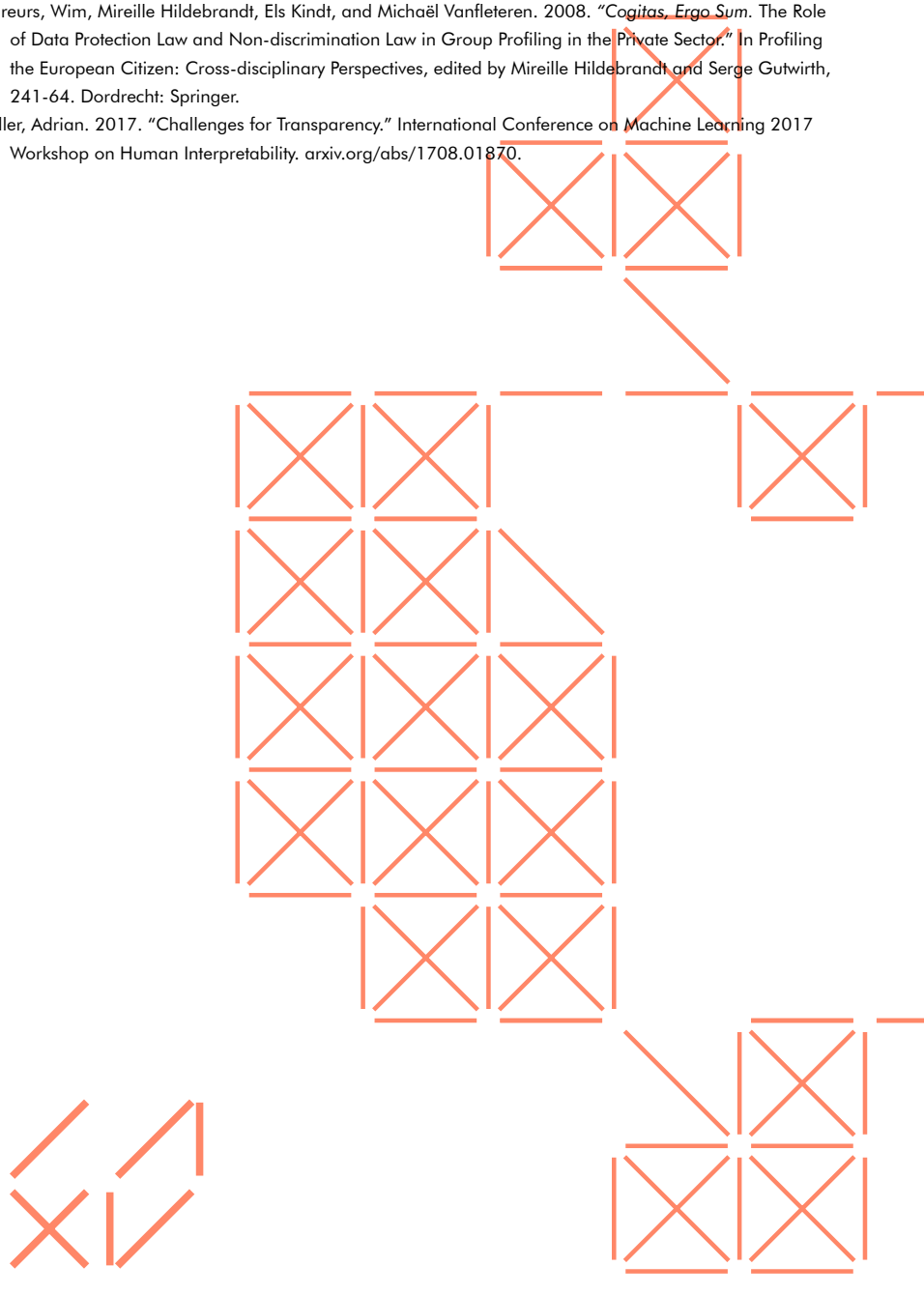
Kafka, Franz, and Breon Mitchell. 1998. The Trial: A New Translation, Based on the Restored Text. New York: Schocken Books.

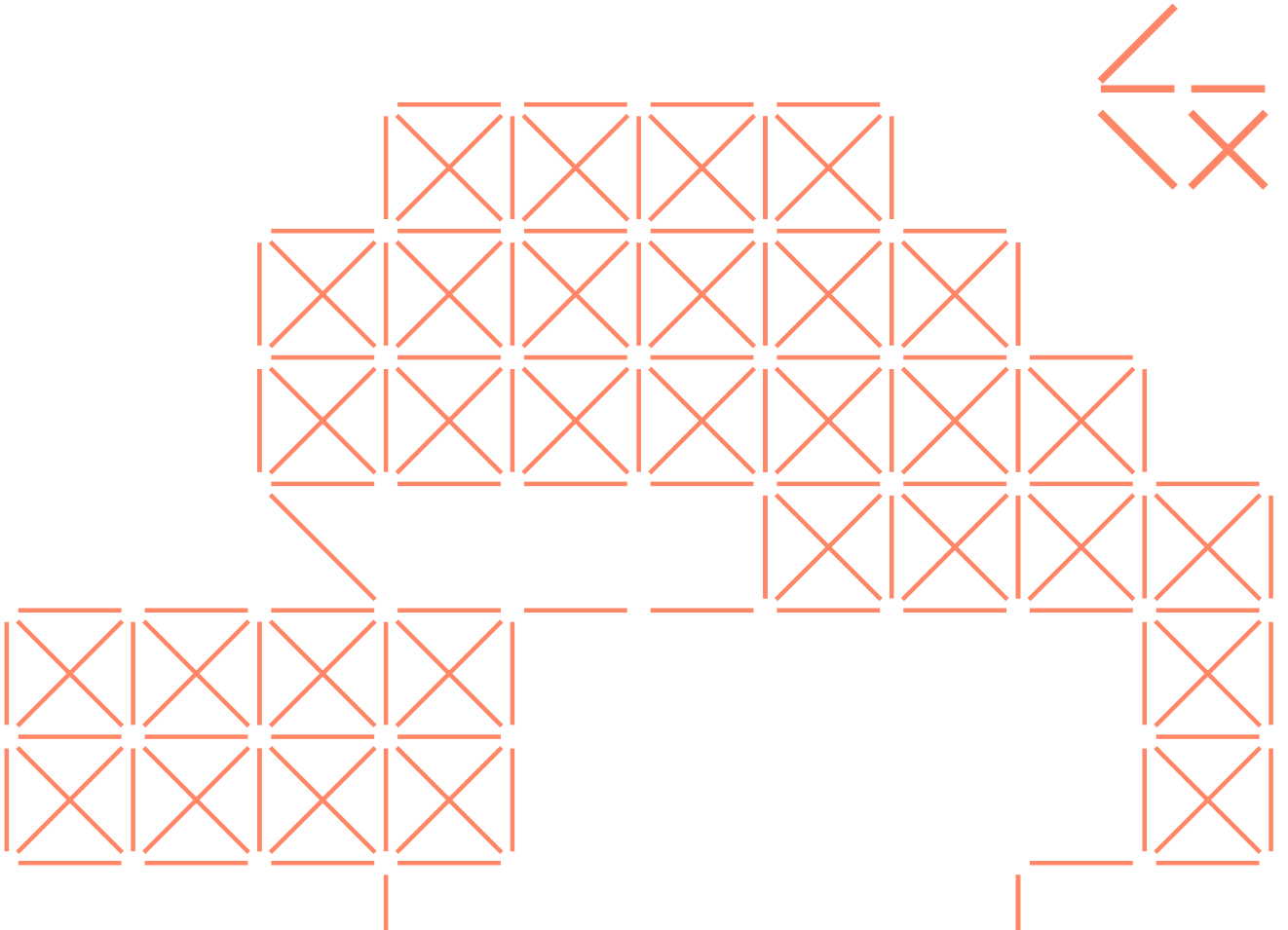
Raz, Joseph. 2011. From Normativity to Responsibility. Oxford: Oxford University Press.

RSA. 2018. "Artificial Intelligence: Real Public Engagement." May 31, 2018. https://www.thersa.org/discover/publications-and-articles/reports/artificial-intelligence-real-public-engagement.

Schreurs, Wim, Mireille Hildebrandt, Els Kindt, and Michaël Vanfleteren. 2008. "*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, edited by Mireille Hildebrandt and Serge Gutwirth, 241-64. Dordrecht: Springer.

Weller, Adrian. 2017. "Challenges for Transparency." International Conference on Machine Learning 2017 Workshop on Human Interpretability. arxiv.org/abs/1708.01870.





WHY DATA PROTECTION AND TRANSPARENCY ARE NOT ENOUGH WHEN FACING SOCIAL PROBLEMS OF MACHINE LEARNING IN A BIG DATA CONTEXT

ANTON VEDDER

Neither data protection nor transparency are effective answers to large part of the social challenges of Machine Learning in a Big Data context (MLBD). Data protection is not enough, because input and output data of MLBD need not qualify as personal data according to the definition stipulated in relevant legislation such as the General Data Protection Regulation (GDPR), nor do they have to be about human beings at all, in order to affect humans in questionable ways. Transparency falls short for another reason. Although the opacity due to technical and contextual dimensions is a basic problem for the solution of ethical and legal problems concerning MLBD (Vedder, Naudts 2017; Burrell 2016; Kroll et al. 2017), transparency can only play a role at the very first start of the deliberations. For the actual observation, articulation and solution of possible problems a broader normative framework (ethical or legal) is needed. What are the problems for which data protection and transparency do not suffice?

Problems not necessarily involving personal data

MLBD, regardless of the types of data processed, can raise problems that have to do with the redistribution of access to information. Think, for instance, of parties obtaining exclusivity concerning technologies or data involved. This may bar others from the new opportunities in a manner that may not be fair or in the interest of society as a whole. Think, for instance of monopolization tendencies with regard to MLBD in agriculture or food production, et cetera.

MLBD may, furthermore, be used to change the ownership of (in part) already existing information by extracting it and relocating it. Take for example very specific data and information concerning optimal manners of growing particular vegetables from seed to crop. Traditionally, such information may be located in the brains, the practices and the communications of specific groups of farmers. Suppose, however, that a data collector would travel to the farmers, and set up an investigation using cameras, sensors and wireless devices for monitoring the vegetables from seed to crop in combination with data about the soil and weather conditions, perform expert interviews and surveys, and in the end have a compendium on how to grow the vegetables in the literally most fruitful way. The question may arise: who should be allowed to sell it, and sell it to whom? Or, the question might become: May the compendium be used to grow the vegetables in other places and by other people than the original farmers, so that the others might start competing with them? What if the original farmers are poor and have no other means of subsistence than exploiting their original expertise?

Problems concerning group profiling but not necessarily involving personal data

MLBD searches for patterns, correlations and commonalities hidden within large datasets. The resulting information can serve as an immediate differentiation ground for discriminating, amongst others, between (groups of) individuals. MLBD can group together individuals on an aggregate level based upon previously unknown shared commonalities found within large data sets. The groups thus created, might not be easily definable, nor in real practice easily recognizable, due to their seemingly random nature.

Where such groups are involved, the resulting group characteristic will often be *non-distributive*, meaning that the characteristic is primarily a characteristic of the group, while it can only be attributed to the individual members of the group in their quality of being members of that particular group rather than to those individuals in their own right. If the latter would be the case, the characteristic would be distributive (Vedder 1999). Take, for example, a group consisting of people who happen to have a red Opel Corsa and a Jack Russell Terrier. Suppose that MLBD shows that this group – coincidentally? – runs an on average relatively high risk of a specific incurable fatal disease. Then, Mary who currently happens to possess both a red Opel Corsa and the Jack Russell Terrier will share in this characteristic. If the characteristic is non-distributive and Mary would get rid of the Opel Corsa or the little dog, or of both, she might not be considered to be at high risk anymore. If the characteristic would have been distributive, she still would have been.

The attribution of a non-distributive property can be true or false depending on the perspective of the assessor. While a person may, as a member of a seemingly random group, run a statistically high risk of developing a disease, she may as an individual in her own right be the healthiest person on earth with a health prognosis for which many would envy her. Due to the non-distributivity in this case of the "being at high risk for the disease" property, both statements – "Mary is at high risk for developing the disease" and "Mary is the healthiest person in the world with excellent health prospects" can be simultaneously true from different perspectives. Their actual use will depend on the context and the perspective of the user (Vedder 1999, 258). The notion of 'data determinism' introduced by Edith Ramirez (2013, 7-8) helps to understand this issue. Seeing and understanding the outcomes of this form of MLBD will be difficult. The groups can often only be identified by those who defined them for a specific purpose, or those who obtain the results of the MLBD directly (Vedder 1999).

MLBD involving data on human beings can very easily result directly or indirectly in discrimination in the sense of prejudicial treatment or judgement. Over the last years, many scholarly works in law and ethics have been dedicated to intentional and unintentional discrimination by MLBD on the traditional grounds for unlawful discrimination: race/ethnic background, gender, sexual orientation, religious/ideological background, political convictions, health condition et cetera (Le Métayer and Le Clainche 2012; Barocas and Selbst 2016; Diakopoulos 2016; Kroll et al. 2017). What has received relatively little attention is that MLBD has the inherent potential to provide for a plethora of new grounds for discriminating among individuals and groups. Not only, however, is it difficult to recognize the exact grounds for differentiation or the definition of the groups distinguished. It often is also difficult to exactly understand the possible unfair or discriminatory character of the attribution of characteristics to individuals and groups if these do not coincide directly or indirectly with the traditional grounds for discrimination mentioned in law (Vedder 2000; Naudts 2017). What many people may intuitively grasp, however, is that adverse judgement (for instance stigmatization) or treatment based on the mere membership of a group or category of people comes very close to the traditional phenomena of discrimination, while judging or treating persons adversely on the basis of a group characteristic that is unknown

to themselves or – due to its non-distributiveness – is dependent on their (seemingly random!) circumstances rather than on themselves in their own right may be unfair and go against the individuality of persons as a fundamental value in its own right.

Finally, group profiles may contain previously unknown and undesired information for all or some of the members of the group involved. Take again the example of Mary. Suppose, Mary happens to have a special interest in epidemiology. Her favourite journal publishes a special issue on the prevalence of as yet incurable fatal diseases. Mary reads about the remarkable discovery of people having a red Opel Corsa and a Jack Russell terrier who run a relatively high risk of developing the incurable fatal disease. Mary looks at the dog, then, through the window, at her car, and trembles…

Ever since the rise of predictive medical testing and screening in the 1980's, patient law in many countries provides people with a right not to know unintended side-results of testing and screening. Should people be protected in similar ways against the exposure to possibly undesired information that results from BDA or should bad news disclosure by MLBD be considered as mere collateral damage?

Distributive group profiles can sometimes qualify as personal data and therefore fall within the scope of data protection laws. This is also the case with non-distributive profiles as soon as they are applied to demonstrable individual persons. Concerning the latter, one should be aware that application to individuals is often not part of the automatic process itself, but an additional step in which humans interpret the outcomes of that process and take decisions on the basis of it. If data protection laws apply, these may provide legal solutions for the problems mentioned. [1] When group profiles cannot be considered as personal data, the problems remain and must be dealt with in another manner.

Broader framework

In the GDPR, transparency is defined as a responsibility of the controllers towards the legislator and towards the data-subject. The responsibilities towards the data-subject receive most attention and a high degree of specification.[2] Of course, this will make some sense in the case of MLBD involving personal data.

In MLBD without personal data, such an obvious addressee is lacking. More importantly, as may have become clear in the previous sections, complexity of the underlying technological process is only one issue to grapple with. The perspective-dependence of the recognisability of profiles is another, while the involvement of a very broad set of possibly relevant values, rights and interests, ranging from fair access to the MLBD infrastructure and information, over individuality and justice, to a right not to know and the rights to data protection and privacy is further adding to the difficulties of finding a satisfactory approach. For that approach transparency is not the end, it is just the beginning.

Given these particularities, a regulatory regime that in the first place enables deliberations about the possible impacts on humans would be desirable. Such a regime

could lay down a basis for those deliberations by assigning accountabilities to the parties that could be identified as the controllers of the BDA data processing. It could stipulate mechanisms of for instance processing records, impact assessments, transparency rules, and obligations to report to data authorities – not merely *personal* data authorities. In order to effectively identify possible rights, legitimate interests, and values affected by data processing activities, broadly composed authorities seem to be called for. In order to help especially with the articulation of possible moral and legal problems, a broadly composed authority should not only consist of representatives from various possible stakeholders, such as corporations and NGOs, but also of ethicists and lawyers.

Notes

[1] The relevant right not to be subject to automated decision-making (Art. 22 GDPR; Recital 71 GDPR) cannot be discussed in this paper for reasons of conciseness.

[2] Art. 12 GDPR.

References

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact" California Law Review 104: 671–732.

Burrell, Jenna. 2016. "How the machine 'thinks': Understanding opacity in machine learning algorithms" Big Data & Society 3(1): 1-12.

Diakopoulos, Nicholas. 2016. "Accountability in Algorithmic Decision Making" Communications of the ACM, 59(2): 56-62.

Le Métayer, Daniel, and Julien Le Clainche, 2012. "From the Protection of Data to the Protection of Individuals: Extending the Application of Non-Discrimination Principles." In European Data Protection: In Good Health?, edited by Serge Gutwirth, Ronald Leenes, Paul De Hert, and Yves Poullet, 315-29. Dordrecht Heidelberg London New York: Springer.

Kroll, Joshua, Joanna Huey, Solon Barocas , Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. "Accountable Algorithms." University of Pennsylvania Law Review 165(3): 633-705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/.

Naudts, Laurens. 2017. "Fair or Unfair Differentiation? Luck Egalitarianism as a Lens for Evaluating Algorithmic Decision-making." Data for Policy. London, 6-7 September 2017. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3043707.

Ramirez, Edith. 2013. "Keynote Address at the Tech. Policy Inst. Aspen Forum, The Privacy Challenges of Big Data." http://www.ftc.gov/sites/default/files/documents/public_statements/privacy-challenges-big-data-view-lifeguard%E2%80%99s-chair/130819bigdataaspen.pdf.

Vedder, Anton. 1999. "KDD: The challenge to individualism." Ethics and Information Technology 1: 275-28

Vedder, Anton. 2000. "Medical data, new information technologies and the need for normative principles other than privacy rules." In Law and Medicine, edited by Michael Freeman and Andrew D. E. Lewis, 441-59. Oxford: Oxford University Press.

Vedder, Anton, and Laurens Naudts. 2017. "Accountability for the use of algorithms in a big data environment." International Review of Law, Computers & Technology 31(2): 206-24.

TRANSPARENCY IS THE PERFECT COVER-UP (IF THE SUN DOES NOT SHINE)

JAAP-HENK HOEPMAN

Even though calls for transparency in a modern form go as far back as the early Age of Enlightenment (Annany & Crawford 2018), perhaps Louis Brandeis can be considered the father of 'transparency theory' because of this famous quote (Brandeis 1914):

> *Publicity is justly commended as a remedy for social and industrial diseases. Sunlight is said to be the best of disinfectants; electric light the most efficient policeman.*

Indeed, transparency is commonly advocated as an important tool to counter the ill effects of automated, data driven, decision-making (Hildebrandt & Gutwirth 2008; Pasquale 2015).

Now Brandeis never used the term 'transparency' itself, but if we read publicity as transparency, I cannot fail to wonder: what if the sun does not shine?... What if we all lived in glass houses but there is no light to see inside? Wouldn't that render transparency useless? Indeed, wouldn't that turn transparency into a perfect cover-up, allowing organisations to hide in plain sight, pretending not to be engaged in any nefarious activities?

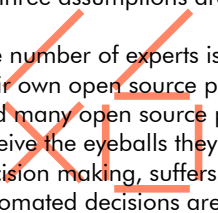
Do many eyeballs make bugs shallow?

It is a common mantra in the open source community: 'many eyeballs make bugs shallow' (Raymond 2000). In fact, it is one of the main arguments why the source code of all software we develop should be open. By publishing the source code of the software, one allows public scrutiny of that code by other, independent, experts. Bugs (i.e. programming mistakes) will be found that would otherwise lay undetected in the source code forever. As a result, systems will become more reliable and more secure (Hoepman & Jacobs 2007). Moreover, fundamental design decisions can be challenged, possibly leading to improved designs.

However…

The mantra assumes three things. First, that an unlimited number of eyeballs, i.e independent experts, is available to scrutinise the growing pool of open source projects. Second, that these experts have an interest or incentive to spend some of their (valuable) time on this. And third, that every open source project is equally likely to attract the attention of a sufficient number of experts.

All three assumptions are unfounded.



The number of experts is severely limited. These experts may often be inclined to start their own open source project rather than contributing to someone else's project. And many open source projects remain unnoticed. Only a few, high profile projects receive the eyeballs they need. Advocating transparency to balance data driven decision making, suffers from the same set of potential problems. Systems that make automated decisions are complex, and require considerable expertise to understand

them (an issue that we will return to further on). Even if all automated decision making by all organisations is done in a transparent way, there will always be only a limited number of experts that can scrutinise and challenge these decisions. Which decisions will actually be challenged depends on the incentives; again, we cannot be sure high-profile cases are likely to attract the attention they deserve.

Transparency by itself is useless without agency

Let's assume transparency works in the sense that 'bugs', i.e. improper data driven decisions, come to light and people want to act. Transparency by itself does not allow them to do so, however. The situation also requires agency, i.e. the ability to address and redress the problem. (Note that for exactly this reason a large class of open source software is in fact *free*, as in free speech. This allows anyone *with the necessary technical capabilities* to change the source code, fix whatever bug they find, and redistribute the solution.)

In many cases you have no agency whatsoever. Computer says no, tells you why, but no matter how you try, you will not be able to successfully challenge that decision. (See Ken Roach's excellent movie "I, Daniel Blake" for a compelling illustration of this point.) This is caused by several factors.

The first, most important one, is the lack of power. A single person, wronged by a decision of a large organisation, is but an itch that is easily scratched. Even if the case involves a larger, powerful, group of subjects that are collectively impacted by the decision, or if the case is taken over by a powerful consumer organisation or a fancy law firm, one would still need laws and regulations that create a (legal) basis on which the decision can be challenged. Finally, the process of appealing a decision may be so cumbersome that the effort to challenge the decision may thwart the benefit of doing so. Individuals easily get stuck into bureaucratic swamps.

The 'house of mirrors' effect

A mirror is made of glass, but it is not transparent. A house of mirrors is a seemingly transparent maze where one easily gets lost. The same problem plagues transparency theory: a decision-maker may be transparent about the decision-making process, but the description may in effect be opaque, hard to understand, hard to access/find, and/or hard to compare with others. For example, many privacy policies are overly legalistic, making them unintelligible by the average user. They are often far to long too, requiring so much reading time that no one ever reads all privacy policies of all sites they visit (McDonald and Cranor 2008).

Even if the decision-maker honestly tries to be transparent about the decision-making process and honestly aims to explain a particular decision to the subject of that decision, this explanation may still be too complex to understand. The explanation may use jargon, may depend on complex rules (if rule-based at all), and may depend on so many variables that data subjects will easily lose track. These properties of transparency may also be put into use disingenuously, to make the explanation unintelligible on purpose, while claiming to be transparent. One can observe a similar

effect in the telecommunications market where mobile phone subscription plans are complex, and where different operators use incomparable tariff plans. As a result, ordinary users have a hard time figuring out which offer suits them best (and a whole market of comparison services was born, not only for the telecommunications market, but also for the health insurance market for example).

Being transparent is hard

It very much depends on the decision-making process whether it is easy to supply a proper explanation for every decision made. In classical rule based expert systems this is certainly possible (by disclosing the rules applied and the facts/data/propositions on which they were applied), but in modern machine learning settings this is much less clear (Burell 2016). In many cases the machine learning system constructs an internal representation 'explaining' the example cases presented to it during the learning phase. But this internal representation, the model of the type of cases the algorithm is supposed to be applied to, is not necessarily close to how humans understand these types of cases and the logic they apply to decide them. A complex vector of weighing factors that represent a neural network does nothing to explain the decision made with that neural network, at least not in how humans understand an 'explanation'.

Challenging a decision is hard

Challenging a decision is hard. Even when given the explanation of the decision and the data underlying the decision, it may be hard to verify that the decision is valid. This is caused by several factors.

First of all, you need the necessary domain knowledge to understand the explanation, and to spot potential problems or inconsistencies in it. For example, to understand whether a decision in, say, environmental law is correct you need to be an expert in environmental law yourself. (This partially overlaps the first argument of the difficulty of finding and incentivising experts to challenge a decision.) Secondly, the validity of a decision depends both on the interpretation of the data on which it is based, and the interpretation of the rules used to arrive at the decision. Moreover, the selection of the rules matters a lot: it may very well be that applying a different set of rules would have led to an entirely different set of decisions. (And all this assumes that the decision-making is in fact rule based to begin with, allowing such a clear interpretation.) Thirdly, the data set may be so large and the model used to 'compute' the decision so complex, that even a basic verification of the consistency of the decision itself (let alone any complex 'what-if' scenario analysis) cannot be done 'by hand' and thus requires access to sufficiently powerful data processing resources. In the worst case the problem is so complex that only the decision-maker itself has enough resources to perform such an analysis. This totally undermines the principle of independent oversight. Lastly, the explanation of the decision may be valid and reasonable, but may not be the *actual* reason for the decision. A common example is the (inadvertent) use of proxies (like home address or neighbourhood) for sensitive personal data categories like race or religion. Sometimes this happens on purpose, sometimes this is a mistake.

Transparency may conflict with other legitimate interests

Even if the system used allows for the proper explanation of all decisions made, publishing these explanations may reveal too much information about the underlying model used to arrive at the decision. Of course, that is the whole point of requiring transparency. However, certain organisations may wish to keep their decision-making logic a secret, and may have a legitimate interest for this. For example, law enforcement or intelligence agencies have every reason *not* to reveal the models they use to identify potential terrorists (for fear that terrorists will change their modus operandi to evade detection). Similar arguments apply to fraud detection algorithms for example. Business, like credit scoring agencies, may not want to reveal their models as these algorithms, these models, may be the only true asset, the crown jewels, of the company.

Conclusion

We have discussed six arguments to show that transparency *by itself* is insufficient to counterbalance the ill effects of automated, data driven, decision making. This is not to say that transparency is useless. To the contrary: the mere fact that decision-makers are forced to be transparent will make them behave more diligently most of the time. But this is not enough. We need new, stronger, models of accountability that take the above limitations of transparency into account (Annany and Crawford 2018). For transparency to work, agency is a prerequisite. We need suitably incentivised experts that can help challenge decisions. Proper enforcement of transparency requirements is necessary, to ensure that the information provided is accessible and intelligible. Using decision-making processes that are hard to explain should be made illegal. And independent verification platforms that make it possible to verify and analyse decisions based on complex models and data sets must be made available. Finally, where transparency conflicts with other legitimate interests, a clear set of principles are necessary to decide when an explanation is not required.

Because without sun, transparency is the perfect cover, hiding in plain sight what everyone fails to see.

Notes

[1] This research is partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no 732546 (DECODE) and by the Netherlands Organization for Scientific Research (NWO) as project 'Patterns for Privacy' (CYBSEC.14.030).

References

Annany, Mike, and Kate Crawford. 2018. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability." New Media & Society, 20(3): 973–89.

Brandeis, Louis. 1914. Other people's money and how the bankers use it. New York: Frederick A. Stokes.

Burell, Jenna. 2016 "How the machine 'thinks': Understanding opacity in machine learning algorithms". Big Data & Society 3(1): 1-12.

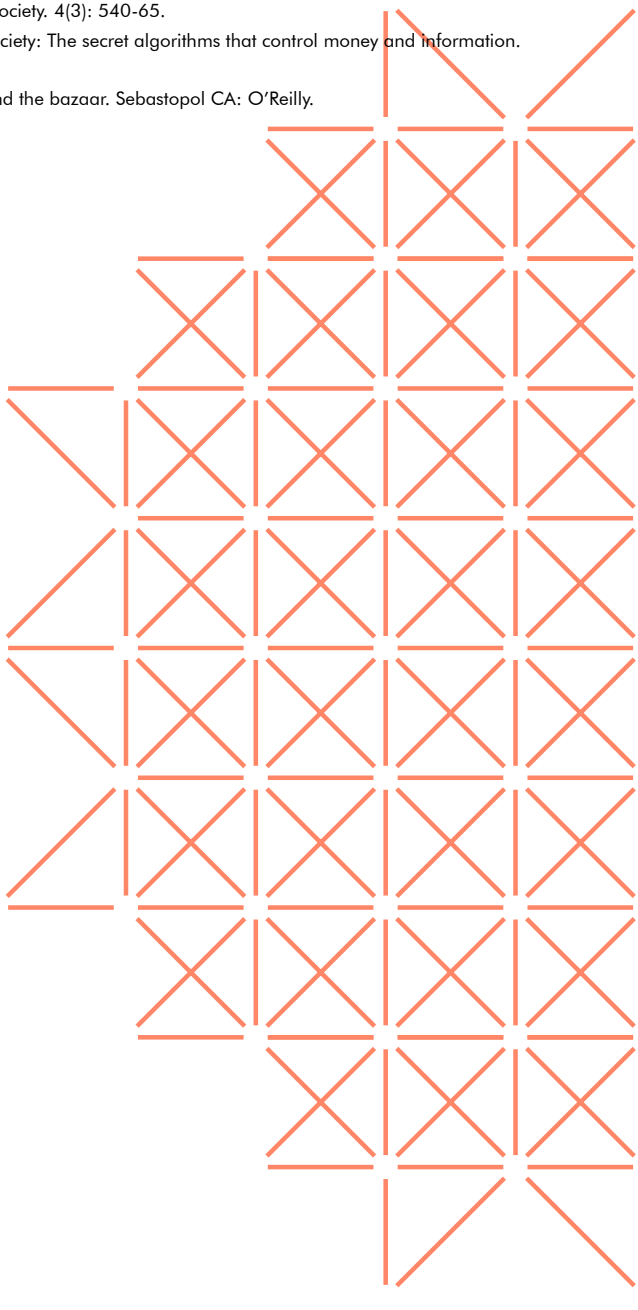
Hildebrandt, Mireille, and Serge Gutwirth. eds. 2008. Profiling the European citizen: Cross-disciplinary perspectives. Dordrecht: Springer.

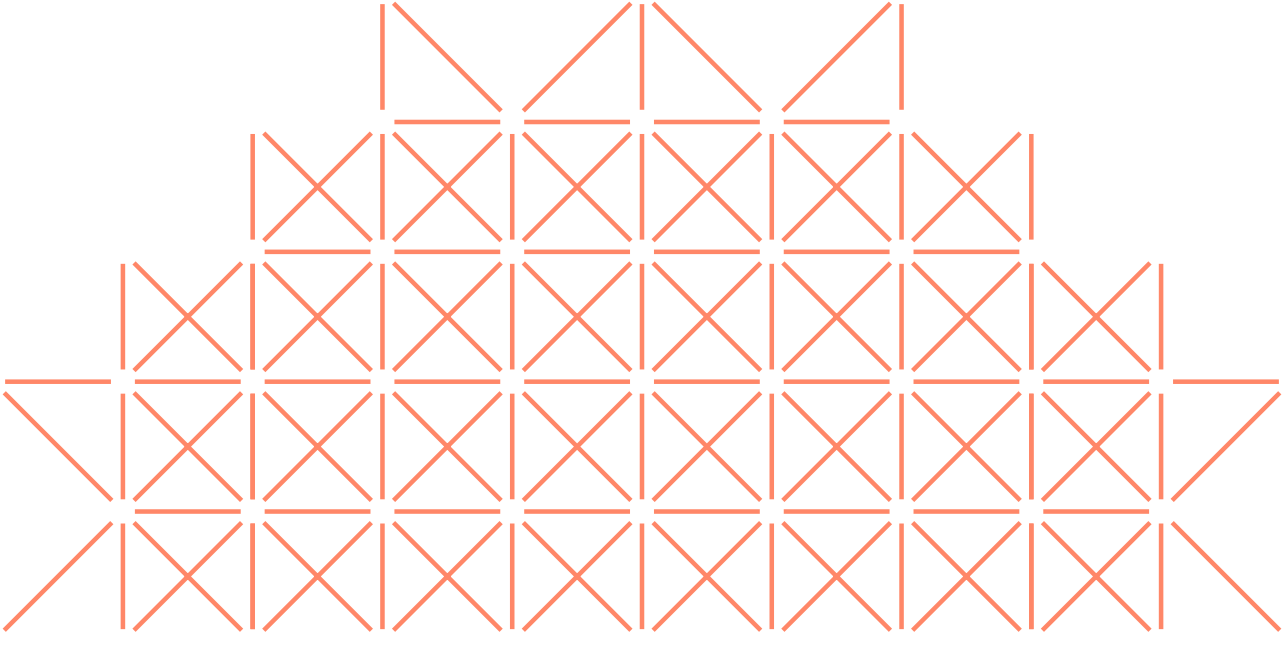
Hoepman, Jaap-Henk, and Bart Jacobs. 2007. "Increased security through open source." Communications of the ACM, 50(1): 79-83.

McDonald, Aleecia M., and Lorrie Faith Cranor. 2008. "The cost of reading privacy policies." I/S: A Journal of Law and Policy for the Information Society. 4(3): 540-65.

Pasquale, Frank. 2015. The black box society: The secret algorithms that control money and information. Boston: Harvard University Press.

Raymond, Eric S. 2001. The cathedral and the bazaar. Sebastopol CA: O'Reilly.





'Oh, I see', said the data subject. And went on to add: 'Yes, I do see why you are collecting all this information about me. I vividly visualize the data you are taking away from my hands. And I can nicely picture with whom you will share it – well, at least the type of people you might, and probably will, share it with. I am delighted now I actually know who you are. I will cherish your contact details while you process all these data, which shall however not be forever, as you somehow melancholically, but certainly accurately, have pointed out. I sincerely appreciate you are able to prove all this processing is lawful – that there is a legal ground, a good reason why this happens, and that, if there was none, all this might still be fine if I freely agree with it, on my own will. I welcome all your kind explanations about the line of reasoning behind the data-driven automatic decisions you will be taking about me sooner or later. They mean so much to me. And I am deeply touched by your efforts in describing how these decisions will make a real difference in my life. I am ecstatic hearing you talk about the existence of a series of rights I have, that I could maybe use. I can almost feel the presence of your data protection officer right here by my side'.

TRANSPARENCY AS TRANSLATION IN DATA PROTECTION
GLORIA GONZÁLEZ FUSTER

This is, perhaps, how some have come to imagine transparency obligations in European data protection law: an act of almost perfect communion between those who decide to process personal data (the 'data controllers') and the individuals linked to such data (the 'data subjects'), during which the latter get to actually see, and properly understand, what is going on with their data, why this is occurring at all, what will happen to them and their data in the near future, and what they could do about it, in case they would like to do something about it. A short moment of illumination of the nevertheless generally unaware individuals that comprise a predominantly ignorant population. The great lifting of the veil of the ever so obscure global contemporary data practices. A ray of light amidst the darkness. The joy of unravelling the precise manner in which you are being profiled. The ecstasy of personal enlightenment, in which 'being aware' (van der Hof and Prins 2008, 119) and 'opening up' (Benoist 2008, 181) are the keywords. The last hope in an increasingly in-transparent world, full of uninformed people.

Breaking open windows

Transparency, as its name suggests, could indeed be about finally being able to see through the shadows of opaque data processing operations. It could, in principle, be about revealing to data subjects the exact nature of what is really going on whenever somebody collects data about them, by bringing those ignorant individuals in direct contact, face to face, with what is happening, and what is - potentially - going to happen at some point. To finally make palpable to everybody the authentic fabric of data processing. To allow you to put your fingers into the spaces between the muscles of the algorithms shaping your existence.

In European data protection law, however, transparency is fundamentally not about a vague, utopic state of objective clarity, but about something else. It is not about letting data subjects sneak into the real life of their data and into the algorithms that move them, but about providing individuals with a certain narrative about all this processing; a narrative de facto constructed for data subjects on the basis of the interests of

the data controllers, and adapted to fit a certain idea of the data subject's presumed needs and ability to discern. At its core, transparency is indeed not about disclosing any hidden practice, or about bringing data subjects closer to anything at all, but about generating and adapting a certain data story to an imagined data reader, that is, about re-creating and triggering new accounts about data, built on some data visions. Transparency is, in this sense, about translating, and creatively transcribing and delivering to data subjects an account of what is being done to their personal data, tailored to a certain idea of what individuals might want to hear, and what they can perceive. It is about being told how you are being profiled, but in a language that inevitably betrays you were already 'being profiled' in order for controllers to decide how they would tell you about it.

The GDPR says it clearly and concisely

Concretely, transparency in European data protection law is an obligation imposed on data controllers to communicate a series of pieces of information, and to communicate them 'in a concise, transparent, intelligible and easily accessible form, using clear and plain language' (Art. 12(1) of the General Data Protection Regulation (GDPR)).[1] Beyond the tautological assertion according to which transparency is about communicating something in a "transparent" way, what the quoted GDPR provision expresses is that transparency is about making an effort to convey information in a way that is objectively short ('concise') and simple ('using clear and plain language'), but also in a manner that is subjectively and contextually adapted to the ability of the addressed data subjects to grasp its meaning, and to make some sense of it. Transparency is, in this way, about a certain reading of who is expected to read transparency notices, and a writing of such reading into the text data subjects will finally get to read.

Complying with the obligation of transparency imposes indeed on the data controller the prior obligation to determine – deliberately or not, consciously or not – who are the targeted data subjects, and what are they supposed to find intelligible and easily accessible. This therefore demands from controllers, first, to take a stand on who might be these individuals (to somehow imagine them, and speculate on their comprehension skills), and, second, to attempt to communicate in a way that presumably matches the intelligibility requirements derived from such imagined/imaginary data subjects.

In this sense, the information provided by controllers to data subjects reflects the controller's perception of the individuals whose data they are about to process; the communication of this information is shaped by such reflection, and sustains it. It is more than just pure plain language, clinically and concisely arranged in an objectively clear manner. It is not an open door towards their own data practices, or an open window into accompanying data protection safeguards. It is not a veil that is lifted, but a veil that is woven. It is a translation to the extent it is framed by the author through an invented data subject/reader, and participates in the further invention of such a subject/reader – it is a 'gesture of appropriation' (McDonald 1988, 152), and an act 'mediated and filtered through the opacity of writing' (Murail 2013).

Tell me you can read me

This translation, technically speaking, shall precede the (second) translation that comes in when the personal data processing at stake actually begins. That is the moment when the data controller can formally start building its own data construction of the individuals whose data it processes, on the basis of the data collected from them, and/or from other sources.

In practice, there is nevertheless often a temporal grey zone surrounding the moment when the data controller starts processing data, on the one hand, and the moment when 'transparent' information is given to the data subject, on the other. Although information shall, in principle, be provided 'at the time when personal data are obtained' from the data subject, it appears that some data controllers do feel entitled (and possibly obliged) to process beforehand at least some data, such as data that will help them determine in which language the data subject needs or deserves, in their view, to be told about the just-about-to-begin data processing practices and correlated data protection safeguards.

Living nearby or inside a linguistic border, and within a linguistically complex reality, it is for instance particularly easy to witness variance in automated language selection decisions, typically unilaterally taken by controllers on often persistently unclear grounds. In my personal case, for instance, the social networking site Facebook has decided I must read their 'Facebook Data Policy' in French, and thus I might repeatedly click and re-click on a link called 'Facebook Data Policy', but I will systematically be automatically directed to a page titled *'Politique d'utilisation des données'*, in French.[2] The digital music service Spotify, on the contrary, initially judged I shall rather read their Privacy Policy in Dutch, and directed me insistently to it for some time, although now it does allow me to cheat and pretend I live in the United Kingdom to access it in English, and thus be able to quote here the beautiful passage where it is stated that my privacy 'is, and will always be, enormously important' to them, and that therefore they 'want to transparently explain how and why [they] gather, store, share and use [my] personal data'.[3]

These are mere examples of choices made by data controllers to define how data subjects can learn about ongoing and upcoming processing operations that affect them and the data connected to them, illustrating that transparency is, foundationally, mediation.

A pixelated mirror in front of a pixelated mirror

Once we agree that to 'transparently explain' is to sustain a certain (pre-)conception of what data subjects need to – and can – understand, this necessarily obliges us to move beyond any simplistic debates about whether what is needed is 'more' or 'less' transparency, or about whether transparency is either 'good' or 'bad'. Transparency is not to be measured by degrees, nor to be celebrated or dismissed as such. It is not about showing, or giving access, but about interpreting and creatively rendering and supporting a certain image of targeted individuals. Transparency is not something that happens to counter the fact that individuals are being profiled, but already about

'being profiled'. Once we realize that transparency is translation, we can move out of naive metrics and binary politics of transparency, towards a critique of how it qualitatively modulates power relations between data controllers and (data) subjects.

Notes

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[2] Politique d'utilisation des données, Date de la dernière révision: 19 avril 2018, https://www.facebook.com/policy.php?CAT_VISITOR_SESSION=c7b73ebc78d1681ade25473632ea e199 [last accessed 10th June 2018].

[3] Spotify Privacy Policy, Effective as of 25 May 2018, https://www.spotify.com/uk/legal/privacy- policy/?version=1.0.0-GB [last accessed 10th June 2018].

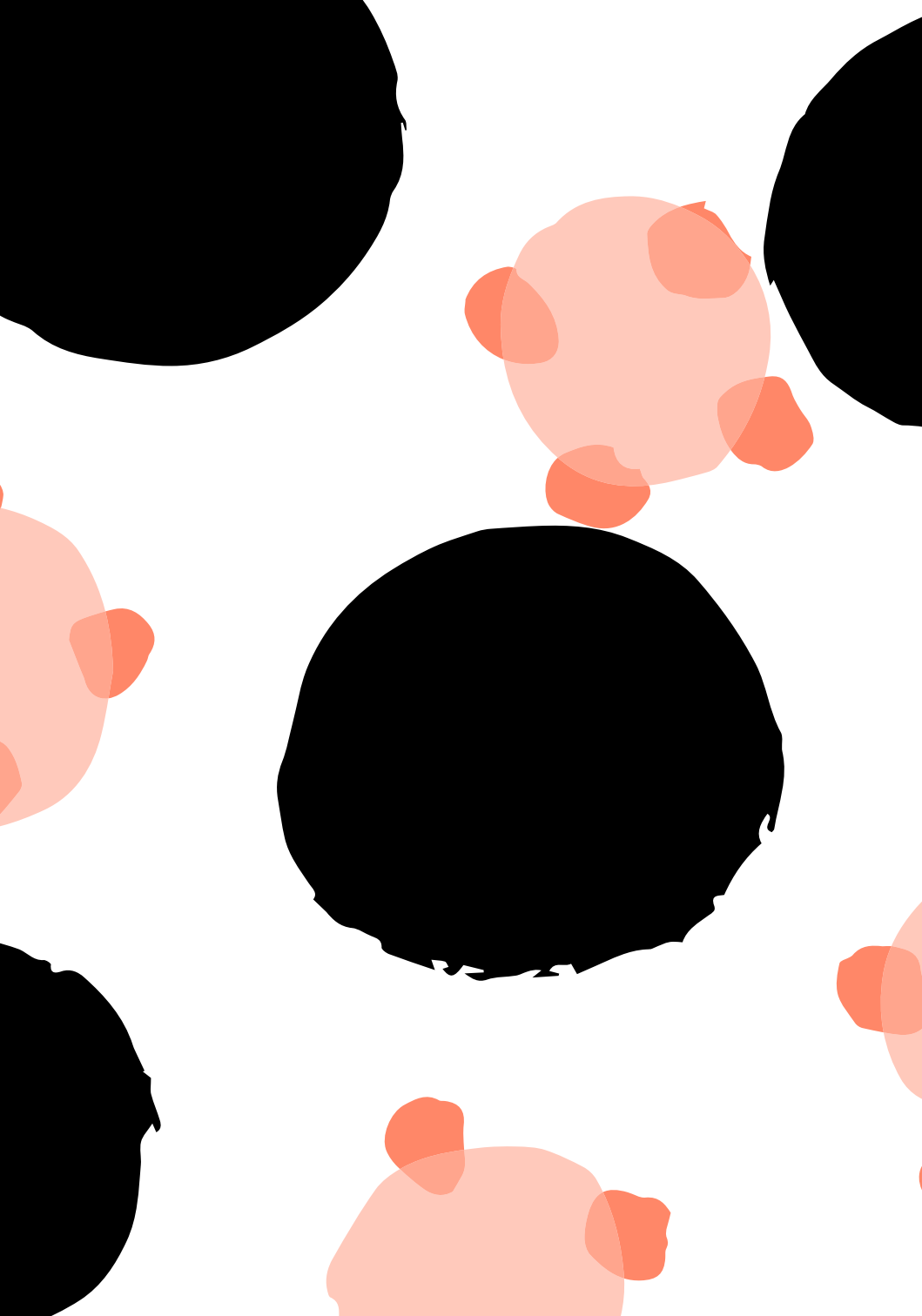
References

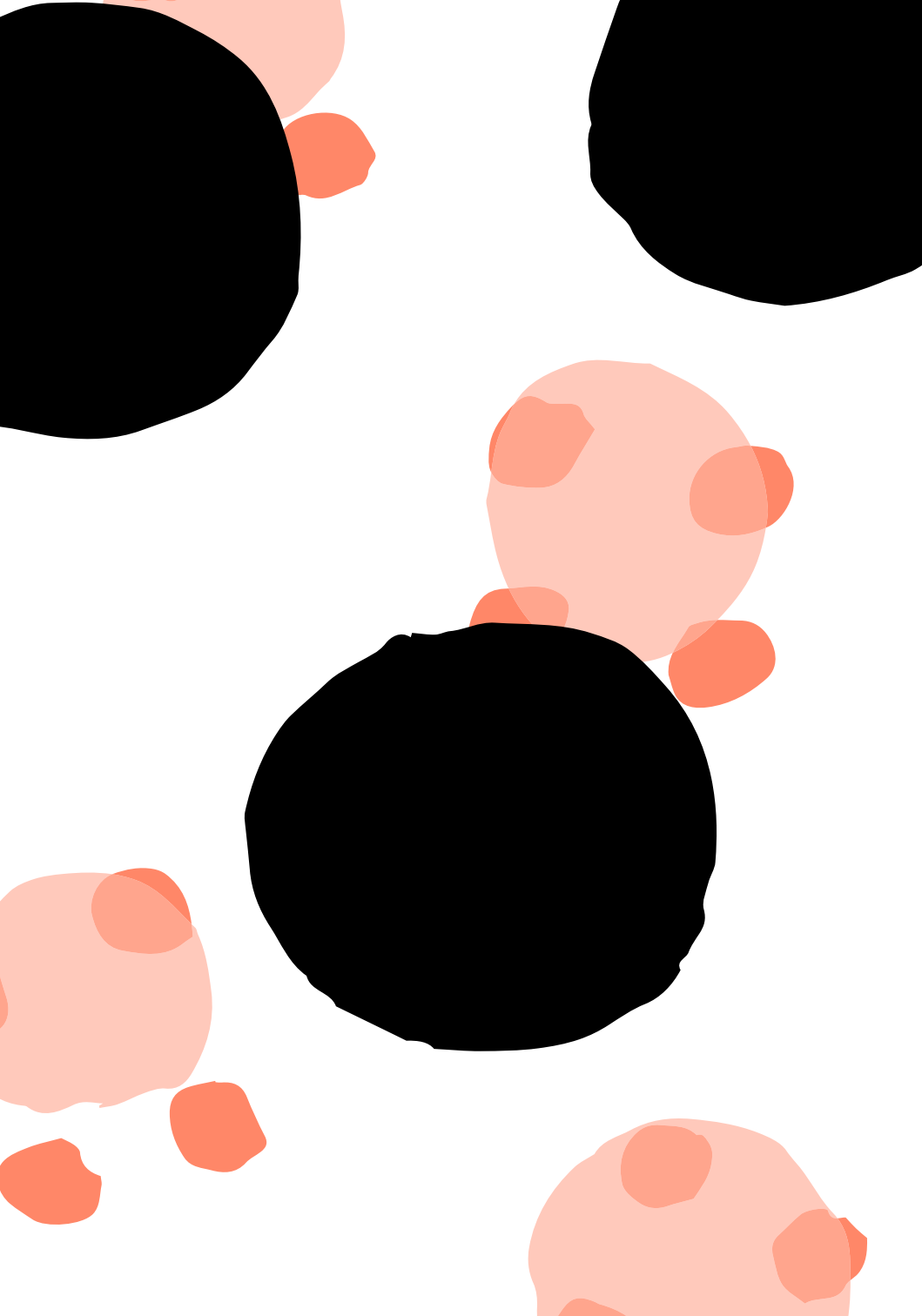
Benoist, Emmanuel. 2008. "Collecting Data for the Profiling of Web Users." In Profiling the European Citizen, edited by Mireille Hildebrandt and Serge Gutwirth, 169–84. Dordrecht: Springer.

Hof, Simone van der, and Corien Prins. 2008. "Personalisation and Its Influence on Identities, Behaviour and Social Values." In Profiling the European Citizen, edited by Mireille Hildebrandt and Serge Gutwirth, 111–27. Dordrecht: Springer.

McDonald, Christie V., and Jacques Derrida. 1988. The Ear of the Other: Otobiography, Transference, Translation; Texts and Discussions with Jacques Derrida. Lincoln: University of Nebraska.

Murail, Estelle. 2013. "The Flâneur's Scopic Power or the Victorian Dream of Transparency." Cahiers Victoriens et Édouardiens (Online) 77 (Spring). https://journals.openedition.org/cve/252.





THE PRESUMPTION OF INNOCENCE'S JANUS FACE IN DATA-DRIVEN GOVERNMENT

LUCIA M. SOMMERER

'Contemporary practices of risk operate in a way that precludes the possibility of a non-dangerous individual.'

Werth (2018, 1).

One reason for the existence of the presumption of innocence (PoI) is to prevent premature, wrongful convictions. Criminal convictions are one of the most serious ways in which a state can diminish its citizens' freedom. Imagining having to endure a conviction as an innocent person triggers a collective fear. Those wrongfully convicted despite procedural safeguards often describe their experience with such phrases as, 'I felt buried alive!'. If no standards existed to ensure that in principle no innocent person is subject to a criminal conviction, we might live in constant fear: a wrongful conviction could happen to any one of us at any time. The PoI sought to remedy this fear by imposing procedural standards on criminal trials. The question that shall be posed here, however, is: in a data driven government is it sufficient to remedy the threat of wrongful convictions by applying the traditional (narrow) PoI? Or is a broader interpretation of the PoI warranted because the contexts in which an innocent person may be 'buried alive' are no longer limited to criminal trials, but have expanded substantially into the time before a criminal trial. General arguments in favor of a broader PoI have been made in the past (Duff 2013; Ferguson 2016), but in this provocation I will focus on the possibly intrinsic need for a broader PoI in data driven governments.

Two readings of the PoI

Traditionally the gaze of the PoI was turned towards the past, i.e. to prior criminal actions. It has applied in cases where someone was accused of having committed a crime, and in the subsequent criminal proceedings. This narrow PoI is thus trial-related, which means it is related to limiting state actions taken only after an alleged criminal act.

A broader reading of the PoI should be adopted. This means a reading not limited to the criminal trial, but instead also related to risk assessments, i.e. to the suspicion that someone will commit a crime in the future. Such a broad reading is needed under data driven governments in which algorithmic profiling of individuals is increasingly used to determine the risk of future criminal activity, and where the criminal justice system attaches materially negative consequences to an individual's high-risk score. Based on pattern matching, it is determined whether someone belongs to a risk group that warrants the attention of the criminal justice system. Statistical profiling, a technique borrowed from behavioural advertising, has increasingly infiltrated the criminal justice sector, along with its existing shortcomings (opacity, discriminatory effects, privacy infringements, false positives) while also creating new ones, notably the risk of prejudgment.

The PoI should thus be a guiding principle not just for the repressive branch of the criminal justice system but must also be applied to preventive pre-trial decisions. Such an extension would provide the PoI with a protective, Janus-faced gaze into the past and into the future simultaneously. This gaze comes in the form of a special standard

of certainty required for both a criminal trial conviction and a high-risk determination triggering pre-trial detention.

A classification of high-risk generally does not claim to predict behaviour with near certitude. Rather, it claims that an individual shares characteristics with a group of people in which higher levels of criminal activity are present compared to the rest of the population. As an illustration, consider the pre-trial risk-assessment tool COMPAS, which is used in bail decisions in the U.S., and which equates an 8% likelihood of being arrested for violent crime in the future with high-risk status (Mayson 2018). A high-risk individual under COMPAS thus shares characteristics with a group of people of which 8% have been rearrested for a violent crime in the past. The nuances of this statistical judgment, including the lack of certitude are, however, ironed out in the application of the statistics. The likelihoods turn into "legal truth" for defendants when a judge at a bail hearing is presented with a high-risk classification (which generally neglects to mention the underlying statistics), and when defendants as a direct or partial consequence are then denied bail. The statistical information that out of a group of say 100 people with the same high-risk characteristics as a defendant only 8 have committed violent crimes turns into the "legal truth" that the defendant is dangerous and must be monitored further through denial of bail. The possibility that the individual belongs to the 92 individuals that have not committed a crime is not considered: the possibility of a non-dangerous individual in this group is, to use Werth's wording, precluded.

In this context one can speak of a dormant penal power embodied in the various data collected and in the profiles created about individuals. This penal power comes to life once a profile is fed into criminal justice algorithms that generate risk scores, which in turn are used as basis for a suspicion or as a justification for further criminal justice measures.

I argue that a mistake (false positive) in this risk context (A is found to be high-risk even though he is not), and in the context of a trial (B is found guilty even though he is innocent), are mistakes of the same kind. In both situations individuals receive a treatment they do not deserve. In the trial scenario, the PoI requires a special standard of certainty (beyond reasonable doubt) to convict someone in order to prevent wrongful convictions. A broad PoI would also require such special, uniform standard of certainty for ranking an individual high-risk and attaching manifestly negative criminal justice consequences (e.g. pre-trial detention) to this risk score. Similarly, a special standard of certainty would also be required for individual-level risk-assessments deployed by law enforcement. The standards in these cases may not be beyond reasonable doubt, but they would most certainly require more than 8% likelihood, and they would at least elicit an overdue public debate over what percentage of false positives society is willing to tolerate in its pursuit of security.

These described algorithmic developments are at the ridge of a deep conceptional shift – the so-called new penology – that appeared in criminal justice in the late 20th century. This shift was marked by the emergence of a then-new discourse on probabil-

ity and risk and a moving away from focusing on the individual offender towards actuarial considerations of aggregates (Feeley and Simon 1992). The recent incorporation of algorithms into this discourse is not merely a linear continuation of the new penology but an exponentiation which brings intrinsic novel challenges to criminal justice. As state actions against citizens are increasingly consolidated towards the preliminary stages of criminal investigations and preventive police-related settings, this provocation argues that legal protections must shift in the same direction to keep pace with technological developments.

Risk colonization: replacing 'action' with 'behaviour'

Another reason to scrutinize and if necessary criticize the integration of automated risk prediction technologies into both law enforcement and pre-trial operations is that it may eventually impact criminal law beyond merely these two contexts. It may just be the beginning of data driven transformations affecting the whole criminal justice system, and it may lead to a shift in the focus of substantive criminal law away from concretely defined criminal offences to the more diffuse category of (algorithmically determined) criminally relevant behaviour and attitudes. The practice of measuring a person against minute data of her past behaviour is not novel in criminal law theory. It is reminiscent of criminal law theories of 'Lifestyle-Guilt' (Mezger 1938, 688) and 'Life-Decision-Guilt' (Bockelmann 1940, 145). These theories disassociated punishability and guilt from a single deliberate or negligent act and attached it to the inner nature of actors reflected in their past life choices. These approaches were popular in Germany during the Third Reich, and are particularly suited to autocratic rule due to their fluidity.

Already today predictive crime technologies and the data collected by such technologies do not remain solely within the realm of pre-trial decisions and policing, but have colonized court settings as well. In the U.S., even sentencing decisions are supported by data driven algorithmic analyses of an offender's future behaviour. It is noteworthy that these analyses were initially developed only for the use in preventive law enforcement measures and only later migrated to a sentencing context (Angwin et al. 2016). A last step of this development may be the migration of algorithmic determinations to the establishment of criminal liability. It is worth noting that the possibility of supporting human rights judgements (Aletras et al. 2016), Supreme Court decisions (Islam et al. 2016), and civil litigation (Katz 2014) with data driven algorithms is already being sounded out. The potential future use of data driven predictions to determine criminal liability should thus not be discounted as a scenario too outlandish to prepare the legal system for.

Outlook: benign amusement and/or bitter reproach

It may well be that future generations, or at least their advantaged elites, will look back on our critical analyses of data driven government and smile benignly at us. Possibly with the same amusement we feel today, when we read about Plato's warnings against the technology of writing (Plato, Phaedrus, 275A). A less-advantaged segment of future societies, however, comprised of individuals who were unable to profit from digital advancements, locked in negative algorithmic presumptions about themselves,

may look back on us more reproachfully, asking why no safeguards for the rule of law and the PoI were pressed for and implemented.

This provocation does not aim to provide a definitive answer as to which of these scenarios is more likely to occur. Rather, it aims to prod the reader to think about whether and how traditional legal protections may be expanded to accommodate the rapid technological evolutions now changing the face of the criminal justice system. Particularly, it wants to prod the reader to consider if the PoI could be understood already today as a two-faced mechanism. A mechanism not just to ensure that a defendant in court will be presumed innocent until proven guilty. But a mechanism which also outside of a trial counters anticipatory data driven risk practices that preclude the possibility of a non-dangerous individual.

References

Aletras, Nikolaos, Dimitrios Tsarapatsanis, Daniel Preoþiuc-Pietro, and Vasileios Lampos. 2016. "Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective." PeerJ Computer Science 2 (October): e93. doi: 10.7717/peerj-cs.93.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "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.

Big Brother Watch. 2018. "Press Release: Police use Experian Marketing Data for AI Custody Decisions." https://bigbrotherwatch.org.uk/all-media/police-use-experian-marketing-data-for-ai-custody-decisions/.

Bockelmann, Paul. 1940. Studien zum Täterstrafrech. Vol. 2. Berlin: de Gruyter.

Duff, Anthony. 2013. "Who Must Presume Whom to be Innocent of What?" The Netherlands Journal of Legal Philosophy 42:170-92.

Feeley, Malcom M., and Jonathan Simon. 1992. "The New Penology: Notes on the Emerging Strategy of Corrections and Its Implications" Criminology 30(4): 449-74.

Ferguson, Pamela R. 2016. "The Presumption of Innocence and its Role in the Criminal Process." Criminal Law Forum 27:131-58.

Islam, Mohammad Raihanul, K.S.M. Tozammel Hossain, Siddharth Krishnan, and Naren Ramakrishnan. 2016. "Inferring Multi-dimensional Ideal Points for US Supreme Court Justices." AAAI'16 Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 4-12.

Katz, Pamela S. 2014. "Expert Robot: Using Artificial Intelligence to Assist Judges in Admitting Scientific Expert Testimony." Albany Law Journal of Science and Technology 24(1):1-47.

Mayson, Sandra G. 2018. "Dangerous Defendants." Yale Law Journal 127(3): 490-568.

Mezger, Edmund. 1938. "Die Straftat als Ganzes." Zeitschrift für die gesamte Strafrechtswissenschaft 57 (1): 675-701.

Oswald, Marion, Jamie Grace, Sheena Urwin, and Geoffrey Barnes. 2018. "Algorithmic risk assessment policing models: Lessons from the Durham HART model and 'Experimental' proportionality." Information & Communications Technology Law 27(2): 223-50. doi: 10.1080/13600834.2018.1458455.

Werth, Robert. 2018. "Theorizing the Performative Effects of Penal Risk Technologies: (Re)producing the Subject Who Must Be Dangerous." Social & Legal Studies Online First: 1-22. doi: 10.1177/0964663918773542.

PREDICTIVE POLICING. IN DEFENCE OF 'TRUE POSITIVES'

SABINE GLESS

Imagine that law enforcement agencies, using a profiling program, stop and search only people of a particular ethnicity within a specific area. The results are astounding: seven drug dealers and two robbers on the run were caught. Once in court, all defendants claim the arrest was discriminatory because of an obviously biased profiling program. Their attorneys argue that the evidence obtained as a consequence to the tainted arrests, must be excluded and that only this removal of the fruit of the poisonous tree will motivate legal authorities to closely monitor policing programs and prevent bias as best as they can. On the other side of the room, victims and members of the public want justice. They consider the use of algorithms to bring perpetrators before the law to have been a success.

Administration of criminal justice and profiling

Predictive policing has triggered a heated debate around the issue of false positives. Biased machine training can wrongly classify individuals as high risk simply as a result of belonging to a particular ethnic group and many agree such persons should not have to shoulder the burden of over-policing due to an inherent stochastic problem (cf. Veale,Van Kleek, and Binns 2018). True positives, or individuals who are correctly identified as perpetrators, do not make headlines. If drugs or other incriminating evidence is found in their possession after being stopped and searched, the fact that such evidence was found using biased profiling is justified because the suspicion turned out to be well-founded. Had the police officer identified them, their colleagues would probably laud them for "good intuition." Scholars have demonstrated that sorting by stereotypes is a form of generalization all humans use routinely (Schauer 2003). However, as Hildebrandt (2008, 30) explained in *Profiling the European Citizen*, with automated profiling the need to effectively constrain such practices in order to prevent a technological infrastructure that practically destroys fairness in criminal justice is eminent.

This provocation argues that the 'true positives' offer the best opportunity to address the issue of biased profiling. The first reason is purely pragmatic – they are already party to a criminal investigation and, as such, have a strong incentive to challenge law enforcement methods and scrutinize policing methods on an individual basis. The second reason is more general (and commonly subscribed to) – that discriminatory stops and searches are inherently unfair, threaten social peace, and frustrate targeted groups (DeAngelis 2014, 43). Use of biased algorithms in policing not only places a burden upon those deemed 'false positives', but also contaminates the 'true positives'. To create an efficient legal tool against discriminatory law-enforcement, defence should be entitled to contest a conviction for biased predictive policing, with a specific exclusionary rule protecting 'true positives' against the use of tainted evidence.

The legal standing of "true positives"

The legal standing of individuals prosecuted following an arrest triggered by biased profiling is unclear. Even an outright illegal arrest may not affect prosecution, although the European Court of Human Rights (ECtHR) has extended certain defense rights to the investigation phase (*Allenet de Ribemont v. France* 1995, 11-13, para. 32-37) and in certain situations, the defense may invoke exclusionary rules with reference to

tainted evidence. The problem is that the exclusion of evidence is a controversial issue (Estreicher and Weick 2010, 950-51) and it remains unclear whether biased predictive policing would actually trigger such exclusion. Generally, where incriminating evidence is found, it is the responsibility of the authorities to clarify the facts and enforce the law. After all, there is public interest not only in bringing criminals to justice, but also in supporting victims.

By contrast, defendants have standing to claim that an arrest was discriminatory and unfair (Gillan and Quinton v. the United Kingdom 2010, 42-45, para. 76-87) and it is in the public interest to stop biased police work and discriminatory arrests. How to resolve the conflict between the interests of the public to obtain justice while simultaneously honouring a defendant's rights depends on the composition of each criminal justice system. However, all systems are faced with the issue of biased policing to some degree and all, to a certain extent, operate on the (yet controversial) premise that a threat of excluding evidence will deter authorities from particular practices (Kafka 2001, 1922-25). Therefore, adopting an exclusionary rule appears to be the obvious solution.

Creating a specific legal remedy for the 'true positives' is the most promising way to deter biased predictive policing. Such individuals are already in the courtroom and can raise appropriate objections while the 'false positives' would have to initiate a new legal action and have little incentive to do so. Similarly, courts or administrative bodies empowered to monitor biased profiling may also lack the incentive to draw attention to biased law enforcement practices in the absence of a powerful legal remedy for 'true positives'.

Exclusion of evidence: A price too high to pay?

Clearly excluding evidence obtained using biased predictive policing techniques will not be a popular remedy in most criminal justice systems. Objections around presumptions of guilt and subverting the interests of justice and the victims would likely be cited. However, if one scrutinizes these arguments, they may turn out to be less convincing than initially thought.

With reference to the first argument, Art. 6 (2) ECHR guarantees European citizens charged with a criminal offence are "presumed innocent until proven guilty according to law." Courts and legal scholars agree that the meaning of the presumption of innocence is broad. What they don't agree on is whether or not the guarantee extends to investigations and other pre-trial actions and it is not explicitly stated in the Convention. However, according to the case-law of the ECtHR, members of the court may not begin criminal proceedings with the preconceived notion that an individual has committed the offence in question (*Barberà, Messegué and Jabardo v. Spain* 1988, 27, para. 77; *Allenet de Ribemont v. France*, 1995, 11-13, para. 33-36). Referring to this line of cases, scholars correctly argue that if the presumption of innocence is not extended to police profiling it will lose its place as a guiding principle in the era of ubiquitous surveillance and big data.

Regarding the second argument, implicit in the objection to an exclusionary rule barring fruit of the poisonous tree is concern that the wheels of justice will lose momentum if a perpetrator is allowed to walk free despite incriminating evidence. This dichotomy is present for every exclusionary rule and invokes our traditional goals of punishment and deterrence. However, there is also the understanding among citizens that authorities will prosecute crimes properly. This involves integrity in both the investigation and subsequent legal proceedings so that individuals against whom the state has a valid case do not walk free. The public's interest in honesty and transparency in investigations provides protection from arbitrary justice and supports the notion that law enforcement agencies should monitor their profiling programs for implicit bias. The EU lawmaker acknowledges this interest with provisions on accountability in prosecution where automated profiling carries the risk of prohibited discrimination (cf. Art. 11 para 3 and Art. 10 Directive (EU) 2016/680).[1]

Support for the exclusion of tainted evidence may also be found in the protection against unreasonable detention. According to Art. 5 (c) ECHR, no citizen's liberty may be deprived except in limited situations, including where there is "reasonable suspicion" that the individual committed an offense. The ECtHR has noted that this requirement that the suspicion be reasonable forms an essential part of the safeguard against arbitrary arrest and detention. More specifically, 'having a "reasonable suspicion" presupposes the existence of facts or information which would satisfy an objective observer that the person concerned may have committed the offence' (*Fox, Campbell and Hartley v. the United Kingdom* 1990, 12, para. 32; Ferguson 2015, 286).

In the case of a human police officer, he or she must identify enough elements, or 'probable cause', to satisfy an objective observer regarding the possible guilt of an individual. In contrast, a police profiling system based upon algorithms does not just monitor one potential subject, but categorizes individuals in a way that assumes certain groups are more likely than others to commit crimes, thus deserving of additional police attention. If such profiling leads to a search and subsequent arrest, no individual law enforcement agent has ex ante identified any probable cause for the arrest. In fact, he or she may never know how the data was formed that resulted in the arrest. This hardly constitutes the reasonable suspicion required by the ECHR.

Justification of an exclusionary rule is also supported by the principle of equality before the law, which is central to any democracy. If police action is based on algorithms that divide a population into groups based upon particular attributes, the result will be a fundamental change to our legal system characterized by an increase in unlawful searches and detainment, in addition to violations of the privacy and liberty of all citizens. It will result in the Orwellian world in which 'some animals are more equal than others' and Big Brother is watching you. That said, one would be mistaken to assume that law enforcement agents were 'colour-blind' prior to the advent of automated profiling, but to date, biased searches by human officers have not paved the way to specific exclusionary rules.

Willingness to pay the price

With predictive policing programs on the rise we must be willing to pay the price of a strong exclusionary rule. A rule barring incriminating evidence found in the possession of a 'true positive' after a discriminatory arrest can be grounded in two lines of reasoning. The first is legal and builds upon the rationale that an overpoliced individual can invoke an exclusionary rule on the basis of an unreasonable search. The second line of argument is as simple and straightforward as it is pragmatic: there is public interest in creating an efficient legal tool against biased profiling and against unmonitored use of such programs (Hildebrandt 2015, 184, 195). Therefore, it is the 'true positives' that offer us the best chance to require authorities to monitor their profiling tools due to the inherent incentive in pointing out potential bias and prohibited discrimination during an ongoing proceeding.

Notes

[1] Directive (EU) 2016/680 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA).

References

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias." 2016. ProPublica, May 23. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

DeAngelis, Peter. 2014. "Racial Profiling and the Presumption of Innocence." Netherlands Journal of Legal Philosophy (1): 43-58.

Estreicher, Samuel and Daniel P Weick. 2010. "Opting for a Legislative Alternative to the Fourth Amendment Exclusionary Rule." University of Missouri-Kansas City Law Review 98: 949-66.

Ferguson, Andrew Guthrie. 2012. "Predictive policing and reasonable suspicion." Emory Law Journal 62(2): 259-326.

Galetta, Antonella. 2013. "The changing nature of the presumption of innocence in today's surveillance societies: rewrite human rights or regulate the use of surveillance technologies?" European Journal of Law and Technology, 4(2). http://ejlt.org/article/view/221/377.

Hildebrandt, Mireille. 2008. "Defining Profiling: A New Type of Knowledge?" In Profiling the European Citizen: Cross-disciplinary Perspectives edited by Mireille Hildebrandt and Serge Gutwirth, 17-30. Dordrecht: Springer.

Hildebrandt, Mireille. 2015. Smart Technologies and the End(s) of Law. Novel Entanglements of Law and Technology. Cheltenham: Edward Elgar.

Kafka, Michael T. 2001. "The Exclusionary Rule: An Alternative Perspective." William Mitchell Law Review 27: 1895-939.

Schauer, Frederick. 2003. Profiles, Probabilities and Stereotypes. Cambridge, MA: Belknap Press of Harvard University Press.

Starr, Sonja B. 2009. "Sentence Reduction as a Remedy for Prosecutorial Misconduct." Georgetown Law Journal 97: 1509-66.

Veale, Michael, Max Van Kleek, and Reuben Binns. 2018. "Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making." Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Paper no. 440. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3175424.

THE GEOMETRIC RATIONALITY OF ALGORITHMIC DECISIONS

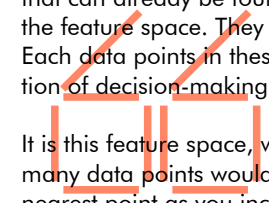
TOBIAS BLANKE

In this provocation, I would like to develop what I call the geometric rationality of algorithmic decisions, which measures social relations using the distance of data points in abstract geometric spaces. This analysis follows on to the work by Claudia Aradau and myself (Aradau and Blanke 2017), where we introduce the concept of 'in-betweenness' in abstract information spaces as a foundation of algorithmic prediction. In this paper, I elaborate how algorithmic innocence, i.e. innocence before an algorithm, is (pre-)decided by a geometric rationality of algorithms. I show how (non)-innocent subjects are created proactively and to be acted upon pre-emptively by algorithmic manipulation of an abstract feature space.

Abstract geometric information spaces

Computational decision-making techniques generally operate with the spatial metaphor of abstract geometric spaces. AI has set off with the idea of abstract information spaces, as an MIT website from the 1990s reveals: 'An information space is a type of information design in which representations of information objects are situated in a principled space. In a principled space location and direction have meaning, so that mapping and navigation become possible' (MIT Artificial Intelligence Laboratory 1998). In the world of AI, we are interested in meaningful information spaces that do not count all available information but only information, which can 'feature' in the calculation of a problem. These are the features describing to algorithms us and all other things in the world. Together these features span an abstract information space using 'vectors' of features. For instance, for people we might think about gender, height, weight and age as features, each of which is a dimension of the problem to be modelled. In this case, we have a four-dimensional feature space. Machine learning techniques that have propagated across different fields can tell us how 'people are materialised as a bundle of features' (Mackenzie 2013).

Decision-making algorithms plot data as points/dots in feature spaces, which thus become a geometrical representation of all the data available to them. Each dot in this space is defined by how much abstract space is in-between it and the other dots in the same space or how distant they are from each other. For practitioners who operate decision-making algorithms, '[d]ata is in some feature space where a notion of "distance" makes sense' (Schutt and O'Neil 2013, 81). In principle, there is no limit to the number of features that can be used to build such an artificial space. Feature spaces can have hundreds, thousands or hundreds of thousands of features/dimensions, depending on how much a computer can process. Machine learning algorithms manipulate this feature space in order to create labels for each example that can already be found in the feature space or that might be found in the future in the feature space. They 'partition' the feature space into zones of comparable features. Each data points in these zones is labelled the same way. Labelling is the materialisation of decision-making by machine learning.



It is this feature space, which drives the (big) data needs in machine learning: 'How many data points would you need to maintain the same minimum distance to the nearest point as you increase the number of inputs of the data? As the number

of inputs increases, the number of data points needed to fill the space comparably increases exponentially' (Abbott 2014, 153), because the number of inputs corresponds to the number of features. The more feature dimensions an abstract information space has the more space there is to fill in this space. The big data drive is a result of the attempt to fill the feature space. In a famous paper Banko and Brill (2001) set the agenda for the big data hype and its rationale. They demonstrated that digital reasoning of all kinds gets more accurate by throwing more data at it, as the feature space gets filled with data points.

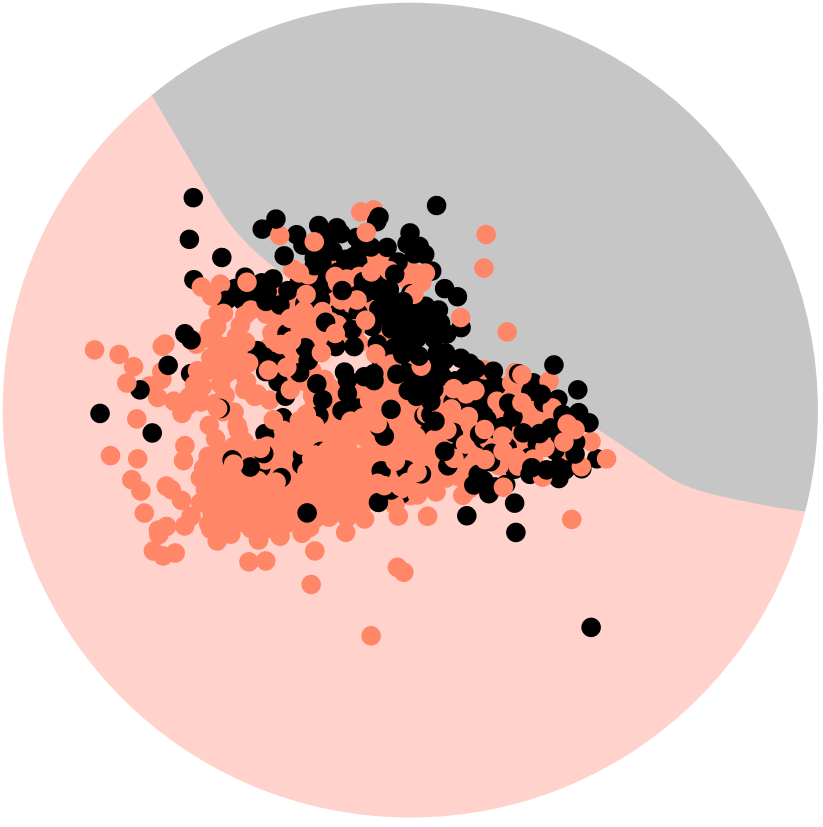


Figure 1. Decision Boundaries

Geometric decision-making

As analysed by Anrig et al. (2008), there are many examples of decision-making algorithms and how they work the feature space. They all 'partition' the space into zones to generate meaning for all data points in the space. Partitioning is geometric decision-making by algorithms. 'Decision trees', e.g., partition the space into decisions made over a subset of features. This way, they collect all points in the space that are, for instance, of gender female, taller than 1.70m, weigh more than 65 kg, etc.

'Clustering' using nearest neighbours partition the feature space into zones of points that share similar feature dimensions and are defined by their border to other zones in the high-dimensional space. 'Regression analysis' as well as '(deep) neural networks' can learn more complex boundaries between zones of similar features. 'Often, different methods are used and the quality of their results compared in order to select the "best one".' (Anrig et al. 2008, 79).

We generated the worked example in Figure 1 to demonstrate a feature space with data points in two classes (black and orange dots). The space is partitioned into two zones by a complex non-linear boundary generated by a neural network. While complex decision boundaries can generate highly accurate zones and partitions, they are known to be difficult to understand. Neural networks are unintelligible compared to the example of decision trees. Cathy O'Neil likens it to the godlike unintelligibility: 'Like gods, these mathematical models were opaque, their working invisible to all but the highest priests in their domain: mathematicians and computer scientists.' (O'Neil 2016, 7). As these models and algorithms are integrated within complex artificial systems, they risk becoming 'black boxes', unintelligible even to the 'high priests' of the digital world. 'In the era of machine intelligence', O'Neil cautions, 'most of the variables will remain a mystery. (…). No one will understand their logic or be able to explain it' (O'Neil 2016, 157).

Such 'unintelligibility' prevents human observers from understanding how well algorithmic geometric rationality works. We cannot be reassured by following the rules of evaluation of the 'high priests' either, as they are made to make the algorithms perform computationally and not socially. This is expressed in what counts for right and wrong decisions in the feature space. Errors and error rates are key sites of the transformation of knowledge, but also sites of controversy with regards to innocence and non-innocence, as decisions by algorithms are contested. In 2018, e.g., the media reported that the company ASI Data Science had developed an extremism blocking tool with government funding of £600,000, which could automatically detect 94% of Isis propaganda with 99.99% accuracy (Lomas 2018). As reported, this is at best confusing information, as nothing else is known about the experiments that led to these error rates. 94% will still be concerning for the security analyst dealing with a system like Facebook and billions of new items a day. 6% missed content can then mean 1,000s of items. We do not know how the accuracy is measured either but 'the government says' for 'one million "randomly selected videos" only 50 of them would require "additional human review".' (Lomas 2018). This means in our Facebook example a block of '50,000 pieces of content daily'. Finally, the tool is also single-minded and does not partition the feature space for all terrorist content but only for ISIS data of a particular time. Complex digital reasoning tends to be single-minded in this way because each feature space is a unique geometry. High accuracy figures for decision-making algorithms should never be enough to reassure us that these algorithms are correct and make wise pre-emptive decisions.

The reader might have also noticed a black dot in the Figure 1 far away in the bottom-right non-innocence blue corner. This is called an outlier and is as such

suspicious/interesting, because we are not just innocent in the feature space by association with other innocent dots close by but also by dissociation with other dots. 'The outliers are determined from the "closeness" (…) using some suitable distance metric.' (Hodge and Austin 2004, 91). We investigate the relations of digital selves and others implied by outlier detection in (Aradau and Blanke, 2018), where we present the real-life security impact outlier methods have. Outliers predetermine innocence in feature spaces as much as closeness does.

The data scientist McCue specialises in outlier-detections in predictive policing. She gives an example from security analytics that demonstrates the power of outlier detection using a cluster analysis (McCue 2014, 102). They monitored conference calls to find clusters of numbers based on geographies and regions. '[I]t would not have been possible to analyze these data without the application of data mining.' (McCue 2014, 104). The resulting two-dimensional feature space exhibits three clusters including one outlying cluster in the bottom-left corner. Features included 'the conference IDs (a unique number assigned by the conference call company), the participants' telephone numbers, the duration of the calls, and the dates' (McCue 2014, 104). '[T]hree groups or clusters of similar calls were identified based on the day of the month that the conference occurred and the number of participants involved in a particular call.' (McCue 2014, 106). The outlier cluster correctly identified a professional criminal network. For McCue, this approach has various advantages. Firstly, one can literally 'see' in the feature space why one cluster is different from the others and an outlier. Secondly, the information used to cluster the participants is not necessarily based on detailed information of individuals in the cluster as it summarises their existence into features, and surveillance can take place without much attention to privacy limitations. Finally, the clusters that are not outliers build a dynamic, algorithmic model of normality. Non-suspicion or innocence is determined by declaring some cluster to be not outliers, while anomalies are outside any cluster. The geometrical distance in the feature space makes outlier dots stand out as outliers.

Conclusion

This short provocation presented ideas on how innocence through algorithms is pre-determined by the position of dots in an abstract feature space and an underlying geometric rationality of distances between dots. We examined the foundations of this geometric rationality, its need for more and more data as well as issues preventing reasoning about errors critically. To be finally counted as innocent, a dot should be close enough to the innocent dots in the abstract space and also not too close or too far away in order not to be suspicious again.

References

Abbott, Dean. 2014. Applied predictive analytics: Principles and techniques for the professional data analyst. New Jersey: John Wiley & Sons.

Anrig, Bernhard, Will Browne, and Mark Gasson. 2008. "The Role of Algorithms in Profiling." In Profiling the European Citizen: Cross-disciplinary Perspectives, edited by Mireille Hildebrandt and Serge Gutwirth, 65-87. Dordrecht: Springer.

Aradau, Claudia, and Tobias Blanke. 2017. "Politics of prediction: Security and the time/space of governmen-

tality in the age of big data." European Journal of Social Theory 20(3): 373-91.

Aradau, Claudia, and Tobias Blanke. 2018. "Governing others: Anomaly and the algorithmic subject of security." European Journal of International Security 3(1): 1-21.

Banko, Michele, and Eric Brill. 2001. "Scaling to very very large corpora for natural language disambiguation." Proceedings of the 39th annual meeting on association for computational linguistics.

Hodge, Victoria J., and Jim Austin. 2004. "A survey of outlier detection methodologies." Artificial Intelligence Review 22(2): 85-126.

Lomas, Natasha. 2018. "UK outs extremism blocking tool and could force tech firms to use it." https://techcrunch.com/2018/02/13/uk-outs-extremism-blocking-tool-and-could-force-tech-firms-to-use-it/, accessed at 31/8/2018.

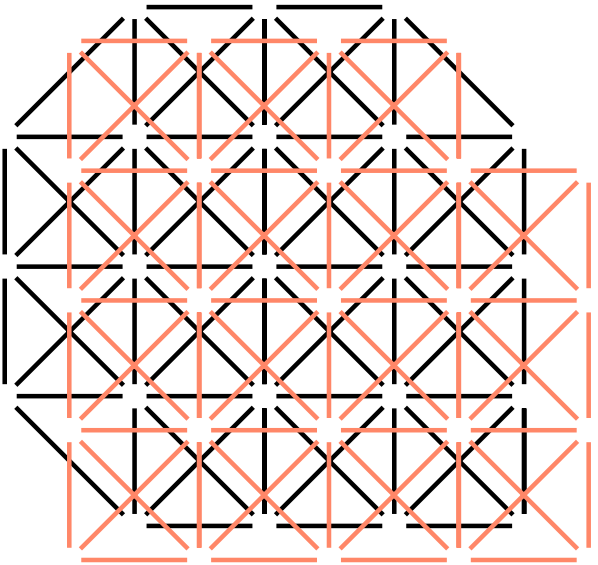
Mackenzie, Adrian. 2013. "Programming subjects in the regime of anticipation: Software studies and subjectivity." Subjectivity 6(4): 391-405.

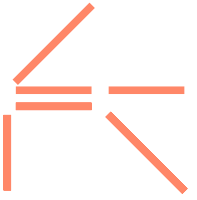
McCue, Colleen. 2014. Data mining and predictive analysis: Intelligence gathering and crime analysis. Oxford: Butterworth-Heinemann.

MIT Artificial Intelligence Laboratory. 1998. The JAIR Information Space. http://www.ai.mit.edu/projects/infoarch/jair/jair-space.html, accessed at 31/8/2018.

O'Neil, Cathy. 2016. Weapons of math destruction: How big data increases inequality and threatens democracy. New York: Crown.

Schutt, Rachel, and Cathy O'Neil (2013). Doing data science: Straight talk from the frontline. New York: O'Reilly Media, Inc.





ON THE PRESUMPTION OF INNOCENCE IN GOVERNMENT: ARE WE ASKING THE RIGHT QUESTIONS?

LINNET TAYLOR[1]

This provocation will argue that the presumption of innocence is a principle of fundamental importance to the rule of law, but that it is of limited use if we wish to protect people from unfairness in data-driven government. As identified by Jacquet-Chiffelle (2008) in Profiling the European Citizen, indirect profiling (assuming, based on data analytics, that an individual fits a particular group profile) was a practice that presented challenges to privacy, autonomy and fairness in 2008. This practice has increased exponentially with the addition of a myriad new data sources, becoming more entrepreneurial (Pasquale 2015) with ubiquitous sensing and distributed data governance. In response, we need new, broader framings of our rights in relation to data profiling. We can no longer assume that the most important data harms relate to individual data subjects as potential rights claimants, and stem from targeted governmental data processing. If we do this, we are effectively searching for our keys under the lamppost, where the light falls.

(Re)defining the question

The presumption of innocence is a key principle that allows us to contest governmental practices of data processing and profiling. Is it all we need to consider, however, given ubiquitous and continual data collection on our behaviour and movements? Or do we need, as well as principles to ensure just treatment of citizens by states, principles to ensure the just treatment of anyone, by anyone with the power to collect and process data? I will use two (semi)hypothetical cases to support my contention, both of which raise questions as to whether the liberal individual framing of rights in relation to data processing is sufficient to help with the challenges we are now facing (Cohen 2019).

In order to discover whether the question the presumption of innocence answers is actually the question that faces us, we should begin from the contemporary landscape of data collection. The proliferation of sensors and the growth in sensing technologies mean that today the majority of digital signals used in profiling come not from individuals engaging consciously with authorities or firms, but from our contact with environments and devices that sense our actions and behaviour.

Under these conditions the presumption of innocence may not be the most useful route to justice: profiling practices that use data collected from environmental sensing, or behavioural and location traces, has a more complex relationship to suspicion or innocence than classic forms of 'volunteered' data such as administrative data gathered by public authorities. Traditionally governmental institutions involved in profiling have identified particular individuals as potential targets based on their membership of a category of interest, resulting in a process of 'blacklisting', 'greenlisting' or 'greylisting' (Broeders and Hampshire 2013). In this case, it is relevant to cite the presumption of innocence as a counter to the clear harm of blacklisting. However, we see processes emerging today on an entrepreneurial basis which instead of starting from established suspect categories (such as people who have downloaded particular documents or belong to particular online groups), mine general data in order to discover anomalies that may relate to suspicious behaviour. The act of focusing on individuals within those categories is only remotely connected to this discovery process, and may not be part of the process at all. The following examples will help to explain this claim.

Two illustrative cases

Case 1: A private-sector consultant to the EU's Space Agency ESA aims to track the paths of undocumented migrants into Europe. The project leaders use various data sources including satellite images of human mobility through North Africa, social media output, local online reports and migrants' mobile calling records as they travel through the desert. The datasets are combined and fed into a machine learning process designed to guess at migrants' place of origin, and thus their likely type of claim to asylum once they make contact with migration authorities. The consultants then sell their consulting services to various of those authorities, and as contractors, use it to determine where to look for migrants who are outliers in terms of successful asylum claims – those from more peaceful or democratic countries – who are predicted to be 'safe' to turn around and send directly home. At no point are migrants targeted as individuals by authorities as a result of the analysis – it is one of several streaming information sources available to various authorities, who make decisions about groups, based on profiles. The system can also identify migrant groups who, if prevented from claiming asylum, would lead to a high-profile human rights problem, and they can be escorted to a place where they can make their claims. It is important to note that this logic is problematic not only because of the right to claim asylum, but also because place of origin is not a valid indicator for the basis of asylum claims.

Case 2: Aminata, a young woman from West Africa, is on a one-day visit to a European city. She stops for a drink on her way through a living lab: an area of semi-public space where the right to surveil the street is awarded by the municipality to any corporation wishing to test its products or services on the public (one example is Stratumseind in the Dutch city of Eindhoven). As Aminata walks down the street, smart lampposts collect data on the way she is moving, her facial expression, skin colour and clothing, the signals her phone is picking up from the nearby antennae and the signals it is sending out to identify itself to those antennae and to local wi-fi points. Aminata is on her way to the airport, without any plans to return to Europe.

Unbeknownst to her, however, a fight has broken out in one of the bars as she passed by. All the data on people in the area at the time is later mined by the company running the living lab, and Aminata shows up as an anomaly in the dataset: her face, skin colour and her phone plan's origin are outliers. The following year this analysis is added to a hundred others by the next company using the living lab. It sells its data on to a national firm, which uses it to train a model designed to algorithmically identify risks to public order in the urban environment. This model is marketed as a service to any organisation interested in this task, including consultants to the police.

Aminata's data are in the model, but by now they only come into play in combination with the data of others, and under certain conditions – particular questions relating to types of incident or urban environment – in a process which would not be transparent to the firm running the model. Where 'N=all incidents and people present' no outlier is excluded, and due to a coincidental mix of conditions, people with some of Aminata's characteristics become flagged by the model as related to violence. This

influences municipal and law enforcement policy towards African migrants negatively in various ways, but is never made explicit in policy or guidelines.

Both these cases demonstrate ways in which data mining is used to create profiles, but where models built on large and diverse datasets to create 'evidence' in ways that are opaque to the user. When data is aggregated and sold by one user to another, it becomes impossible to check its original meaning. Yet completeness often becomes a synonym for reliability: a dataset reflecting all the violent incidents in a street, or all the migrants passing through North Africa, is likely to be seen as more reliable than one showing just a few, or one that disaggregates individual event data to understand more about causation. Aggregation facilitates decision-making at the same time as concealing meaning. Furthermore, though, the individual is neither identifiable nor individually analytically important in the dataset. It is their characteristics in relation to the larger group that provide the means for prediction (for more on this, see Taylor, Floridi, & van der Sloot 2017). It is also in relation to these generalisations (based on 'types, not tokens', Floridi 2014) that decision-making is done. Increasingly, when we are affected by data-driven governance it is not because of our own data but because of others', which has travelled amongst users and through models entirely within the private sector. The question of presumption of innocence becomes less relevant in the more vague, diffuse practice of risk prediction, based on data whose origin can no longer be traced, and which has never been attached to a single identity.

What exactly, then, is being challenged by such profiling? It may be best phrased as the right to resist inclusion in the database – any database. This is not a right data protection can address: instead it relates to privacy, and is fundamentally a political question. I should be able to choose how components of my identity are used by others, and to resist their arbitrary inclusion in processes that involve exerting power over anyone's options and behaviour. Although framed as individual rights, in this case privacy and autonomy must extend beyond the individual and also become conceptualised in relation to all of us at once – the people in this street, the people in that region of the desert. In cases such as these the need to exert our rights materialises in relation to the final destination and purpose of data about us, but cannot be predicted at the moment the data is collected. The mobile network operator, the living lab's temporary director, or the social media company cannot predict how the data they provide will be used. Conversely, the state (if it is involved) becomes less likely to know the reliability or origin of the data it is using in relation to a particular problem, or the way in which its processing affects its ability to answer the question.

Combating invisible harms

It is salutary to remember that Edward Snowden made his revelations about government surveillance while working for the private consulting firm Booz Allen Hamilton, which was providing commercial consulting services to the US government. Increasingly, problems for which governments are answerable will have many invisible authors against whom we have few enforceable rights.

I have argued that where profiling is an entrepreneurial service and risk assessment and pre-emption the aim, the presumption of innocence becomes relevant at the end of a long line of actions. The single, identifiable subject, the single identifiable watcher and the auditable data supply chain that ends with a governmental actor are increasingly a fiction. Instead we are seeing the emergence of an entrepreneurial free-for-all which conceals data's origins, paths, purposes and reliability. In relation to this, we will need human rights that can be claimed against any data collector, that are pre-emptive and that are powerfully enforced by government. Given the billions generated by this growing market for data, however, a remedy seems as far away as it did when Hidebrandt and Gutwirth published their volume in 2008.

Notes

[1] This research has received funding from the European Research Council under the European Union's Horizon 2020 Programme / ERC Grant Agreement n. [757247].

References

Broeders, Dennis, and James Hampshire. 2013. "Dreaming of Seamless Borders: ICTs and the Pre-Emptive Governance of Mobility in Europe." Journal of Ethnic and Migration Studies 39(8): 1201–18. doi: /10.1080/1369183X.2013.787512.

Cohen, Julie E. 2019. "Turning Privacy Inside Out." Theoretical Inquiries in Law 20(1): 8–36.

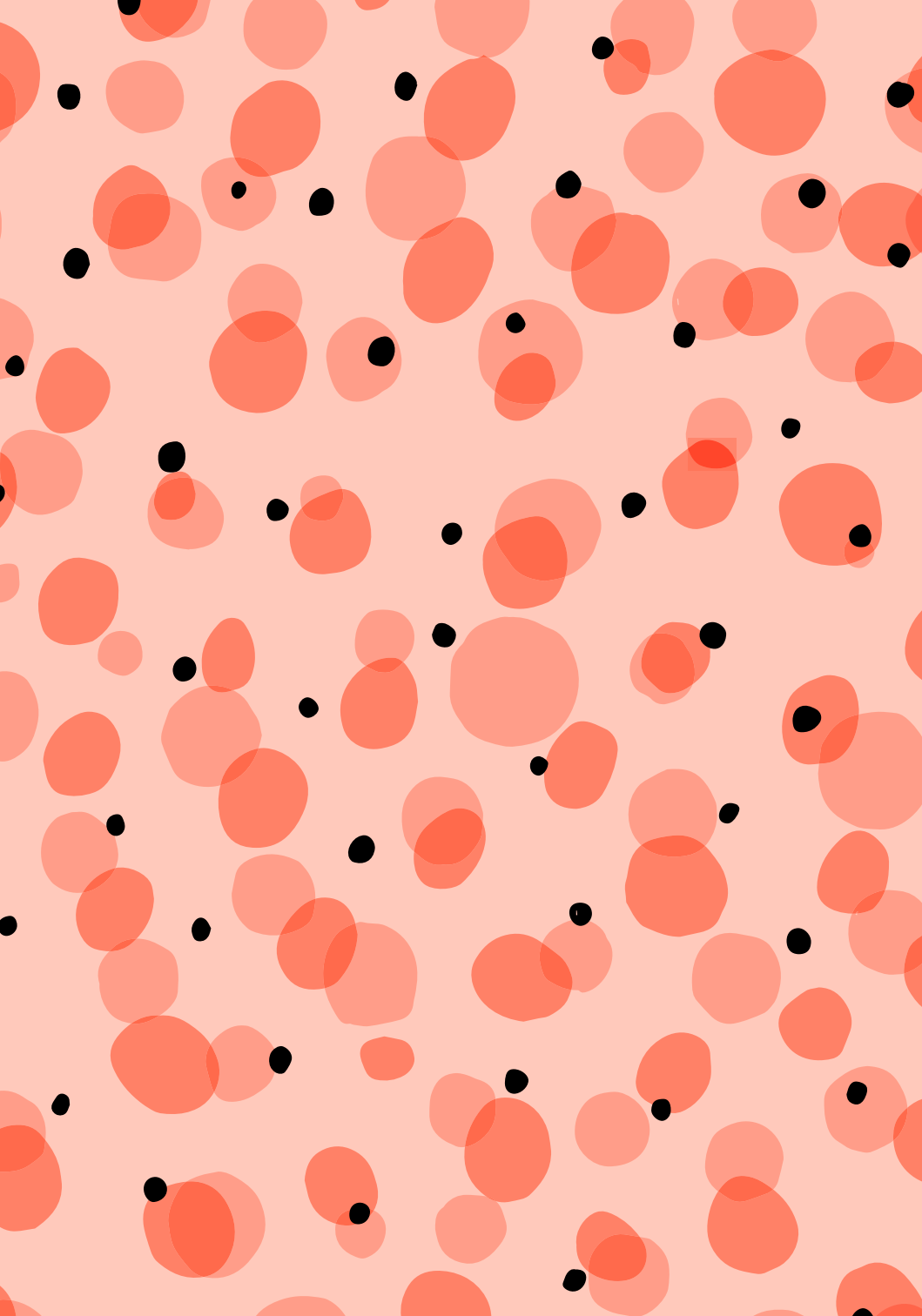
Economist. 2018. "Does China's Digital Police State Have Echoes in the West?" The Economist, May 31, 2018. https://www.economist.com/leaders/2018/05/31/does-chinas-digital-police-state-have-echoes-in-the-west.

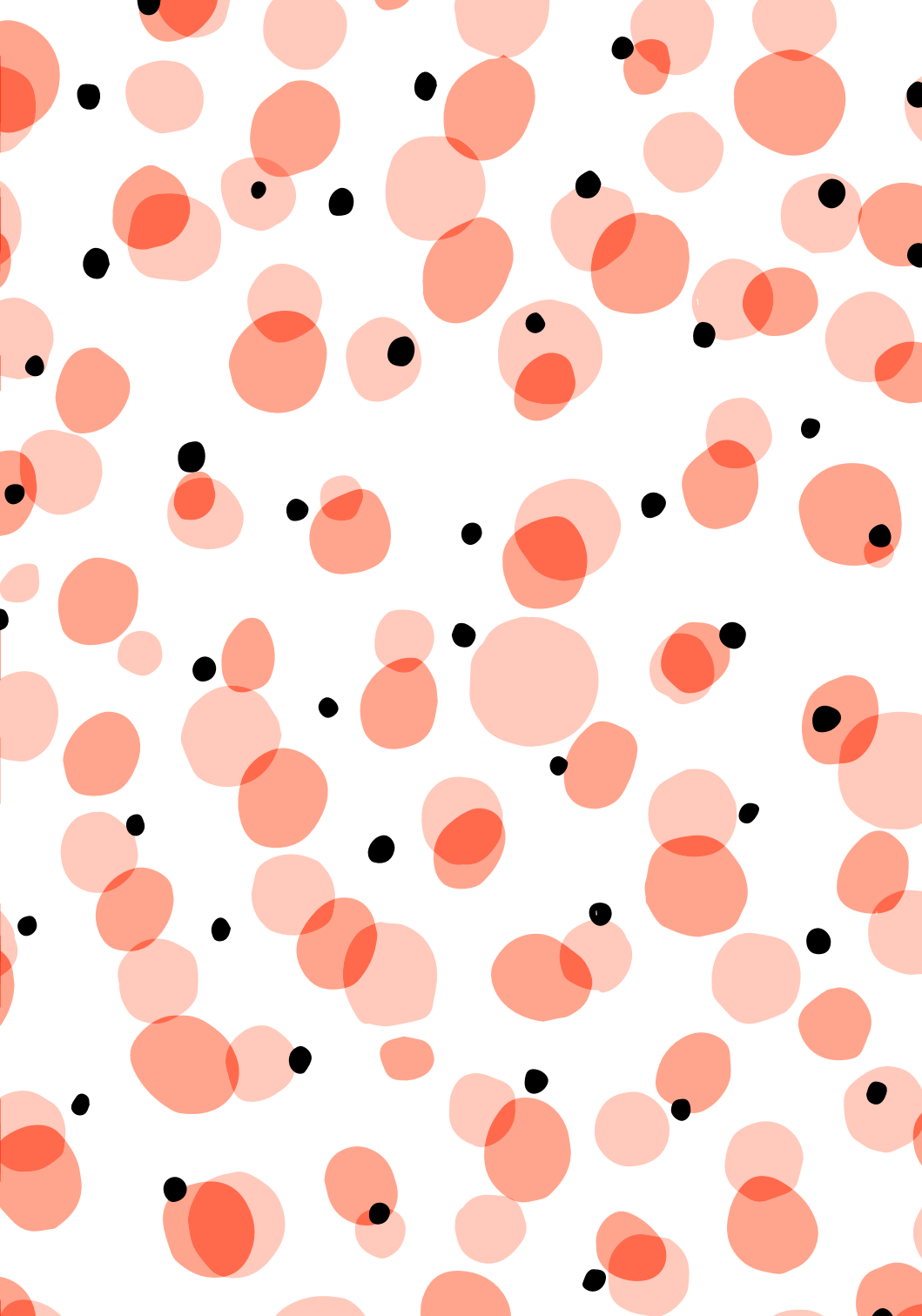
Floridi, Luciano. 2014. "Open Data, Data Protection, and Group Privacy." Philosophy & Technology 27(1): 1–3. https://doi.org/10.1007/s13347-014-0157-8.

Jaquet-Chiffelle, David-Olivier. 2008. "Reply: Direct and Indirect Profiling in the Light of Virtual Persons. To: Defining Profiling: A New Type of Knowledge?" In Profiling the European Citizen, edited by Mireille Hildebrandt and Serge Gutwirth, 17–45. Dordrecht: Springer.

Pasquale, Frank. 2015. The Black Box Society. Cambridge, MA: Harvard University Press.

Taylor, Linnet, Luciano Floridi, and Bart van der Sloot, eds. 2017. Group Privacy: New Challenges of Data Technologies. Dordrecht: Springer.





A LEGAL RESPONSE TO DATA-DRIVEN MERGERS

ORLA LYNSKEY

Personal data lead a double life in digital society: they are a digital reflection of our physical and spiritual selves while also being of economic value, given the revealing insights and predictions that are gleaned from them. Amassing a large volume and variety of data is therefore a core commercial objective of firms involved in data analysis and profiling.

This 'data grab' is facilitated in two primary ways. First, it is legitimised by the data protection framework, for instance when individuals consent to personal data processing or such processing is necessary for the performance of a contract. These legal bases for the collection of large volumes and varieties of data have long been, rightly, criticised by data protection advocates but their limits – including whether an individual can be excluded from a service if she refuses consent - are only now being tested before administrative authorities and courts. The fruit of these challenges has yet to be borne. Similarly, the principle of 'data minimisation', key to ensuring the proportionate aggregation of data, has thus far remained dormant, although calls mount for it to be rendered operational (see, for instance, Hildebrandt 2018).

The second way in which such a data grab is facilitated – through mergers and acquisitions – has received less critical attention. Competition law is the main legal instrument available to public authorities to curtail the exercise of private power. Yet, competition law remains generally unconcerned with the *acquisition* of power, provided that power does not significantly impede effective competition on relevant markets. Thus, merger control provides a facilitating framework for mergers motivated by data acquisition (data-driven mergers). This is despite the fact that such mergers often undermine data protection policy.

This provocation acknowledges that competition law cannot be instrumentalised to achieve data protection aims yet argues that it should not be applied in a manner that actively hinders the achievement of effective data protection, as is currently the case (section 1). Rather, there must be realistic and holistic oversight of data-driven mergers to limit the long-term implications of such transactions on the effectiveness of data protection rights (section 2).

The role of competition law in undermining data protection

Competition law is designed to promote consumer welfare, which is enhanced when consumers receive lower price and better quality products as well as more choice and innovation. Yet, mergers and vigorous competition limit choice by reducing the number of firms operating on any given market. Thus, it is arguable that competition law does not seek to preserve choice as such. Rather, competition authorities only intervene in acquisitions where effective competition is likely to be significantly hindered. Data protection and privacy concerns frequently fall into the blind spot of such economic analysis, as the *Facebook/WhatsApp* transaction vividly illustrates.

That transaction received attention on both sides of the Atlantic. The US Federal Trade Commission approved the merger, subject to the proviso that WhatsApp continue to honour its existing commitments to users (FTC 2014). In practice, this meant that

WhatsApp users were given notice, that Facebook would transfer their names and phone numbers to Facebook, and a choice – a take-it-or-leave-it choice. In the EU, the transaction was similarly approved, with the Commission noting that 'any privacy-related concerns flowing from the increased concentration of data within the control of Facebook…do not fall within the scope of the EU competition law rules but within the scope of the EU data protection rules' (*Facebook/WhatsApp* 2014, para 164). Data protection and privacy concerns stemming from the data aggregation were overlooked, with two elements of the Commission's reasoning meriting further attention.

First, while it may seem trite to state, where firms are not competing, mergers will ordinarily not be deemed to lessen effective competition. Thus, for instance, the Commission held that as consumers 'multi-home' (by using several applications) on the market for consumer communications applications, the acquisition would not negatively impact upon competition on that market. Pursuant to such logic, the acquisition by data giants of firms in markets in which they are not yet operating, or where they do not yet face a competitive constraint, would not be problematic. A good example is Google's acquisition of mapping company Waze, approved in the UK on the grounds that Waze did not yet exercise a competitive constraint on Google. Such reasoning facilitates defensive acquisitions of nascent competitors before they reach the scale to disrupt the status quo in the market (Stucke and Grunes 2016). Facebook's acquisition of WhatsApp, a 'maverick' firm offering individuals superior data protection, could be viewed from this perspective. The Commission did not however examine the impact of the merger on the future quality of the 'privacy' offered by both, despite acknowledging that Facebook Messenger and WhatsApp could be differentiated on the basis of privacy policies (*Facebook/WhatsApp* 2014, para 102).

Secondly, personal data are treated solely as an economic asset, with the proliferation of data viewed positively. A key concern in data-driven mergers is that the aggregation of data by the merging parties will constitute a barrier to market entry for potential competitors. Thus, competition authorities often consider whether potential competitors will have difficulty accessing a sufficient volume and variety of data following the transaction. Viewed through a competition lens, the availability of data on secondary markets, through data brokers and other sources, is a boon while such availability is difficult to square with core data protection principles such as purpose limitation and data minimisation. In *Facebook/WhatsApp*, the Commission held that Facebook's use of WhatsApp data to improve its targeted advertising would not give Facebook a competitive boost as a large quantity of valuable Internet user data for advertising purposes would continue to be available post-merger.[1]

Not only does such reasoning overlook data protection concerns, the end result is that one limb of digital policy – competition law – poses a significant challenge to the effectiveness of another limb – data protection. While proponents of a clear delineation between these legal spheres argue that the ex post application of data protection law should suffice to ensure data protection, such reasoning fails to acknowledge,

and even exacerbates, well-documented structural impediments to individual choice (Solove 2013; Lazaro and Le Métayer 2015). In this context, while individuals may 'consent' to having the merged entity aggregate their data, and continue to use its services, their choice – and control over their personal data – is nevertheless curtailed. Indeed, consent in situations of power asymmetry must be treated with caution (Borgesius et al 2017). The GDPR has the potential to improve the status quo and to deliver individuals more effective data control however, as noted above, its provisions have yet to be interpreted by courts. Furthermore, while technically free to abstain from, for instance, using Google products and services, such freedom requires considerable resources and initiative.

An assessment of the externalities of data-driven mergers

As the European Data Protection Board recognises, the time has come 'to assess longer-term implications for the protection of economic, data protection and consumer rights whenever a significant merger is proposed' (EDPB 2018). This provocation argues that such an assessment requires both a realistic and a holistic approach to data-driven mergers.

To date, most competition authorities have remained wilfully oblivious to the data protection implications of merger transactions, claiming that data protection law should remedy any eventual data protection concerns. Yet, there are avenues available to incorporate data protection into existing competitive analysis, without extending or distorting the aims of competition law.

First, there is now wide recognition that the level of data protection offered to individuals is an element of product or service quality and a competitive parameter on which companies can compete. How such quality if affected by a merger can be examined.

Secondly, when making competitive assessments, competition authorities take into consideration the existing legal framework on a market as contextual background. Data protection forms part of this legal landscape and should be taken into consideration in this way. Yet, the mere existence of data protection regulation should not lead to the assumption that existing market structures reflect individual preferences. In 2012 Farrell, a competition economist, documented a 'dysfunctional equilibrium' on data-driven markets, confirmed more recently by Which? (a UK consumer organisation). Which? suggests that some individuals are 'rationally disengaged': the perceived benefits of searching for data-protection friendly services are outweighed by the costs, making it rational for individuals not to engage in this search (Which? 2018). The GDPR's stronger substantive protection and more effective enforcement provisions have the potential to disrupt and reconfigure this vicious cycle. Competition authorities should nevertheless engage with behavioural economic analysis in order to understand how consumer choice is actually exercised and what inhibitors influence decision making (Fletcher 2017). The law in practice often deviates from the law on the books and therefore, for instance, the mere possibility of data portability under the GDPR does not necessarily prevent 'lock-in' to particular digital services (as assumed by the *Commission in Sanofi/Google/DMI*).[2]

In addition to being realistic when assessing data-driven mergers, it is also necessary to take a holistic approach to such transactions. Competition agencies are not well placed to do this: each notified transaction is examined on its facts alone and many transactions are never notified to these agencies if the turnover of the acquired company is not significant (as is frequently the case with technology start-ups). Thus, technology giants have been able to engage in a strategy of incremental acquisition that when viewed collectively paint a concerning picture of consolidation. Since the early 2000's Google, for instance, has acquired many of its household brands through takeovers: Android; YouTube; Doubleclick; Deepmind; and Nest Labs, to name but a handful (Reynolds 2017).

Recommendation

In light of the complex economic, social and political ramifications of personal data aggregation, a more cautious approach to data-driven mergers must be adopted. This could entail a moratorium on all acquisitions by certain technology giants, as proposed by the Open Markets Institute (Lynn and Stoller 2017). A preferred solution, hinted at by the EDPB in its statement on economic concentration, would be to subject data-driven mergers to a separate 'non-competition' assessment running in parallel to the competitive assessment. Such an assessment is currently afforded to media mergers in many countries, in recognition of the broader implications such transactions can have on media plurality. Given that the volume and variety of data aggregated by technology companies similarly entails broad societal implications, the case for a similar non-competition assessment in data-driven mergers must now be made.

Notes

[1] Case M.7217 – Facebook/WhatsApp [2014] OJ C-417/4, para 189.

[2] Case M.7813 - M.7813 - Sanofi/Google/DMI JV [2016] OJ C-112/1, para 69.

References

Borgesius, Frederik, Sanne Kruikemeier, Sophie Boerman and Natali Helberger. 2017. "Tracking Walls, Take-It-Or-Leave-It Choices, the GDPR, and the ePrivacy Regulation". European Data Protection Law Review 3(3): 353-68.

European Data Protection Board, "Statement of the EDPB on the data protection impacts of economic concentration" August 27, 2018. https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_statement_economic_concentration_en.pdf.

Farrell, Joseph. 2012. "Can privacy be just another good?" Journal on Telecommunications and High Technology Law 10: 251-64.

Fletcher, Amelia. 2017. "Exploitation of Consumer Decision-Making and How to Address it: Lessons from Past Demand-Side Interventions." Journal of European Competition Law and Practice 8, no. 8: 517-23.

FTC, "FTC notifies Facebook, WhatsApp of privacy obligations in light of proposed acquisition" April 10, 2014, https://www.ftc.gov/news-events/press-releases/2014/04/ftc-notifies-facebook-whatsapp-privacy-obligations-light-proposed.

Hildebrandt, Mireille. 2018. "Primitives of Legal Protection in the Era of Data-Driven Platforms." Georgetown Law Technology Review 2(2), 252-273.

Lazaro, Christophe, and Daniel Le Métayer. 2015. "Control over Personal Data: True Remedy or Fairy Tale?" SCRIPTed 12(1): 3-34. https://script-ed.org/?p=1927.

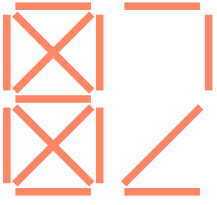
Lynn, Barry, and Matt Stoller. 2017. “How to stop Google and Facebook from becoming even more powerful.” The Guardian, November 2, 2017.

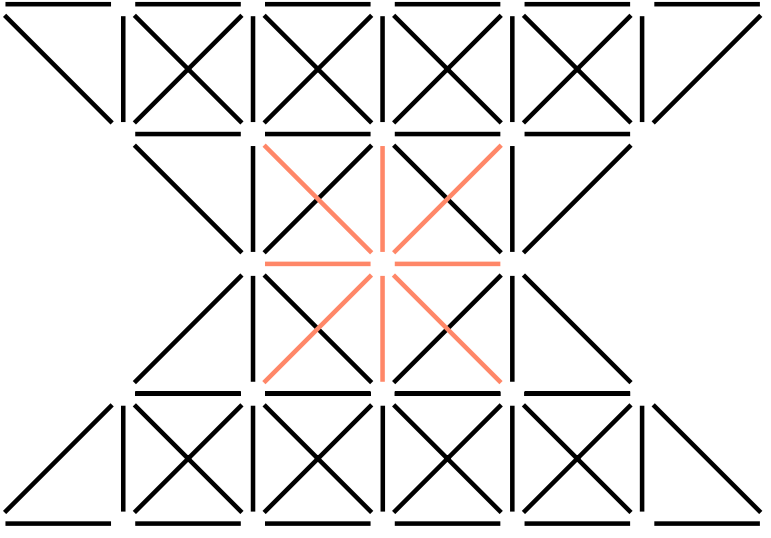
Solove, Daniel. 2013. “Privacy Self-Management and the Consent Dilemma.” Harvard Law Review 126: 1880-903.

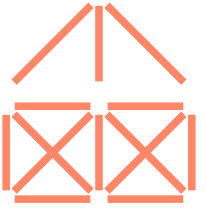
Stucke, Maurice, and Allen Grunes. 2016. Big Data and Competition Policy. Oxford: Oxford University Press.

Reynolds, Matt. 2017. “If you can’t build it, buy it: Google’s biggest acquisitions mapped.” Wired Magazine, November 25, 2017. https://www.wired.co.uk/article/google-acquisitions-data-visualisation-in-foporn-waze-youtube-android.

Which? “Control, Alt or Delete? The Future of Consumer Data”. June 4, 2018. https://www.which.co.uk/policy/digitisation/2659/control-alt-or-delete-the-future-of-consumer-data-main-repor.







ETHICS AS AN ESCAPE FROM REGULATION, OR ETHICS-WASHING 1 TO ETHICS-SHOPPING?2

A strange confusion among technology policy makers can be witnessed at present. While almost all are able to agree on the common chorus of voices chanting 'something must be done,' it is very difficult to identify what exactly must be done and how. In this confused environment it is perhaps unsurprising that the idea of 'ethics' is presented as a concrete policy option. Striving for ethics and ethical decision-making, it is argued, will make technologies better. While this may be true in many cases, much of the debate about ethics seems to provide an easy alternative to government regulation. Unable or unwilling to properly provide regulatory solutions, ethics is seen as the 'easy' or 'soft' option which can help structure and give meaning to existing self-regulatory initiatives. In this world, 'ethics' is the new 'industry self-regulation.'

Rigorous ethical approaches?

This approach does not do justice to many of the proponents of ethical approaches to technology who think long and hard about ethical frameworks for technology development. It is however indicative of the increasingly common role of technology ethics in political debates. For example, as part of a conference panel on ethics, one member of the Google DeepMind ethics team emphasised repeatedly how ethically Google DeepMind was acting, while simultaneously avoiding any responsibility for the data protection scandal at Google DeepMind (Powles and Hodson 2018). In her understanding, Google DeepMind were an ethical company developing ethical products and the fact that the health data of 1.6 Million people was shared without a legal basis was instead the fault of the British government. This suggests a tension between legal and ethical action, in which the appropriate mode of governance is not yet sufficiently defined.

Ethics / rights / regulation

Such narratives are not just uncommon in the corporate but also in technology policy, where ethics, human rights and regulation are frequently played off against each other. In this context, ethical frameworks that provide a way to go beyond existing legal frameworks can also provide an opportunity to ignore them. More broadly the rise of the ethical technology debate runs in parallel to the increasing resistance to any regulation at all. At an international level the Internet Governance Forum (IGF) provides a space for discussions about governance without any mechanism to implement them and successive attempts to change this have failed. It is thus perhaps unsurprising that many of the initiatives proposed on regulating technologies tend to side-line the role of the state and instead emphasize the role of the private sector. Whether through the multi-stakeholder model proposed by Microsoft for an international attribution agency in which states play a comparatively minor role (Charney et al. 2016), or in a proposal by RAND corporation which suggests that states should be completely excluded from such an attribution organisation (Davis II et al. 2017). In fact, states and their regulatory instruments are increasingly portrayed as a problem rather than a solution.

Case in point: Artificial Intelligence

This tension between ethics, regulation and governance is evident in the debate on

artificial intelligence. To provide just one example, here the position of the European Commission is most telling, in which it states that:

> *Draft AI ethics guidelines will be developed on the basis of the EU's Charter of Fundamental Rights, following a large consultation of stakeholders within the AI Alliance. The draft guidelines will build on the statement published by the European Group of Ethics in Science and New Technologies (European Commission 2018a).*

This statement is so confusing on numerous levels that it deserves a closer analysis. The EU intends to build ethics guidelines on the basis of the existing EU Charter of Fundamental Rights. However, if this is the EU's intention, why not simply call for the implementation of fundamental rights in digital technologies?

At the same time, the ethics guidelines will also "build on" the recommendations of the work of the main EU body on ethics - The European Group on Ethics in Science and New Technologies (EGE) - which have developed a set of 'Ethical principles and democratic prerequisites' as part of their report on the Ethics of Artificial Intelligence (European Group on Ethics in Science and New Technologies (EGE) 2018). The principles developed by the EGE cover numerous aspects related to fundamental rights such as human dignity, but also introduce completely unrelated aspects such as sustainability, while entirely leaving out other aspects such as freedom of assembly or cultural rights.

From fundamental rights to potential rights

This leads to a considerable blurring of lines in regard to both ethics and rights. Ethics—even in an applied sense—is distinct from the law and human rights. At the same time EU fundamental rights are not understood as fundamental rights but rather as ethical imperatives to be complied with in a non-binding fashion. While admittedly the European Commission does threaten more strict regulation of AI, it does not specify under what conditions this would take place or what this legislation would look like. Such legislative specification is however urgently necessary.

In this sense these are 'potential fundamental rights', developed under the shadow of hierarchy of the European Commission. They certainly cannot be claimed at present and if these potential fundamental rights are 'violated' (whatever that means in the context of ethical commitments to uphold fundamental rights) they would be no legal recourse of any kind available. Indeed, it is in fact likely that these rights would actively need to be violated frequently and these violations would need to be made public widely, in order for the European Commission to be willing to do anything about their actual violation. In that sense, these potential rights serve as an inspiration for potential action rather than a commitment to their implementation.

The same confusion applies to any potential ethical behaviour based on these potential fundamental rights. Should actors who wish to uphold such an ethical

framework actively violate the rights frequently in order to ensure that the European Commission turns potential rights into actual fundamental rights? How should they act ethically under a shadow of hierarchy expecting their conformity? The EGE acknowledges at least some of these challenges in suggesting that there is a danger of 'ethics shopping' in the approach followed by the European Commission, in which "regulatory patchworks" (European Group on Ethics in Science and New Technologies (EGE) 2018, 14) are seen as the source of this problem. In this context language, AI ethics are essentially a quasi-binding instrument, which will be made binding only if it is sufficiently violated.

Beyond myths of law, ethics and technology

In a masterful book on Technology and the Trajectory of the Myth, David Grant and Lyria Bennet Moses argue thinking about the law as "more than simply a 'roadblock' on the road to greater technological innovation" (Grant and Moses 2017, 215). While acknowledging that this is the case, ethics is evidently more than a value-laden framework to pre-empt or evade the law. Both law and ethics exist in parallel and can contribute to positively influencing human behaviour. Where they can and should meet is in the design process of technologies, which itself can enable certain forms of human behaviour. Here the idea of Value based Design, Privacy by Design (Cavoukian 2009), Legal Protection by Design (Hildebrandt 2016), Human Rights Based Communications Infrastructure (Wagner 2012) and Ethical Design (Balkan 2017) align to a considerable degree on many of their practical recommendations for the development of technology. Evaluating Rights and Ethics in Practice.

Yet the possibility to implement these solutions in technical design does not answer a more difficult question: how then to differentiate the many ethical frameworks out there and decide which are more likely to deliver appropriate ethics? How to ensure that ethics shopping or ethics washing does not become the default engagement with ethical frameworks or rights-based design?

Broadly speaking, I argue that it is possible to differentiate between 'thin' and 'thick' approaches to technology design and development, regardless of whether these are ethical or human-rights based. In order for these ethical approaches to be taken seriously as 'thick' approaches they should at minimum conform to the following basic criteria:

1. External Participation: early and regular engagement with all relevant stakeholders.
2. Provide a mechanism for external independent (not necessarily public) oversight.
3. Ensure transparent decision-making procedures on why decisions were taken.
4. Develop a stable list of non-arbitrary of standards where the selection of certain values, ethics and rights over others can be plausibly justified.
5. Ensure that ethics do not substitute fundamental rights or human rights.
6. Provide a clear statement on the relationship between the commitments made and existing legal or regulatory frameworks, in particular on what happens when the two are in conflict.

While this list is relatively straightforward, many initiatives are not able to respond to these challenges. As has been discussed above both attempts at developing ethical technologies by Google DeepMind and AI ethics guidelines by the European Commission have not managed to address many of the challenges above. This is particularly confusing as in other areas like the profiling of European citizens (Hildebrandt and Gutwirth 2008), the EU takes a much stronger regulatory fundamental rights-based approach. This approach is most prominently found in the EU's General Data Protection Regulation (GDPR) and has many similarities to the 'thick' approach to technology design and development described above.

Thus, is a world in which ethics-washing and ethics-shopping are becoming increasingly common, it is important to have common criteria based on which the quality of ethical and human rights commitments made can be evaluated. If not, there is a considerable danger such frameworks become arbitrary, optional or meaningless rather than substantive, effective and rigorous ways to design technologies. When ethics are seen as an alternative to regulation or as a substitute for fundamental rights, both ethics, rights and technology suffer.

Notes

[1] Chris Marsden at the 20th FIPR Conference in Cambridge on 29 May 2018: http://youtu.be/LRiAcbvSA3A?t=1h8m20s.

[2] Acknowledgements: I am very grateful for the excellent comments received by Mireille Hildebrandt, Sarah Spiekermann, Linnet Taylor, Emre Bayamlıoğlu and the excellent seminar at Vrije Universiteit Brussel (VUB) on 10 Years of Profiling the European Citizen for which this article was originally developed.

References

Balkan, Aral. 2017. "Ethical Design Manifesto." Ind.Ie. 2017. https://2017.ind.ie/ethical-design/.

Cavoukian, Ann. 2009. "Privacy by Design: The 7 Foundational Principles. Implementation and Mapping of Fair Information Practices." Information and Privacy Commissioner of Ontario, Canada. https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf.

Charney, Scott, Erin English, Aaron Kleiner, Nemanja Malisevic, Angela McKay, Jan Neutze, and Paul Nicholas. 2016. "From Articulation to Implementation: Enabling Progress on Cybersecurity Norms."Microsoft Corporation, June 2016. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/REVmc8.

Davis II, John S., Benjamin Boudreaux, Jonathan William Welburn, Jair Aguirre, Cordaye Ogletree, Geoffrey McGovern, and Michael S. Chase. 2017. Stateless Attribution. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2081.html.

European Commission. 2018a. "A European Approach on Artificial Intelligence". http://europa.eu/rapid/press-release_MEMO-18-3363_en.htm.

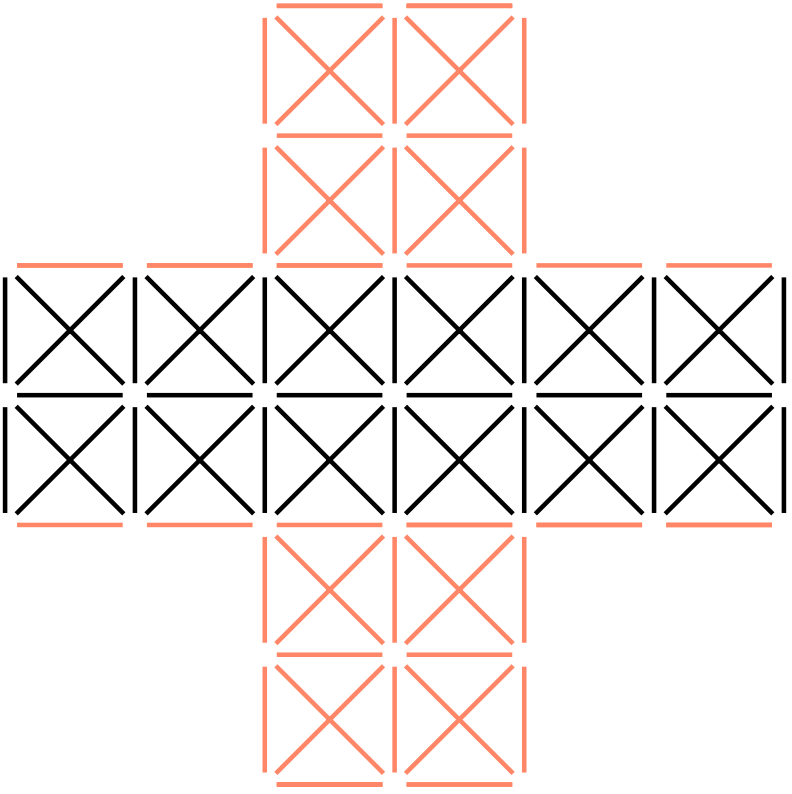
———. 2018b. 'Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions (COM(2018) 237 Final)'. http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51625.

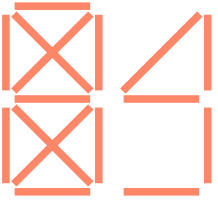
European Group on Ethics in Science and New Technologies (EGE). 2018. "Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems". https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

Grant, David, and Lyria Bennett Moses. 2017. Technology and the Trajectory of Myth. Cheltenham: Edward Elgar Publishing.

Hildebrandt, Mireille. 2015. Smart Technologies and the End(s) of Law. Novel Entanglements of Law and

Technology. Cheltenham: Edward Elgar Publishing.
Hildebrandt, Mireille, and Serge Gutwirth, eds. 2008. Profiling the European Citizen: Cross-Disciplinary Perspectives. Dordrecht: Springer Science.
Powles, Julia, and Hal Hodson. 2018. 'Response to DeepMind'. Health and Technology 8(1–2): 15–29. doi: 10.1007/s12553-018-0226-6.
Wagner, Ben. 2012. 'After the Arab Spring: New Paths for Human Rights and the Internet in European Foreign Policy'. Brussels, Belgium: European Union. doi: 10.2861/9556.





CITIZENS IN DATA LAND

ARJEN P. DE VRIES

My provocation in the panel on *Legal and political theory in data driven environments* at the workshop '10 Years of Profiling the European Citizen' started with a quote from the closing chapter of Profiling the European Citizen (Hildebrandt and Gutwirth 2008):

> *For individual citizens to regain some kind of control over the way they live their lives, access is needed to the profiles applied to them. This will require both legal (rights to transparency) and technological tools (the means to exercise such rights).*

Looking at progress with respect to these two requirements, European citizens have been successful in creating a legal framework that gives people the power to claim substantial rights in their personal data. Even if we have not yet gained much experience with the law being tested on its practical usefulness, serious restrictions have been imposed upon the parties that control the processing of personal data (e.g., data minimisation, data portability). Switching our perspective to the technological tools however, I am much less optimistic. Wouldn't it be so much easier to exercise our right on e.g. data portability if we actually knew who has our data, in what form, on what server, and how to access and manipulate that data – and not merely transfer this data from one service that we do not control to yet another one?

Profiling

Take a look at the original rendering of my provocation for the online workshop proceedings:

Arjen P. de Vries Citizens in Data Land

Profiling

The informed reader has recognised the use of TEX and infers correctly that this provocation is written by a computer scientist. The author is indeed

As you read in the Figure already, the informed reader would recognise immediately the use of the LaTeX typesetting system and infer, correctly, that this provocation is written by a computer scientist.[1] The author is indeed trained as computer scientist and the first thing he had to do upon receiving the invitation to join the workshop with a provocation was to look-up the meaning of that term, using a search engine (I might as well share my ignorance with you, the reader, given that *I shared this information already with one of the largest tech companies in the world*). The title of the panel revealed more gaps in my background knowledge, because my immediate association with "political theory" is the title of a Coldplay song. Wikipedia came to the rescue, although I would tell my students not to simply rely on the information in the online encyclopaedia when it concerns my area of expertise... At this point in my provocation, you know most of the information about me that you would have learned also from

my bio on one of the various social media sites where I have an account.[2]

Now, the simple fact that you can find this personal information about me via a web search by name (you need to include the middle initial) is no issue of concern; the bio is a public self-description I contributed voluntarily to the online world, as a 'citizen of data land', advertising why to connect to me. What does (and should) raise objections is the detailed information that I gave away implicitly, mostly unaware, through usage of online services such as the search engine. And it is not easy to escape hidden forms of profiling if I want to stay a 'citizen of data land'; a recent analysis of the CommonCrawl 2012 corpus found that the majority of sites contain trackers, even if websites with highly privacy-critical content are less likely to do so (60% vs 90% for other websites) (Schelter and Kunegis 2018). I learned from an independent blogger that her commissioning parties demand Google Analytics based statistics: to generate any income as an online writer, sharing visit data from your blogging site with Google has become a *de facto* prerequisite, even if you keep your site free from advertisements. The way the Web has evolved, accessing online information implies being profiled.

Civic responsibility in 'data land'

Will the new legal rights (transparency and control) help enforce a new balance? We should not sit back and expect the GDPR to save our privacy from organisations' hunger for data. If only 'citizens of data land' had the means to take control of their data, including the traces they leave online; alas, we have seen less progress with regard to the technological tools necessary to exercise our new rights.

The current situation is that 'we the people' give those who run online services a *carte blanche* to collect our data. The legal framework will make this collection more transparent (we hope), but it cannot change the status quo if we do not act ourselves. It is – to a large extent – our own personal choice (if not to say mistake) that we let a few, very large and omnipresent organisations build their business model on harvesting personal data *en masse*.

If we do not modify our online behaviour, the GDPR creates an improved legal context, sure; but the balance of power between individual citizens and the (public and private) organisations they deal with online shifts back just a tiny fraction of how it could shift back to the citizen, if only we were more responsible in taking care of our data.

Our data, our devices

We have been seduced to give up, voluntarily, the control over our personal data, in exchange for convenience: the convenience of having services managed for us, in the cloud, seemingly for free. We give away our data without much consideration of their value, or the long-term consequences of doing so. We might try to claim back our data with the re-gained legal rights, or at least exercise control over the ways our data is used – but would it not be so much easier to "simply" keep our data for ourselves?

We create our personal data ourselves, and, at least initially, on our own devices.

Instead of handing over that data to an external organisation that runs an information service for us, I put my cards on two design principles to help establish a renewed, better balance, where the people who create the data exercise a significantly larger degree of ownership over their data.

Personal web archives

The first principle is to build systems for online information interactions such that they **keep data where it originates:** in your own device.

As a proof of concept, consider the personal web archive and search system called WASP,[3] that archives and indexes all your interactions with the Web and enables effective re-finding (Kiesel et al. 2018). Those searches remain completely local (and therefore private). While WASP did not yet address the case of a user managing multiple devices (like a smartphone and a desktop computer), this is resolved with Prizm, a small personal device that acts as a gatekeeper between your edge devices and the outside world (Lin et al. 2016).

A more radical version of the design principle (of keeping all your personal Web interactions local) would be to expand those interactions, as a seed to a personal crawl that captures also the information for highly likely future interactions, while also storing a significant fraction of the Web as a snapshot local to your device, instead of in your favourite search engine's data centres.

Practical implementation of this idea raises many interesting technical questions (exciting for the computer scientist in me), where I imagine a role for commercial and/or non-profit organisations too. They could, for instance, package recent web crawls for distribution, sliced per topic of interest.[4] People could then subscribe to regular updates of their own personal search engine index without the need to crawl the Web themselves; the GDPR helps us trust those organisations to keep subscription information private and secure.

Decentralised social media

Obviously, whenever we want to share information with others, we cannot keep that data on our own infrastructure. The second design principle would therefore be to **decentralise online services** (or, better, to *re*-decentralise the Web).

The recent rise of decentralised alternatives to existing centralised social media services is especially promising. ActivityPub[5] is a W3C standard that has been granted the status of 'recommendation' (since January 23rd, 2018) and has already been implemented in an increasing number of open source projects. For example, Mastodon is essentially a 'decentralised version of Twitter' where ActivityPub facilitates the communication among thousands of Mastodon instances that together host over 1 million registered users. Other community projects have created decentralised alternatives for Instagram (PixelFed), YouTube (PeerTube), and Medium (Plume).

This cooperation of decentralised online services that exchange social information

using ActivityPub has been called the Fediverse (a partial blend of federated and universe). Members of the Fediverse interact freely with each other, even if their accounts reside on different so-called 'instances'. This enables communities to organise themselves, independent from large corporations that would like to collect this data in a huge centralised database. Examples of Mastodon instances that serve a community include the recent Mastodon instance created for 'all people with an email address from University of Twente', an MIT instance, and, an instance I created myself, aiming to be a new online home for the Information Retrieval community.[6]

Closing statement

The directions in which I seek a solution for better technological support are still a long way from empowering the 'citizens of data land'.

A hurdle to take is how to get these new solutions in a state so that 'data land' ends up under 'the rule of the people'. Managing your own personal data is a '21st century skill' that the 'citizens in data land' will have to master. If we do not pay attention, we end up replacing one 'aristocracy', of an elite of large tech corporations, by another one, consisting of tech savvy people who know how to operate their own data infrastructure, thus excluding others from exercising the same level of control over their data.

The exciting technological developments that underpin the two principles of data ownership and decentralisation create an opportunity to exercise a higher level of control over the decision as to who gains access to our data. However, we need to pay for this control in the form of an investment in personal computer infrastructure and the effort to acquire the skills to manage this infrastructure.

Are we, the people, willing to make that effort? Paraphrasing Hildebrandt and Gutwirth (2008, 365):

> *Citizenship, participation in the creation of the common good and personal freedom cannot be taken for granted, they presume that citizens 'acquire the competences to exercise control over what is known about them and by whom'.*

Notes

[1] The format of the text in the Figure is another, more subtle hint that the author might be a computer scientist.

[2] 'Computer scientist and entrepreneur. Information access & integration of IR and DB. And Indie music'.

[3] https://github.com/webis-de/wasp/.

[4] Consider a new service provided by The Common Crawl Foundation, http://commoncrawl.org/, or, alternatively, a new community service provided via public libraries.

[5] ActivityPub, https://www.w3.org/TR/activitypub/.

[6] Visit https://idf.social/ or https://mastodon.utwente.nl/ for more information.

References

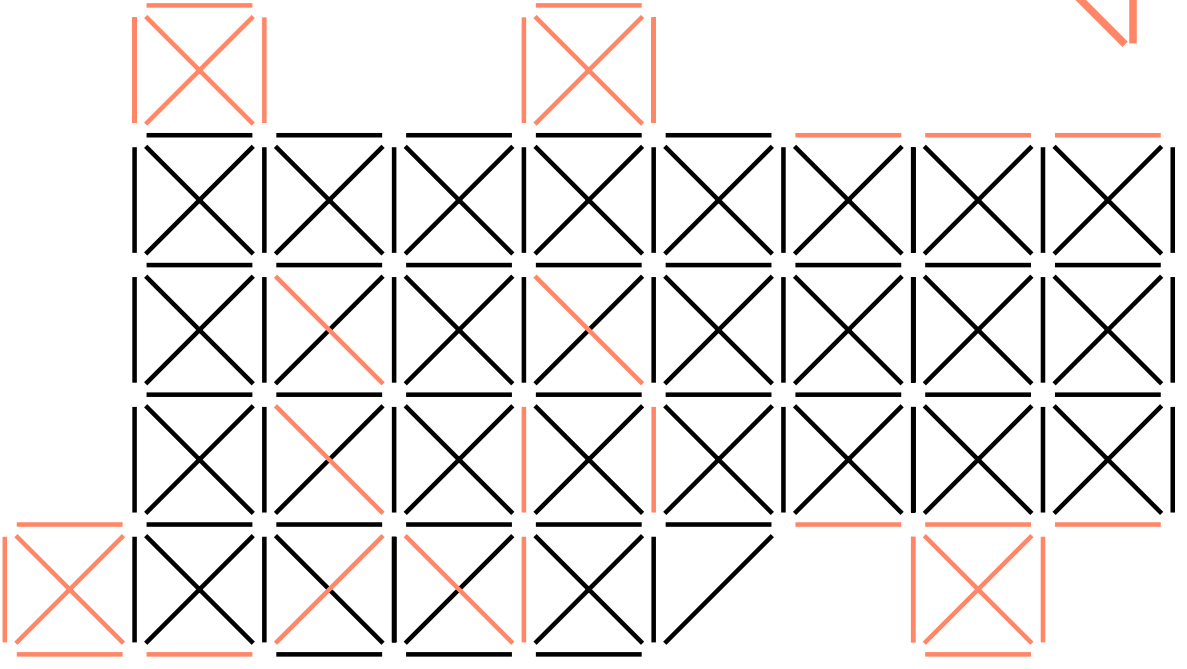
Hildebrandt, Mireille, and Serge Gutwirth. 2008. "Concise conclusions: Citizens out of control." In Profiling the European Citizen: Cross-Disciplinary Perspectives, edited by Mireille Hildebrandt and Serge Gutwirth,

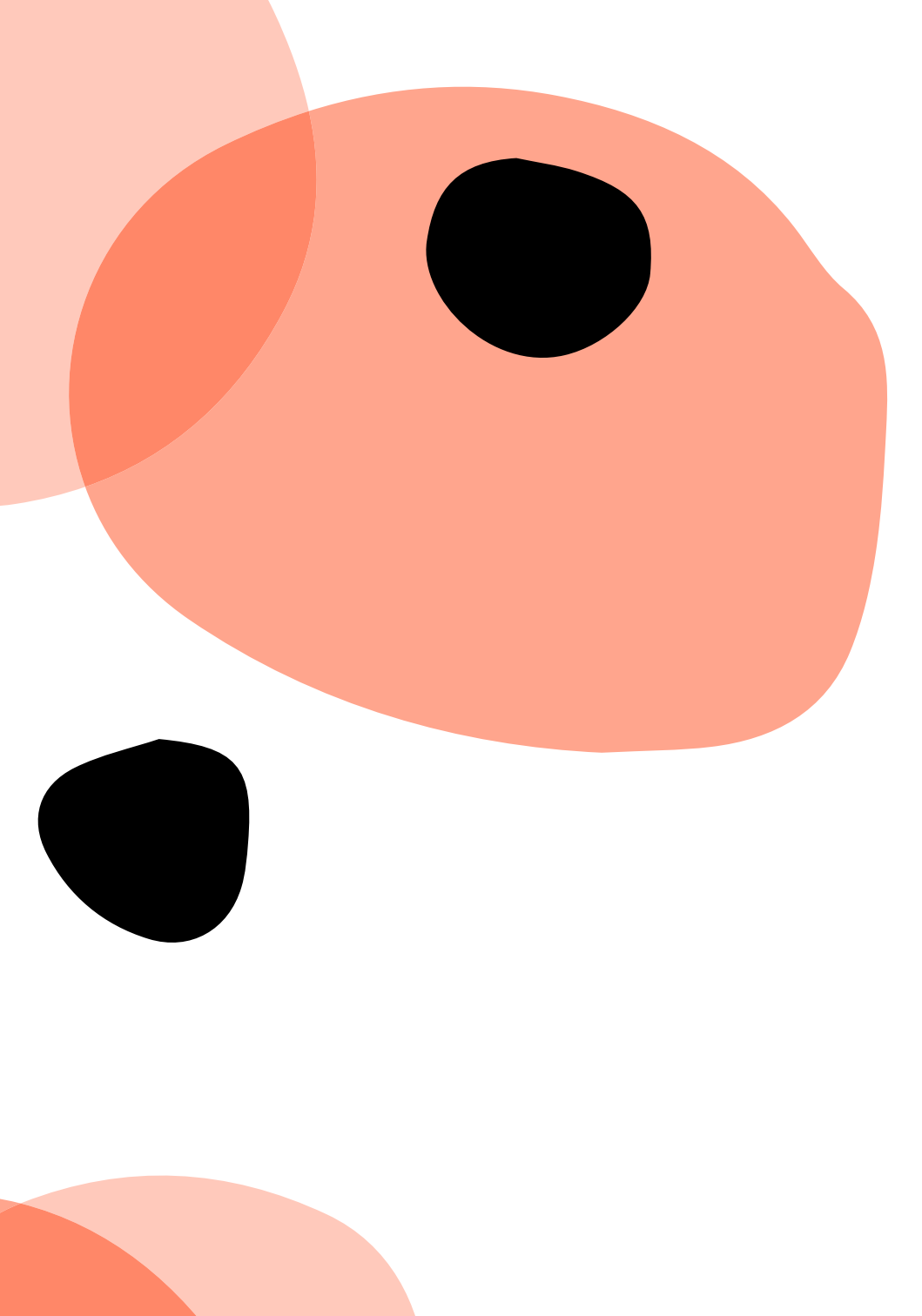
17-45. Dordrecht: Springer.

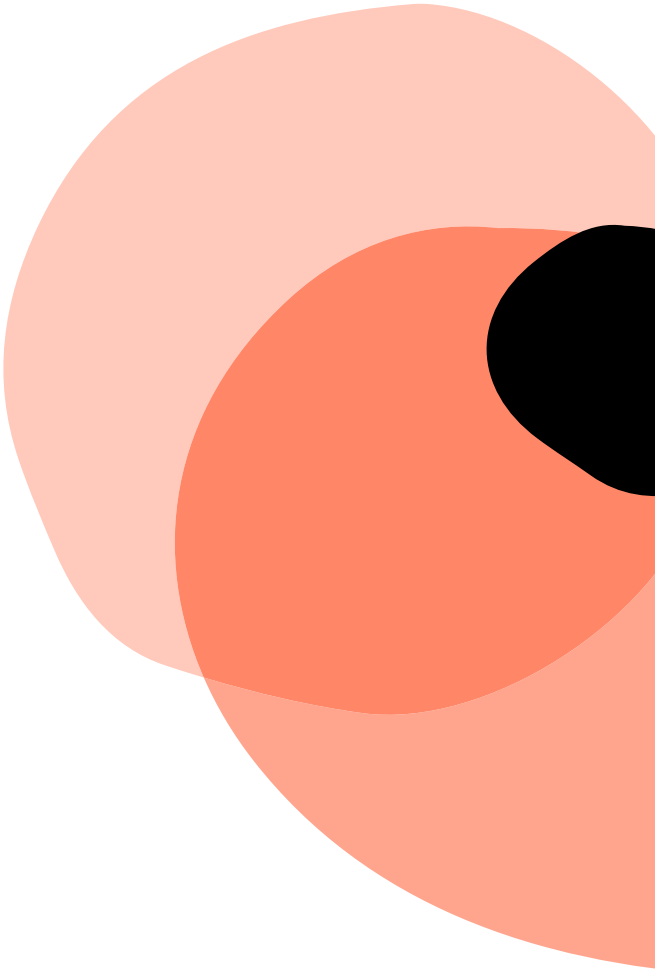
Kiesel, Johannes, Arjen P. de Vries, Matthias Hagen, Benno Stein, and Martin Potthast. 2018. "WASP: Web Archiving and Search Personalized." In Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems, Bertinoro, Italy, August 28-31, 2018. CEUR Workshop Proceedings 2167: 16-21.

Lin, Jimmy, Zhucheng Tu, Michael Rose, and Patrick White. 2016. "Prizm: A Wireless Access Point for Proxy-Based Web Lifelogging." In Proceedings of the first Workshop on Lifelogging Tools and Applications (LTA '16). ACM, New York, USA: 19-25.

Schelter, Sebastian, and Jérôme Kunegis. 2018. "On the ubiquity of web tracking: Insights from a billion-page web crawl." The Journal of Web Science 4(4): 53–66.







FROM INTER-SUBJECTIVITY TO WILDERSUBJECTIVITY: KNOWLEDGE CLAIMS AND THE DIGITAL CONDITION

FELIX STALDER

One of the consequences of digitization is a deepening crisis of epistemology, caused by the proliferation of social, biological and machinic actors that overwhelm established methods of generating and organizing knowledge (Stalder 2018). And, since there is a close relationship between epistemology and politics, between ways of knowing and ways of managing the world, we are also in a deep political crisis. This manifest itself not the least in a populist rejection of 'science' and 'facts' (Manjoo 2008). This crisis of the established – let's call it modern-liberal – epistemic-political order has created a space for the establishment of a new one, which doesn't yet have a name, even if its outlines are already visible.

The epistemology of the modern-liberal era

The basic structure of epistemic-political order that created the modern era in the West was established in the mid 17th century. Not only defined the peace treaty of Westphalia in 1648 the secular nation-state as the pinnacle of power and ultimate sovereign, but the Royal Society in London, founded in 1660, established a new mode of asserting matters of fact. Basically, matters of fact were henceforth to be asserted by the observation of independent individuals, organized as communities of peers. These communities were bounded in two respects. First, the domain of knowledge in which the peers could assert facts with authority was limited to what would later be called a scientific discipline, over time, these boundaries got ever more narrow as the number of disciplines increased. Second, it was bounded by an agreement on the methods of knowing, these methods would define the other dimension of 'discipline' (Schapin and Schaffer 1985).

The first boundary not only led to the establishment of different scientific disciplines, but also a separation of powers, so to speak, between science, politics and religion. Each with its own internal segmentation, but, above all, separated from each other. The second boundary, the agreement of methods, rather than on outcomes, constructed science not only as an open-ended enterprise capable of revising its own paradigms (Kuhn 1962), but also demanded from its practitioners that they had no interest in specific outcomes, rather that they would accept whatever the method yielded. And the results were to be accepted, if, and only if, other members of the community shared the same observation. This was made easier, or perhaps even possible at all, by the aforementioned separation of domains. The knowledge thus produced concerned the 'other', that is, nature and it was possible to be disinterested towards the 'other'. Thus, it became possible that, say, a Jewish Marxist chemist could easily reach consensus with, say, a Christian conservative chemist, as far as chemistry was concerned.

Along with the methods, a new place for the observation of nature was created, the lab. The main advantage of the lab was that it was a controllable environment, that is, in it, it was possible to reduce the complexity and isolate a limited number of relationships to be manipulated and observed in a reproducible manner. The fact that the natural environment outside the lab was far more complex was acknowledged through the formula of ceteris paribus, the assumption that while a set of elements were manipulated, 'all other things being equal.'

Thus, the modern scientific practice has been based on principles of 'inter-subjectivity' (the position of the observer played no role in the observation), 'distance' (in the double sense that the observer was disconnected from the observation and that the object of the observation, nature, was the 'other'), 'disinterestedness' (the results of the observation did not directly affect the observer) and 'reduction' (theory-driven separation of important from unimportant variables).

Limits of the modern epistemology

For a long time now, all of these principles have come under sustained critique (Lyotard 1984). Second-order cybernetics showed that the observer is part of the system that he/she/it is observing (Von Foerster 2003). Increased scope and complexity of the problems made the idea of the 'view from nowhere' highly questionable, if not impossible. But if the observer is inside the problem, then the position of the observer becomes crucial and resulting observation cannot be complete but is necessary partial and needs to be complemented with other partial observations. In cultural studies, this is called 'positionality', meaning that any statement, even statement of fact, is related to a position from which this statement is made (Hall 1990, 18). If the observer is inside the problem then the problem domain can no longer be constructed as the 'other'. This means that there can be no disinterested description, but matters of fact become, as Bruno Latour put it, "matters of concern" (Latour 2004). If we take climate science as an example, then every statement about the climate is also a statement about the society that is now understood as producing this climate. Hence every description becomes a prescription. Thus, the most urgent question turns from how an external object really 'is', to how we can, have to, or want to relate to it and in this relation, how we transform the very thing we are observing. Thus, the principle of knowledge moves from independent truth to dependent utility. Which immediately raises the question: useful for whom?

Last, but not least, with a rising dynamism and complexity, that is sharp increase in the number of actors interacting with one another and ways in which this interaction can unfold, the question becomes ever more critical: which variables are the important ones, and which are the one that can be ignored? The effects of this increasing difficulty of distinguishing between variables to include and variables to exclude are, on the one hand, a crisis of replicability that seems to be plaguing the sciences, and, on the other hand, the mounting costs of ignored variables reasserting themselves in things like climate change.

None of this is new. Second order cybernetics is from the early 1970s, Lyotard's observation of the transformation of science from seeking truth to creating utility is from 1980, Latour's fundamental critique of the modern epistemology that is separating of society and nature is from 1988.

Deepening the crisis and going beyond it

But what is new, then? Machine-driven analysis of large data sets is introducing a new way of doing science. In this, it is answering to this crisis while, at the same time, deepening it. This is the case even if it works according to its own program (I

will ignore here practical issues such as quality of data, issues of modelling and so on). For example, the claim to be able to processes large quantities of unstructured data, can be seen as avoiding the problem of reductionism. Rather than relying on a sample size of questionable representativeness, or on a controlled laboratory environment, or on theory-driven hypotheses, the approach (at least in its ideal) is to take in all data without any prior separation of important from unimportant aspects of the problem. This separation is done now through machine learning, and the less assumptions go into the processes, the higher the chance to find something new. Yet, the opacity and the complexity of the tools of analysis re-introduces problems of replicability with a vengeance. Because, the problem of reductionism has turned into a more fundamental problem of method, the very core of science itself. By focusing on 'relations that work' (while continuously adapting the question until it yields a statistically significant answer), on utilitarian effects (accurately predicting the short-term), rather than fundamental causation, machine driven analysis dispenses with the notion of a disinterested search for an external truth and fully concentrates on relationships that can be manipulated for pre-determined ends. But since the actor who does the analysis – most clearly in the case of social media companies – is a core element of the situation he/she/it is analysing, and is thus inside the problem rather than outside of it, result of the analysis can immediately be fed back into the situation changing its composition or dynamics. From the point of view of the company paying for the research, this is not a bug, but a feature.

In some way, this is an old problem of the social sciences, now on steroids. Max Weber argued already that what distinguishes social science from other forms of research is that the ideas people have about society, in part derived from social science, affect the dynamics of society. Noortje Marees (2017) sees this kind of 'interactivity' as one of the core elements of new field of digital sociology. This problem seems to plague ever more sciences because of the aforementioned breakdown of separation between scientific process and the object of analysis. Machine-driven analysis takes this as a starting point, accelerating the processes by feeding its results back into the 'object' and claims to overcome it by reducing the temporal scope of analysis making it, in effect, a continuous process, rather than a one-time event.

Acknowledging utility, positionality and partiality

This suggests to me that it might be more productive to think of machine-driven 'data science' as a new mode of knowing, one that breaks with fundamentals of scientific method that defined the modern-liberal era. This need not be a bad thing, because modern science produced not just knowledge, but also as Ulrich Beck (1992) observed, a lot of risk. Thus we need new methods that can deal with the dynamism and complexity of the problems we are not just facing, but in which we are in over more complex ways, also implicated in. There is a need to find new ways to make scientific facts transparent and democratically accountable. Rather than trying to defend traditional ideals of science – disinterestedness, distance, inter-subjectivity – we would acknowledge that science is ever more interested. This is not to advocate an 'anything goes' attitude, or a superficial relativism or post-modern claim about the constructedness of science, but it might be a first step to develop tools and methods to

account for the necessary positionality of any knowledge claim that concerns complex, dynamic systems in which the observer is directly implicated.

This is all the more urgent for political reasons. The number of actors who have access to very large data is sharply limited. In effect, nobody can do research on social media data the way Facebook can do it. And here, it's obvious that this research is interested and a source of social power. In such a context, claims of 'scientific objectivity' are likely to serve as a way to abdicate responsibility for the research and its consequences. To highlight the positionality and partiality of any claim, also and in particular in data science, would render more obvious the need to combine competing claims into new ways of understanding the world that is not so much inter- but rather multi-subjective. Each of these claims, in order to be understood as science, needs to be rigorous, fact-based and transparent to others, but they cannot claim to be disinterested or separated from outcomes.

References

Beck, Ulrich. 1992. Risk Society: Towards a New Modernity. London: Sage Publications.

Hall, Stuart. 1990. "Cultural Identity and Disapora." In Identity: Community, Culture, Difference, edited by Jonathan Rutherford, 222-37. London: Lawrence & Wishart.

Kuhn, Thomas. 1962. The Structures of Scientific Revolutions. Chicago: Chicago University Press.

Latour, Bruno. 2004. "Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern." Critical Inquiry, no. 30: 225–48.

Lyotard, Jean-François. 1984. The Postmodern Condition: A Report on Knowledge. Minneapolis: University of Minnesota Press.

Manjoo, Farhad. 2008. True Enough: Learning to Live in a Post-Fact Society. Hoboken, NJ: Wiley.

Marres, Noortje. 2017. Digital Sociology: The Reinvention of Social Research. Malden, MA: Polity.

Schapin, Steven, and Simon Schaffer. 1985. Leviathan and the Air-Pump: Hobbes, Boyle and the Experimental Life. Princeton, NJ: Princeton University Press.

Stalder, Felix. 2018. The Digital Condition. Cambridge, MA: Polity Press.

Von Foerster, Heinz. 2003. "Cybernetics of Cybernetics." In Understanding Understanding: Essays on Cybernetics and Cognition, 283–86. New York: Springer.

PREREGISTRATION OF MACHINE LEARNING RESEARCH DESIGN AGAINST P-HACKING

MIREILLE HILDEBRANDT

ML is based on the idea that intelligence concerns the ability to learn from experience, rather than the ability to apply ready-made knowledge. In that sense it favours inductive rather than deductive inferences. In the domain of artificial intelligence many voices now warn against overestimating the effectiveness of inductive learning, without however disqualifying its potential achievements (Brooks 2017, 2018; Marcus 2018).

The Mechanics of ML

It is interesting to note that human intelligence thrives on what Peirce called abductive inferences (Peirce and Turrisi 1997, 241-56), which are neither inductive nor deductive. Abductive inferencing basically entails an informed guess as to the explanation of a set of observations. Building on Peirce, scientific research can be framed as starting with an abduction based on observation, generating an explanation (theory) from which a hypothesis (prediction) is deduced about subsequent observations, after which the prediction can be inductively tested against new observations. Building on Popper's theory of falsification,[1] hypotheses should be developed in a way that enables the rejection of the explanation – not merely its verification. A theory that explains why all swans are white should not just be verified by detecting ever more white swans, but tested against its potential falsification by searching for black swans.

ML has been defined as 'improving automatically with experience' (Mitchell 1997, 1). More precisely '[a] computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E' (Mitchell 1997, 2). A crucial role is played by the so-called hypotheses space, i.e. a set of mathematical functions to be tested as accurate descriptions of patterns in the data on which the algorithm is trained (the training data). These hypotheses can be seen as abductively generated by the human developer of the ML research design, even if the system can be designed such that it generates further hypotheses (mathematical functions). Feeding the system with 'big data' can be seen as inductive testing. What is missing here, is the explanation. Normally, abduction generates an explanation, from which hypotheses can be deduced, which are then tested. In the case of ML, abduction does not concern an explanation but a set of mathematical functions that may or may not accurately describe statistical patterns in the data. The missing link in ML is the explanation or interpretation of the world that is supposedly represented by the data (Anderson 2008; Hofman, Sharma, and Watts 2017; Pearl and Mackenzie 2018; Hildebrandt 2015, 37-57).

Machine experience is limited to digital data

To understand what this means, it is pivotal to note that the 'experience' of the machine consists of digital data (when referring to data I mean digital data). Machines do not experience light, temperature, sound or touch, nor do they have any understanding of natural language; they merely process data. Whereas such data may be a trace of, or a representation of light, temperature, sound, touch or text, it should not be confused with what it traces or represents. The choice of data, the kind of translation it implies, the type of error it may contain and the way it has been curated all impact the accomplishments of ML. For instance, developers may use 'low hanging fruit', i.e. data that is easily available but not necessarily relevant or complete. This

may result in bad ML applications (garbage in, garbage out or GIGO), and can be remedied either by obtaining other and/or more data, or by accepting that the data needed for the task cannot be harvested.

Before training their learning algorithm ('the learner') on the data, developers will attempt to remove irrelevant or incorrect 'noise', depending on the goal of the operation. They always run the risk of removing highly relevant data, even though the risk can be reduced by testing on differently curated data sets.

However, we must also remind ourselves that data-driven applications necessarily feed on the reduction of real world experience to sensor data or to natural language processing, translating the flux of a life world into variables that enable measurement and calculation. Such translation may lead to computational artefacts (bugs), taking note that any quantification requires prior qualification (as the same type of data). In the case of real-time interaction with data-driven systems this may lead to strange responses, such as taking a pedestrian for a plastic bag – resulting in the death of the pedestrian (Gibbs 2018).

Finally, bias in the real world may be reinforced due to the use of statistics, often resulting in what has been coined 'disparate accuracy', which may further entrench existing discrimination (Barocas and Selbst 2016).

The mathematical target function

A more fundamental point is that the goal of ML can be summarized as detecting relevant 'bias' in a dataset, where 'bias' refers to the patterned deviation from a random distribution (Mitchell 1997, 20-51). Unless a dataset has a random distribution – which is highly improbable – an algorithm that is trained to detect 'bias' will always come up with patterns. The more interesting point then is to figure out whether the bias is either spurious or relevant.

The detection of relevant 'bias' in a dataset can be defined as the approximation of a mathematical target function that best describes the relationship between input and output data. To enable such approximation a so-called hypothesis space is developed with sets of mathematical functions that may or may not succeed in describing this relationship. The better the function expresses this relationship, the higher the accuracy of the system.

Machine learning can thus also be defined as a type of compression. Instead of a huge set of data, we now have a mathematical function that describes the data, noting that the same data can be compressed in different ways, depending on the task and/or the performance metric. As should be clear, the shaping of the hypotheses space is critical for a proper description of the data; a well-developed space is hoped to generate a hypothesis that does well if tested on new data.

A core problem is that a detailed hypothesis space may do very well on the training set, but very bad on out-of-sample test data, as it 'overfits' with the training data in

a way that weakens its ability to generalize to new data. A less detailed hypothesis space, however, may generate a function that does well in generalizing to new data, but 'overgeneralizes' in a way that results in overlooking crucial connections, thus missing relevant features. If the environment is static and translates well into data, these problems can be resolved by iterant experimentation. If the environment is dynamic such iteration may not work.

Especially where human agents and societal institutions respond to their behavioural data traces being tested, machine learning algorithms face a double feedback loop as the anticipation of human and societal agents may invalidate the findings of 'the learner'. That is why a game with fixed and closed rules such as Go can be learnt based on the brute force (computing power) of programs such as AlphaZero (Collados 2017), whereas the adaptive nature of complex social phenomena remains elusive even when a system is trained on unprecedented volumes of data. This means that the fundamental assumption that underlies any ML system, i.e. that reality is governed by mathematical functions, does not necessarily hold for human society.

P-hacking

Next to bias in the data and the hypotheses space, the outcome of an ML application may be biased due to cherry picking with regard to the performance metric (P). This metric determines the accuracy of the system, based on the task (T) the system aims to perform and the data (E) it trains on. As one can imagine, if some metric P1 achieves 67% accuracy, whereas another metric P2 achieves 98% accuracy, the temptation to use only P2 and boast high accuracy is formidable. I will call this P-hacking, as it seems to be the twin sister of p-hacking (Gollnick in this volume, Berman et al. 2018). Especially in systems that are difficult to interpret high accuracy does not mean much, as the system may be getting things wrong despite the accuracy. The opacity of the underlying causality (e.g. in the case of medical diagnosis) or reasoning (e.g. in the case of quantified legal prediction) easily hides potential misfits.

For instance, a system that was meant to predict death after pneumonia qualified chest pain, heart disease and asthma as indicators of low risk, contrary to reality (Caruana et al. 2015). Any doctor can tell you that these three indicators correlate with high risk. Nevertheless, the accuracy was very high – within the dataset on which the algorithm was trained. Because the indicators were visible it was easy to figure out what went wrong: patients with chest pain, heart disease or asthma are sent to a hospital and monitored so well that their risk is lowered due to the fact that they are treated as high risk. If, however, the rating had been based on a neural network it might have been far less obvious which of the features caused the system to attribute a low risk. This makes reliance on such systems dangerous, as it may take time (and unnecessary death) before the mistake is uncovered.

So what?

Based on the analysis of ML research design, I propose that whoever puts an ML application on the market should pre-register the research design that was used to develop the application (including subsequent updates). This will contribute to

the contestability of claims regarding the safety, security, and reliability of such applications, while also enabling the contestability of decisions based on such applications in terms of potential violations of fundamental rights such as privacy, data protection, freedom of expression, presumption of innocence and non-discrimination. If such preregistration were to become a requirement, e.g. in an updated Machinery Directive,[2] it would also be a good example of 'legal protection by design' (Hildebrandt 2015, 218).

Notes

[1] Peirce's fallibilism, as well as Popper's related theory of falsification demand that scientific theory is restricted to explanations that enable testing in a way that enables their refutation.

[2] DIRECTIVE 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery, and amending Directive 95/16/EC (recast).

References

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." Wired Magazine 16(7).

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." California Law Review 104: 671–732.

Berman, Ron, Leonid Pekelis, Aisling Scott, and Christophe Van den Bulte. 2018. 'P-Hacking and False Discovery in A/B Testing'. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3204791.

Brooks, Rodney. 2017. "Machine Learning Explained." MIT RETHINK. Robots, AI, and Other Stuff (blog). August 28, 2017. http://rodneybrooks.com/forai-machine-learning-explained/.

———. 2018. "My Dated Predictions – Rodney Brooks." MIT RETHINK (blog). January 1, 2018. https://rodneybrooks.com/my-dated-predictions/.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721–1730. KDD '15. New York, NY, USA: ACM. doi: 10.1145/2783258.2788613.

Collados, Jose Camacho. 2017. "Is AlphaZero Really a Scientific Breakthrough in AI?" Medium (blog). December 11, 2017. https://medium.com/@josecamachocollados/is-alphazero-really-a-scientific-breakthrough-in-ai-bf66ae1c84f2.

Gibbs, Samuel. 2018. "Uber's Self-Driving Car Saw the Pedestrian but Didn't Swerve – Report." The Guardian, May 8, 2018. http://www.theguardian.com/technology/2018/may/08/ubers-self-driving-car-saw-the-pedestrian-but-didnt-swerve-report.

Hildebrandt, Mireille. 2015. Smart Technologies and the End(s) of Law. Novel Entanglements of Law and Technology. Cheltenham: Edward Elgar.

Hofman, Jake M., Amit Sharma, and Duncan J. Watts. 2017. "Prediction and Explanation in Social Systems." Science 355 (6324): 486–88. doi: /10.1126/science.aal3856.

Marcus, Gary. 2018. "Deep Learning: A Critical Appraisal". arxiv.org/abs/1801.00631.

Mitchell, Thomas. 1997. Machine Learning. New York: McGraw-Hill Education.

Pearl, Judea, and Dana Mackenzie. 2018. The Book of Why: The New Science of Cause and Effect. New York: Basic Books.

Peirce, Charles S., and Patricia Ann Turrisi. 1997. Pragmatism as a Principle and Method of Right Thinking: The 1903 Harvard Lectures on Pragmatism. Albany: State University of New York Press.

INDUCTION IS NOT ROBUST TO SEARCH

CLARE ANN GOLLNICK

A data scientist's goal is one of translation: from data to knowledge to action. After defining a hypothesis but before making a decision, a critical step in the process is transforming data into evidence and assessing the quality of that evidence. Data does not speak for itself. The same observation in disparate contexts will support a disparate set of conclusions. Thus formalized inductive logic, including statistical inference and machine learning, require quantification of both the data and the context. Probabilities and probabilistic reasoning are used almost exclusively to quantify evidence as these frameworks combine observation with context into question-agnostic, cross-disciplinary metrics (1 in 100 chance, 95% confident, 99% accurate etc.).

Automated intelligence (AI) is based on search and selection

As data analysis evolved from a means to an end to a profession in and of itself, there have been substantial efforts to automate and scale the inductive process. Basic statistical inference requires a hypothesis, observed data, and a probabilistic measure of evidence. Automated intelligence additionally requires: 1) searching, testing many hypotheses or candidate models at the same time or in quick succession and 2) selection, choosing among the candidate hypotheses the 'best' one to be used in decision making.

Search and selection make it difficult to accurately represent context. The role of context is best understood with an example. A scientist believes that a coin is fair. She performs an experiment in which the coin is flipped repeatedly, recording the outcomes. The scientist observes eight 'tail' outcomes consecutively. The data (eight consecutive tails) is raw observation. The total number of times the coin was flipped is relevant context. If the coin was flipped a total of ten times, the series of eight consecutive tails is substantial evidence the coin is not fair. However, if the coin was flipped ten thousand times, one series of eight consecutive tails is still consistent with the hypothesis of a fair coin. If the number of total flips is unknown, it is difficult to make any statement with respect to the hypothesis.

Automated intelligence uses probabilistic reasoning and inductive logic outside the confines of controlled experiments or defined context. Multiple hypotheses and uncontrolled variables are tested simultaneously. With coin-flip data, for example, the goal may be to predict the outcome of the next coin flip without knowing details of the experiment. Potential predicative hypotheses may include: 1) the flipping process is biased; 2) the coin is changed mid-experiment to a new coin at random; 3) both the coin and flipping process are biased but with different degrees and direction; 4) the coin was made of chocolate and bias was influeced by room temperature. The inclusion of bizarre hypotheses is used to drive the point: with sufficient contextual uncertainty, a historical data set is likely to be consistent with multiple contradictory explanations. To make decisions, it is necessary to define a selection criterion by which to choose a 'best' hypothesis or model (the model most likely to generalize into knowledge). These selection criteria also take the form of probabilistic estimates of evidence, requiring context of their own.

Reproducibility in scientific literature

Experimental scientists, particularly in theory-poor and hypothesis-rich fields such as biology and psychology, have learned the hard way what happens when inductive logic is automated and scaled. Currently, most published, peer-reviewed studies in these fields describe a result that would not occur again if the experiment were repeated (Baker 2016; Ioannidis 2005). This problem has become known as the 'reproducibility crisis'. Reproducibility is a core tenet of the scientific method; excessive irreproducibility undermines the credibility and minimizes the impact of the output of scientific endeavours.

The reproducibility crisis is often attributed to the misuse of statistical hypothesis testing (Nuzzo 2014), but is better understood as an unavoidable outcome of scaling induction (search and selection of evidence). Briefly, a hypothesis test is an algorithm that calculates the probability that the differences between the experimental and control groups would have occurred if the experimental modulation had no effect (the null hypothesis). A common output metric is the 'p-value'. If a p-value is sufficiently small, the null hypothesis is rejected. The experimental result is considered statistically significant. Statistical significance has become the default selection criterion at which an experiment is published, thereby making its way into scientific literature.

The problem emerges as researchers try multiple similar experiments. A statistically significant result is unlikely to occur in one experiment, but is likely to occur eventually if an experiment is repeated. The scale required to produce a false-positive is smaller than one might imagine. A common scientific threshold of statistical significance is $p<0.05$ (less than 5% chance of occurring due to chance alone); using this threshold, the number of experiments needed to create where a scientist is more likely than not to observe a false-positive result is on the order of twenty experiments.[1] In practice, a scientist could see this false-positive result in an early iteration and perceive it as strong evidence of a true effect.

Scientists can exacerbate the problem by using search-based strategies within their own research process. This is known as p-hacking or data dredging. Scientists are incentivized to seek out unexpected anomalies and patterns. In fact, a scientific career is considered successful only if a scientist publishes statistically significant results regularly and repeatedly. There is external pressure to design a research process that maximizes the likelihood of finding a statistically significant result with minimal time or effort. For example, a researcher could design a method to screen hundreds of chemicals as drug candidates at the same time, increasing the likelihood that one or a few will have a statistically significant result. A researcher could collect many covariates and test every combination of covariates for combinatorial effects on an experimental outcome, performing thousands of hypothesis tests either explicitly or implicitly. Counter-intuitively, while a scientist is hired to run experiments, the more productive (by number of experiments) a researcher is when attempting to support a given hypothesis, the less reliable the evidence generated by any one experiment. At the extremes, data dredging methods can be used to support nearly any conclusion: including arguing that the mind of a dead salmon can be read using fMRI data

(Bennett, Wolford, and Miller 2009) or that people can age in reverse become younger by listening to a Beatles' song (Simmons, Nelson, and Simonsohn 2011).

The reproducibility crisis is a tangible demonstration of the limits of induction. In fact, the degradation of the quality of scientific literature was predictable from an understanding of statistical inference and scale alone. In 2005, Dr. John Ioannidis of Stanford University demonstrated using a Bayesian framework that scientific literature would become more unreliable over time as many scientists repeatedly tested similar, but ultimately incorrect, hypotheses (searching) and only reported significant results (selection) (Ioannidis 2005). Importantly, an individual scientist may not be aware that the same experiment was performed in the past, is currently being performed in other laboratories, or that an analogous experiment was performed using other methods. Yet this invisible context is critical to accurately assess the quality of the probabilistic evidence provided by their study (Ioannidis 2005; Nuzzo 2014; Simmons, Nelson, and Simonsohn 2011). As such, the reproducibility crisis is a problem observed not by individual scientists or within a single study, but upon examination of the complete body of scientific literature or from the perspective of a population [2]. Unsurprisingly, pharmaceutical companies that depend on academic research to identify drug candidates were one of the first to quantify the magnitude of the reproducibility crisis in biology. Bayer reported less than 30% of attempts to reproduce findings resulted in successful replication (Prinz, Schlange, and Asadullah 2011). Amgen reported less than 11% (Begley and Ellis 2012).

Saving machine learning from p-hacking

Machine learning is not foundationally different from hypothesis testing. While p-values have been replaced with other probabilistic metrics of evidence, most machine learning models are still well approximated by the model of 'many hypothesis tests performed simultaneously'. The training phases of machine learning are highly iterative, relying on the same methods of statistical inference used in all types of induction (search). A hypothesis (candidate model) is generated, tested, updated and tested again until some stopping condition is met (selected). The model that most closely aligns with a previously chosen metric of success is selected, much in the same way that a particularly successful experiment is selected for publication based on the success of the experiment. The practice of running many iterative analyses on the same data and choosing the model that performs 'best' is indistinguishable from p-hacking, except that most of the steps are performed by a computer. In fact, due to automation, it occurs faster and more obviously than in traditional science experimentation. Rather than using the term p-hacking, the machine learning terminology is 'overfitting'. Overfitting is evidence of having too many degrees of freedom, considering through too many potential hypotheses (searching) to find (select) a model that appears to perform well.

Much of the manual work that goes into generating a machine learning algorithm is focused on mitigating the damage done by scaling induction. Cross validation (separating into training and testing) mimics the 'one hypothesis' to 'one experiment' standards of the ideal scientific method. Regularisation (penalising complex explana-

tions) is meant to limit the number of models an algorithm implicitly considers, thereby reducing the amount of data dredging. Yet, just like well-meaning scientists seeking statistically significant results for their research projects, data scientists may break these protections using excessive search and selection in their own workflow. Repeated training, testing, training, and testing will create models that appear to work (perform better than chance), but are fitting noise. Much like publication of scientific experiments, the incentive structure around an individual data scientist's performance is often not aligned with success of a data science initiative overall.

If not already there, automated intelligence and machine learning will develop a reproducibility crisis of its own. Early research wins and models announced with much publicity will not generalize and eventually fail. Businesses will perceive their data science teams as underperforming, or not worth the investment. Practitioners and strategic leaders would benefit from understanding the limits of inference. Models built based on strong theoretical foundations (existing knowledge, context), based on rules that have already shown substantial predictive value, will outperform models developed largely by inference, based on excessive search and selection.

Notes

[1] Author acknowledges an over-simplification. The exact number depends on number experimental variables such as variability within the population but often falls on this order of magnitude.

[2] An often-proposed solution to the reproducibility crisis is to publish all experiments regardless of outcome (negative or positive). This proposal solves a problem of selection, but also changes the nature and intent of scientific literature. Scientific literature would no longer represent a body of knowledge, but a public record of experiments. As such, it only pushes the problem of scaling induction to a later stage of the scientific inference process.

References

Baker, Monya. 2016. "1,500 Scientists Lift the Lid on Reproducibility." Nature 533 (7604): 452–454. doi:10.1038/533452a.

Begley, C. Glenn, and Lee M. Ellis. 2012. "Raise Standards for Preclinical Cancer Research." Nature 483 (7391): 531–33. doi:10.1038/483531a.

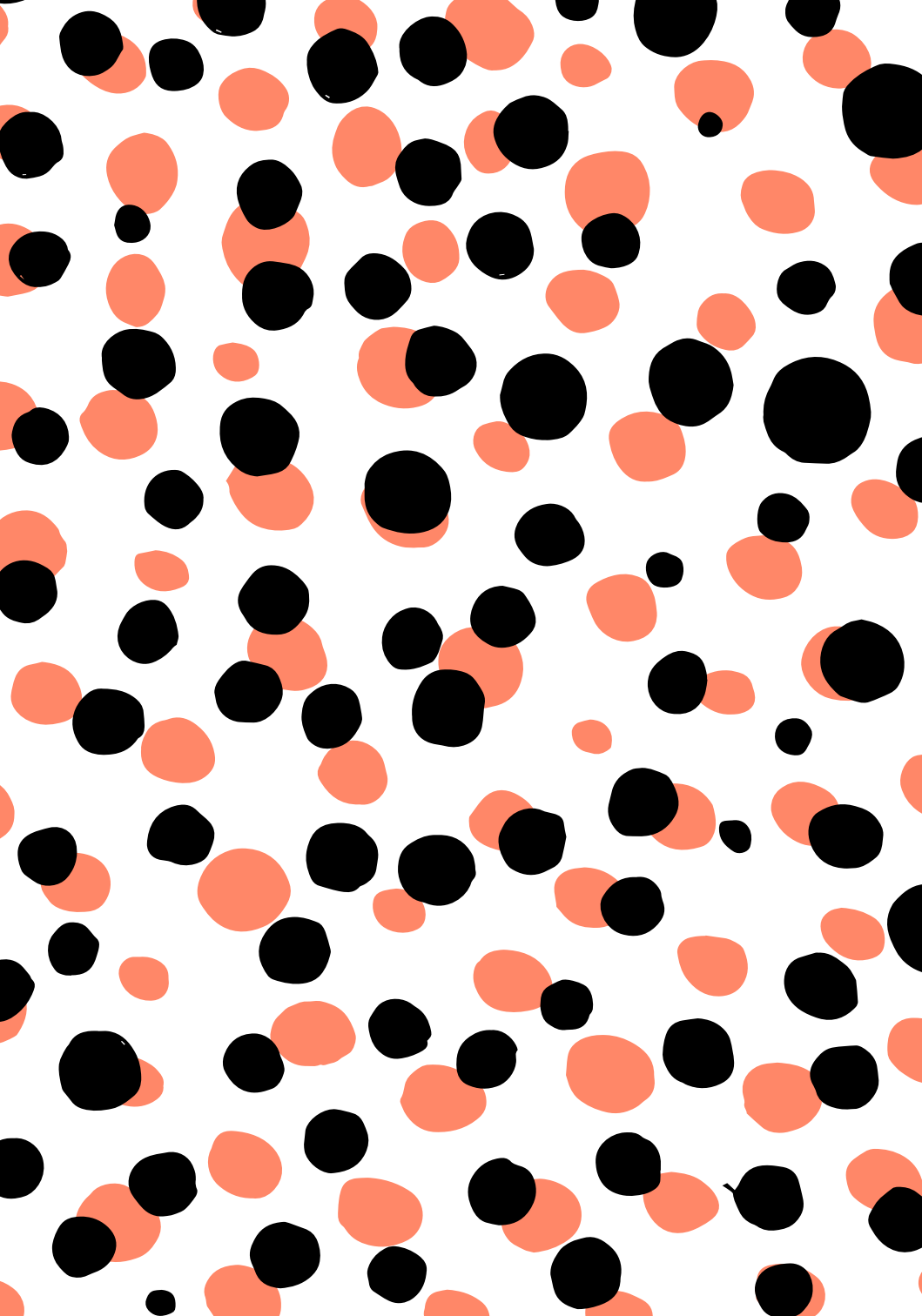
Bennett, Craig M., George L. Wolford, and Michael B. Miller. 2009. "The Principled Control of False Positives in Neuroimaging." Social Cognitive and Affective Neuroscience 4 (4): 417–22. doi:10.1093/scan/nsp053.

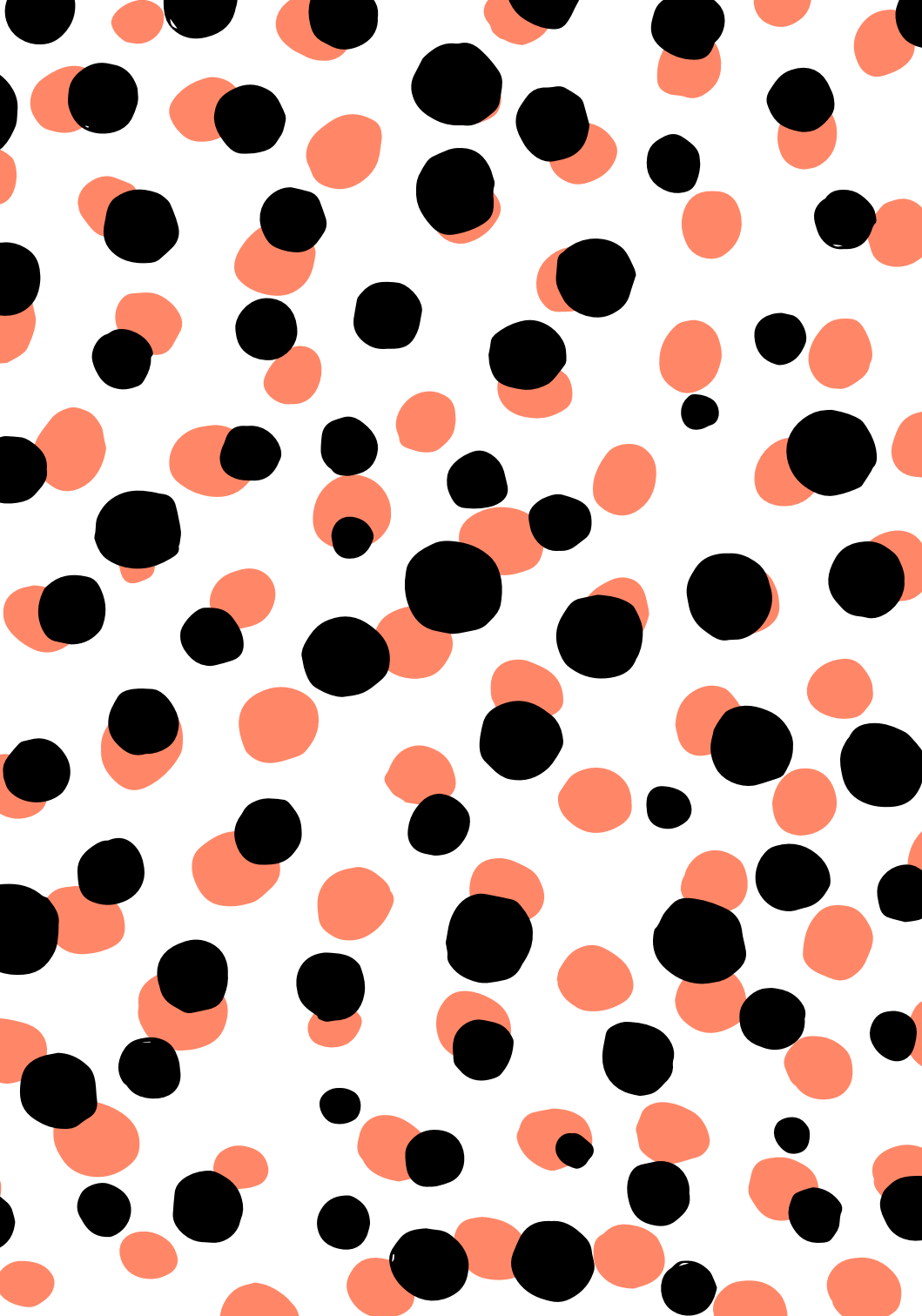
Ioannidis, John P A. 2005. "Why Most Published Research Findings Are False." PLoS Medicine 2 (8): e124. doi:10.1371/journal.pmed.0020124.

Nuzzo, Regina. 2014. "Scientific Method: Statistical Errors." Nature 506 (7487): 150–52. doi:10.1038/506150a.

Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. "Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?" Nature Reviews. Drug Discovery 10 (9). Nature Publishing Group: 712. doi:10.1038/nrd3439-c1.

Simmons, Joseph P., Leif D Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." Psychological Science 22 (11): 1359–66. doi:10.1177/0956797611417632.





In the data economy, many companies try to gain a competitive edge by extracting profiles and other hidden knowledge from large amounts of data via data mining and machine learning (Custers and Bachlechner 2018). This can be seen as an input-output process, in which knowledge (i.e., profiles) is extracted from raw data (Fayyad et al. 1996). Many kinds of data are used as 'ingredients', but often the analysis (i.e., 'the recipe') is supposed to remain confidential, as it may constitute the core business secrets of companies. Also the outcomes, i.e., the resulting profiles or extracted knowledge (such as new target groups or risk assessments), are often supposed to remain confidential, as it may be valuable commercial information for companies.

Profiles extracted from large datasets are often regarded as useful knowledge for subsequent decision-making and micro-targeting (Hildebrandt and Gutwirth 2008; Zarsky 2003). In this provocation, a different perspective is taken, in which profiles are not regarded as knowledge, but rather as (new) data, namely as inferred data. Using this perspective, it is shown that profiles are not only an end result or an end product, but can also be reused as ingredients for further data analytics. In this way, profiling processes may function as amplifiers, amplifying bias and inaccuracies via positive feedback loops, that further entrench consequences for data subjects.

Profiling: ascribing inferred data to people

Personal data is the basis of each process of profiling people, either as individuals or as groups (Custers 2013). After data is collected, it is analysed, usually in automated ways, using tools like data mining and machine learning. The data used for input is gathered in different ways: large volumes of data are generated by people themselves (e.g., via social media) as well as by technology, including sensors (e.g., cameras, microphones), trackers (e.g., RFID tags, web surfing behaviour) and other devices (e.g., mobile phones, wearables for self-surveillance/quantified self). In this way, profiles can be inferred from all kinds of data, including behavioural biometric data (Yannopoulos et al. 2008), location data (Fritsch 2008) or anonymised data (Schreurs et al. 2008).

Basic profiling techniques like regression, classification or clustering essentially ascribe attributes to people. This means that new attributes are inferred from available attributes, either from the same person or from other persons. These inferences may be precise (e.g., inferring age from the data of birth) or estimates (e.g., inferring intelligence or happiness from Facebook likes) (Kosinski et al. 2012). In this way, attributes a data subject does not want to disclose or attributes a data subject does not know can be predicted via data analytics and ascribed to that person. The key characteristic of inferred data is that it is data inferred from other data and not data directly or indirectly provided by data subjects.

Depending on factors like the total population, existing privacy laws and maturity of the data economy, it may differ from country to country in how many databases people are represented. In the EU it is reasonable to assume that people have their personal data in several hundreds or even several thousands of databases. Usually people are not aware of this and neither are they aware which data it concerns, for

which purposes the data are processed, and how any resulting profiles may lead to decisions about them (Eurobarometer 2015). Many of these data have not been obtained directly from the data subjects, but are data obtained via data sharing and data reuse, sometimes via so-called data brokers as intermediaries (Custers and Ursic 2016).

Reusing inferred data: positive feedback loops

The reuse of inferred data may have advantages. Inferring data can be a tool to fill gaps in incomplete datasets or check the correctness of available data by matching inferred data with the contested data. In this way, datasets enriched with many inferred attributes are likely to have higher levels of completeness and accuracy. In big data analytics, completeness and correctness of data is not a strict condition, but obviously may contribute to getting more accurate and reliable results.

At the same time, reusing inferred data as input for data analytics, particularly profiling processes, may turn profiling processes into amplifiers with positive (i.e., self-reinforcing) feedback loops. Effects of small disturbances (like incorrect or incomplete data, or flaws in the data analysis) may lead to an increase of the magnitude of perturbations. Obviously, this may have some serious consequences for data subjects. Profiling based on datasets from data brokers that contain large amounts of inferred data, may propagate any existing biased patterns, leading to disparate impact (Barocas and Selbst 2016). For instance, for profiling insurance premiums, a dataset with income data (directly obtained from data subjects) is less valuable than a dataset further enriched by the data broker with credit scores (inferred data). However, the credit scores may already be based on the income data, which means the insurance premium profiles are influenced twice by the original income data: directly and indirectly via the inferred credit scores. The reuse of inferred data may thus lead to self-fulfilling prophecies – a phenomenon well-known in profiling (Custers 2013). In case of inferred data, however, the effect might be much stronger: because of the self-reinforcing effect, patterns may be amplified and become much more entrenched. These effects may amplify inequality, undermine democracy and further push people into categories that are hard to break out (O'Neil 2016).

Inferred data under the GDPR

The GDPR provides data subjects with an extensive number of data subject rights, like rights to information, access, erasure and more. With regard to profiling, most of these rights seem to focus on the input data. The term inferred data occurs nowhere in the text of the GDPR, which clearly focuses on (personal) data, not on knowledge. A few rights, such as the right to object to profiling under certain conditions (Art. 21) and the right not to be subjected to automated individual decision-making (Art. 22) relate to the profiling process.

Data controllers should inform data subjects (upon request) about the existence of profiling processes and provide meaningful information about the logic involved and its consequences for the data subject (Art. 13.2f, 14.2g, and 15.1h). There is an extensive debate on how far this 'right to explanation' actually extends (Wachter et al.

2017; Veale and Edwards 2018; Selbst and Powles 2017; Kaminski 2018). However, few argue that there is an obligation for data controllers to disclose (1) the actual algorithms used, (2) the actual weighting of the data subject's data, and (3) data of other data subjects used in the profiling. Without such information, it is impossible for data subjects to check whether data is inferred correctly.

Companies may not be very keen to share algorithms and profiles as these can be considered trade secrets of vital interest, constituting their competitive edge. These companies may also suggest that profiles are corporate secrets because they may, via reverse engineering, enable disclosure of their analyses and software (Hildebrandt 2011, 23).

If inferred data is ascribed to groups or categories, it may not be personal data. However, in micro-targeting inferred data will often be ascribed to an identified or identifiable natural person, yielding personal data. If inferred data is personal data, there may still be practical issues with data subject rights. For instance, the right to rectification requires that data subjects show that the data are wrong. Proving that inferred data are wrong, is impossible for data subjects without access to analysis tools and the data of other data subjects used in the analysis. Obviously data subjects may object to the profiling altogether, but this may be too rigorous.

Data subjects may also consider transferring their data to other data controllers that provide more transparency on their profiling processes. This can be done via the right to data portability, prescribing that a data subject has the right to receive the personal data concerning him or her in a structured, commonly used and machine-readable format. However, this right does not include inferred data, as it is limited to only personal data which he or she has provided to a controller. This includes observed data, but not inferred or derived data (WP29 2016). In fact, a data controller may further limit the right to data portability by inferring data while deleting the original data on which the inferences are based, even if this is done in reversible ways (Madge 2017).

Wrap-up

Profiles are usually considered as knowledge extracted from data, but they can also be considered as (inferred) data that can be used as input for other profiling processes. Reuse of inferred data may contribute to improving the completeness and correctness of datasets. However, the reuse of inferred data may also turn profiling processes into amplifiers with positive (i.e., self-reinforcing) feedback loops. This may lead to propagation of existing biases in datasets and resulting patterns, amplifying inequalities and other issues related of profiling even stronger than in regular profiling practices. Looking at the GDPR, inferred data may or may not be personal data. If so, people have a right to access the inferred data and to receive meaningful information about the logic involved in the data analytics. However, since data subjects have no right to access the algorithms and data of other data subject used in the analyses, it is impossible for them to check whether data is inferred correctly.

References

Barocas, Solon, and Andrew Selbst. 2016. "Big Data's Disparate Impact" California Law Review 104(3): 671–732.

Custers, Bart. 2013. "Data Dilemmas in the Information Society." In Discrimination and Privacy in the Information Society, edited by Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky, 3-26. Heidelberg: Springer.

Custers, Bart, and Helena Ursic. 2016. "Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection." International Data Privacy Law 6(1): 4-15.

Custers, Bart, and Daniel Bachlechner. 2018. "Advancing the EU Data Economy: Conditions for Realizing the Full Potential of Data Reuse" Information Polity 22(4): 291-309.

Eurobarometer Survey 431. 2015. Attitudes on Data Protection and Electronic Identity in the European Union. Brussels, June 2015.

Fayyad, Usama, Gregory Piatetsky-Shapiro and Padhraic Smyth. 1996. "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM, 39(11): 27-34.

Hildebrandt, Mireille, and Serge Gutwirth, eds. 2008. Profiling the European Citizen: Cross-Disciplinary Perspectives. Dordrecht: Springer.

Hildebrandt, Mireille. 2011. "The Rule of Law in Cyberspace?" Inaugural Lecture, Nijmegen, Radboud University. https://works.bepress.com/mireille_hildebrandt/48/.

Kosinski, Michal, David Stillwell, and Thore Graepel. 2012. "Private Traits and Attributes are Predictable from Digital Records of Human Behaviour." Proceedings of the National Academy of Sciences USA 110: 5802–5.

Fritsch, Lothar. 2008. "Profiling and Location-Based Services." In Profiling the European Citizen, edited by Mireille Hildebrandt and Serge Gutwirth, 147-68. Dordrecht: Springer.

Kaminski, Margot. 2018. "The Right to Explanation, Explained." University of Colorado Legal Studies Research Paper No: 18-24.

Madge, Robert. 2017. "Five loopholes in the GDPR" My Data Journal, August 27, 2017. https://medium.com/mydata/five-loopholes-in-the-gdpr-367443c4248b.

O'Neil, Cathy. 2016. Weapons of Math Destruction; How big data increases inequality and threatens democracy. New York: Crown.

Schreurs, Wim, Mireille Hildebrandt, Els Kindt, and Michaël Vanfleteren. 2008. "*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, edited by Mireille Hildebrandt and Serge Gutwirth, 241-64. Dordrecht: Springer.

Selbst, Andrew and Julia Powles, 2017. "Meaningful Information and the Right to Explanation." International Data Privacy Law 7(4): 233-42.

Veale, Michael, and Lilian Edwards. 2018. "Clarity, Surprises, and Further Questions to in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling" Computer Law & Security Review, 34(2):398-404.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. "Why a Right to Explanation of Automated Decision-Making Does not Exist in the General Data Protection Regulation." International Data Privacy Law 7(2): 76-99.

WP29. 2016. Guidelines on the right to data portability, Article 29 Data Protection Working Party. 242. Brussels.

Yannopoulos, Angelos, Vassiliki Andronikou, and Theodora Varvarigou. 2008. "Behavioural Biometric Profiling and Ambient Intelligence." In Profiling the European Citizen: Cross-disciplinary Perspectives, edited by Mireille Hildebrandt and Serge Gutwirth, 89-110. Dordrecht: Springer.

Zarsky, Tal. 2003. "Mine Your Own Business! Making the Case for the Implications of the Data Mining of Personal Information in the Forum of Public Opinion." Yale Journal of Law and Technology 5(1): 1-57.

A PROSPECT OF THE FUTURE: HOW AUTONOMOUS SYSTEMS MAY QUALIFY AS LEGAL PERSONS

LISA JANSSENS

> *'SAM employs water kefir grains to produce a beverage, acting as a small scale automated food production system. This hybrid entity is both technological and organic, and strives to earn a living in the human world raising questions on ethics and machine rights.'*
>
> (Jense and Caye 2017)

The robot SAM is autonomous. However, it pays water and electricity bills, and it also employs and pays people. SAM has a bank account and, last but not least, SAM pays taxes. According to the artists, Arvid Jense and Marie Caye, this makes SAM into an independent economic entity. SAM comprises both organic and technological components. This is what makes it so difficult to classify SAM. The artwork questions the futuristic idea as to whether or not autonomous systems, when they are independent entities, can become susceptible to having rights and obligations.

The idea of legal personhood for robots, in a society that is increasingly interwoven with autonomous systems, is becoming ever more relevant. SAM's ability to act as a legal subject (e.g. to contract) depends on SAM qualifying as a legal person. In the future we may decide to attribute legal personhood to entities such as SAM. In the light of such developments, the European Parliament published a report with *'recommendations to the Commission on Civil Law Rules on Robotics'* (from now on titled: the report),[1] in which legal challenges surrounding autonomous systems are reviewed comprehensively. The report urges the European Commission to further explore whether or not the attribution of legal personhood to robots may be a possible solution to such challenges. This was stated as shown below:

> *(S.) Whereas the more autonomous robots are, the less they can be considered simple tools in the hands of other actors (such as the manufacturer, the owner, the user, etc.); whereas this, in turn, makes the ordinary rules on liability insufficient and calls for new rules which focus on how a machine can be held partly or entirely – responsible for its acts or omissions; whereas, as a consequence, it becomes more and more urgent to address the fundamental question of whether robots should possess a legal status.*

Bryson et al. (2017) argue that the case for electronic (legal) personhood is weak and that its application will also present us with certain issues. They advise us to take caution and to reflect on the problems, such as corruption, that have arisen in the past with the arrival of novel legal persons. Other examples of novel legal persons include entities that are accountable but unfunded, or fully financed but unaccountable. According to Bryson et al (2017) these examples illustrate the weakening of the legal protection for humans versus artificial persons.

In this provocation, the concept of legal personhood is explored as a possible solution to the challenging problems of a future in which autonomous systems interact more and more with the world. For example, when mistakes or failures occur during these interactions, the question arises who is liable. I will illustrate the complexity of this question by investigating the deadly accident with one of Uber's self-driving cars. By exploring this case, I will explore the question as to whether or not the attribution of legal personhood to autonomous systems could be one of the conceptual legal frameworks in which responsible innovation, with application of artificial intelligence, is made possible.

Autonomous systems and meaningful control

Autonomous systems relate to the research field of artificial intelligence; one of the primary goals in this field is to replicate human intelligence in machines. The hope to quickly match human intelligence to its fullest extent disappeared once it became clear how big and complex this ambition turned out to be. Successes and breakthroughs in the research field of artificial intelligence occurred only gradually and at a slow pace (Brooks 1991). Two questions arose with regards to how applications of artificial intelligence influence the human-machine interaction: how does autonomy of systems relate to human autonomy? And, to what extent does human autonomy change because systems become autonomous?

Autonomy concerns the attribution of meaningful control. Meaningful control relates to power and insight. Without this, there cannot be a form of meaningful control over how to carry out an operation or action. When we look at the implementation of meaningful control in autonomous systems, we could interpret that Artificial Intelligence/Machine Learning models in systems can be controlled, and monitored, when they are transparent. In order to create the possibility of meaningful control, transparency refers not only to the makers – the insiders – it explicitly refers to others who can check and understand the models as well. Pasquale (2016, 191) states:

> *Black boxes embody a paradox of the so-called information age: Data is becoming staggering in its breadth and depth, yet often the information most important to us is out of our reach, available only to insiders. Thus the novelists' preoccupation: What kind of society does this create?*

After all, transparency in *optima forma* concerns the entire AI/ML model. Which datasets were used? Which performance metrics were applied? How has the data been labelled, and which algorithms have been selected – and why (Hofman, et al. 2017)? Taking these questions into account, it is hard to fathom that parties would hide behind the so-called black-box algorithms. Ultimately, how else can we know whether the value of analyses through these models should represent a value in reality?

Who is liable for accidents?

Around 9:58 a.m. on Sunday, March 18, 2018, an Uber test car with software from

Volvo hit a 49-year old woman on the northbound Mill Avenue, Arizona. The woman did not survive the accident. The lethal accident caused by the self-driving test car was – in all probability – caused by a software error, according to the National Transportation Safety Board:

> *(…) the self-driving system software classified the pedestrian as an unknown object, as a vehicle, and then as a bicycle with varying expectations of future travel path. At 1.3 seconds before impact, the self-driving system determined that an emergency braking maneuver was needed to mitigate a collision.*

What is meaningful control? Does it entail the control of people over systems? Or does it mean that autonomous systems themselves have restricted autonomy, with a strict margin and assessment framework in which they are allowed to evaluate, judge and act? In this accident, various parties were involved; therefore, the question of who, or what, was in control concerns different actors. Is it Volvo that delivered the emergency brake software that did not work correctly? Is it Uber who has purchased this software? Or is it the operator who monitored the test-car, as he was looking at a monitoring screen instead of the road just before the accident? Perhaps it is the autonomous (test) car itself?

These questions address the matter of liability: who is responsible for how the architecture of the software works? And who can, for example, be held responsible for meaningful control over the systems? Failures and errors are in the news on a regular basis (Burkitt 2018; Feed 2018; Holley 2018; Marshall 2018). They are mainly caused by the enormous amount of complex traffic situations that must be assessed by the software. Cruise Automation from General Motors had problems with the blocks for roadworks, and even more worrying were GM's prototypes that tried to change lanes to the opposite side of the road. Google's Waymo was involved in an accident in Arizona, where the car tried to drive into streets that were too narrow. At Telenav, the prototype confused a roundabout for a stationary vehicle. Nissan faced a shutdown of the entire autonomous system. The more autonomous systems function, the more they will make their own assessment frameworks and rules, and the more complicated it becomes to address the responsibility question.

After these accidents and the ensuing discussion about liability, several brands have stopped testing their autonomous vehicles on public roads. This implies that if the question of responsibility cannot be addressed sufficiently, it can, in turn, inhibit innovation. Meaningful control is connected to successful innovation. A successful innovation is, amongst other things, a responsible innovation for society. On the one hand, to stimulate responsible innovations companies that develop or use autonomous systems could be held accountable for the performance and transparency of the AI/ML models. Still, the question remains as to whether legal personhood for autonomous systems solves more problems than it initiates. On the other hand a possible solution can be trace back the control on existing legal (corporate) entities.

SAM by Arvid Jense & Marie Caye

After all, citizens and consumers need to be able to rely on products and services developed through AI/ML models that result from responsible innovation. This, in turn, would mean that applications based on simulations with AI/ML must fit reality, and if they are not – and this results in accidents – the brands would be liable. To deviate from norms which entail responsible innovation, is not acceptable and could lead to liability.

Can an autonomous system become a legal person?

Will the attribution of legal personhood to autonomous systems provide a useful legal framework for solving liability issues? The system of law is flexible and as such has the possibility to create new entities in the existing system of law. The question of the European Parliament on this matter is stated in the report as below:

> *(T.) Whereas, ultimately, robots' autonomy raises the question of their nature in the light of the existing legal categories –of whether they should be regarded as natural persons, animals or objects –or whether a new category should be created, with its own specific features and implications as regards the attribution of rights and duties, including liability for damage;*

When an autonomous system is granted legal personhood, this creates a reality that can possibly give direction to the questions of meaningful control. Nevertheless, addressing legal personhood in the context of autonomous systems is complicated and leads to new challenges. Who identifies which systems qualify for legal personhood? How and under what conditions should this be done? What are the consequences when the system is disconnected and no longer exists?

Just as you can hold a company liable the European lawmaker could also create a reality in which an autonomous system can be liable. We need to be cautious and to reflect on the problems, such as abuse: it may be useful for the brands who are using autonomous systems to declare the system liable and walk away, without paying damages.

When we apply the attribution of legal personhood to the previously discussed Uber case, the above questions become concrete. Who will represent the autonomous system in court? Is it Uber or Volvo? If the self-driving car is a legal person, with—in this case—representatives of Uber and Volvo, the challenging question is: does this mean a distributed control and liability that stretches out over these actors? Tackling this problem, both practically and legally, is crucial to bridge the gap between human control over systems and the increasing autonomy of those systems.

Notes

[1] Committee on Legal Affairs, European Parliament, Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)), 27 January 2017.

References

Brooks, Rodney Allen. 1991. Intelligence without representation, Artificial Intelligence 47: 139-159.

Joanna J. Bryson, Mihailis E. Diamantis and Thomas D. Grant, 2017. "Of, for, and by the people: the legal lacuna of synthetic persons." Artificial Intelligence and Law, (25) 3: 273-291.

Burkitt, Bree. 2018. Waymo self-driving vehicle involved in Arizona Crash, USA Today, accessed online September 2018.

Feed, Benjamin. 2018. New report reveal self-driving cars struggling with software, highway overpasses, and even the sun, Statescoop, accessed online September 2018.

Hofman, Jake M., Amit Sharma and Duncan J. Watts. 2017. Prediction and explanation in social systems, Science 355: 486-488. https://doi.org/10.1126/science.aal3856.

Holley, Peter. 2018. After crash, injured motorcyclist accuses robot-driven vehicle of 'negligent driving', The Washington Post, accessed online September 2018.

Jense, Arvid, and Marie Caye. 2017. Thesis: free the means of production, accessed online September 2018 arvidandmarie.com.

Marshall, Aarian and Alex Davies. 2018. Uber's Self-Driving Car Saw The Woman It Killed, Report Says, Transportation, accessed online September 2018.

National Transportation Safety Board, 2018. Preliminary report highway, Accident ID: HWY18MH010.

Pasquale, Frank. 2016. Black Box Society. Cambridge, Massachusetts, London: Harvard University Press.

PROFILES OF PERSONHOOD: ON MULTIPLICITY OF REPRESENTING SUBJECTS

NIELS VAN DIJK

The European Parliament recently recommended 'electronic personhood' as a special legal status for robots to directly attribute them liability for caused damage, moving this idea from science fiction to legislative possibility. This 'provocation' will use this proposal to reflect upon the notion of personhood, not to analyse its singular nature, but to study persons as a *multiplicity of doubles* for individuals according to various modalities: dramatic, legal, political, statistical, digital. This turns this text into a gallery of masks, or a 'Hall of Faces' as presented in the TV series *Game of Thrones*. We will draw up several 'profiles of personhood' to explore the diverse ways this concept has been given conceptual meaning and visual sense. This juxtaposition is not meant to recognize patterns of similarity, but to put them in contrast and see how their attributes and functions differ.

Persona: A mask on stage

The etymology of the term 'person' goes back to the Latin *persona*. It refers to the mask that actors used to wear in Roman theatrical plays and which visually indicated which roles they were assuming. The mask allows one individual to impersonate another individual, to play their character and to speak and act in their name. This theatrical technique makes it possible to detach the human subject from the person. It was also used as a metaphor for other phenomena. Cicero used persona to understand the idea of representation both in a political sense when a magistrate acts in the name of the public community, and in a legal sense when the lawyer speaks for a client (Cicero 1967).

[W]hat can be so unreal as poetry, the theatre or stage-plays? And yet, … I myself have often been a spectator when the actor-an's eyes seemed to me to be blazing behind his mask (Cicero 1967, 337).

Roman Masks, Comic and Tragic. Author of Image unknown, Source: Parton, James. Caricature and other Comic Art. New York: Harper. 1877.

Erbore African Man. Image by YellowMonster, Source: https://pixabay.com; adaptation by Victor Bornia.

The subject is double: … to the extent that a subject is invested [by the law] with a function he is called 'person' (Thomas 1998).

The juristic person shares this theatrical meaning as a legal mask. It sets up a double for an individual, distinguishing it from the human being of flesh and blood. These two levels have often been confused by taking this juristic person in a symbolic sense, imbued with essential attributes (will, consciousness, life). In law however, 'personification' is often used to abstract from physical details, or even to introduce presumptions against the natural order (denaturalization). The *persona* has a 'fictive' existence in law. It is a legal artifact that institutes a '*point of imputation*' for legal relations, a foothold within the legal system for attributing certain rights and obligations (Thomas 1998). This pointillist mask, not unlike African or Balinese variants, hereby allows an entity to become an actor in legal processes and perform legal actions. In law, this relation between individual and person is divisible. The same individual can assume *personae* of several people (e.g. as their agent), whereas several different individuals

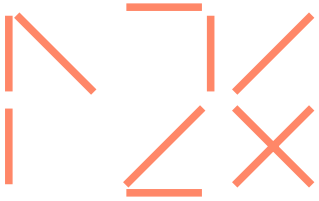
can assume one single persona (*e.g.* for a heritage). This mutual divisibility of the juristic person hinges on the type of legal relations implied, which can vary in kind and intensity. Furthermore, various non-human entities have also been granted this legal status and personhood for robots fits this line. Non-human entities can however not claim rights in their own name. They have to be represented, often by a lawyer.

Public persons. Unifying a multitude

Personification also became applied to publics, most famously in Hobbes' Leviathan. The public does not pre-exist as a coherent community. The multitude of people is only *unified* into one person through the mechanism of the social contract. The sovereign bears this public person and is authorized to speak in the name of the people and become their representative. This personification of the state is also clearly represented in the famous frontispiece to Hobbes' Leviathan. This is a *composite picture* depicting a multitude of single individuals that become unified in the main character, carrying the sword of supreme power. It depicts the unification of the composite body politic in a single sovereign person.

Average persons. Statistical realities

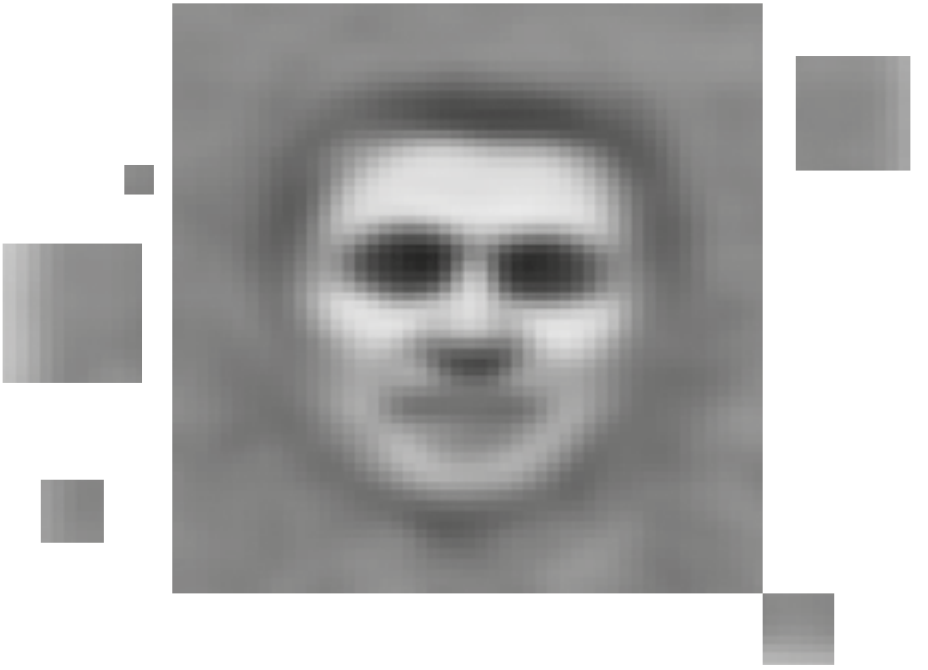
In the 18th century, there is an evolution away from a governmental regime focused on Hobbesian legal sovereignty. Through the rise of statistics in State administration, the *population* appeared as 'a new subject', with its own regularities and problems (Foucault 1994). The application of statistics to citizen behavior spurred a quest for 'social laws' governing people. Quetelet observed that large quantities of data about human attributes had certain distributions that allowed calculating a 'mean' and its deviation (Quetelet 1842). He here introduced the term 'average man' not as the quality of a real person, but as the real quality of a certain population. Galton strengthened this development by observing that many of these human traits were mutually correlated. This work was closely linked to his anthropometrical studies to identify certain types of humans from outer appearance. He invented the technique of *composite photography*, superimposing successive images of different individuals on the same photographic plate to generate a single portrait. When these images were taken from a certain 'class' of people, they formed a certain 'type' of person, e.g. a criminal and healthy type, and showed its common physical traits. This provided a visual instantiation of average persons as statistical realities of populational classes. The goal of this new statistical expertise was not only to obtain knowledge, but to devise policies to improve populational development towards desirable types and away from undesirable ones (Galton 1907).



Frontispiece to Leviathan. Image by Abraham Bosse, Source: (Hobbes 1998).
A multitude of men, are made *one* person, when they are by one man, or one person, represented; so that it be done with the consent of every one (Hobbes 1998, 98).

One may ask if there exists, in a people, *un homme type*, a man who represents this people by height, and in relation to which all the other men of the same nation must be considered as offering a deviation (Quetelet 1845, 258).
Specimens of Composite Portraiture [fragment]. Image by Francis Galton, Source: (Galton 1907).

Digital persons. Dividual data portraits



The optimal stimulus according to numerical constraint optimization. Machine-generated image, Source: (Le et al. 2012).

The 20th century saw the rise of artificial intelligence, machine learning and data mining, which share methodology with statistics. Self-learning algorithms can reiteratively search for patterns in data sets until arriving at optimal 'clusters' with their own mean or 'centroid'. When applied to people, the resulting correlations between data can be used to represent a human subject as a member of an existing *community*, or of a new *virtual grouping* of people. One field of application is image recognition, where Google researchers built an algorithm that learnt the pattern of the average human face (Le et al. 2012). This ghostly facial archetype shows a remarkable resemblance to Galton's 'composite portraits'. Such algorithmically inferred profiles form one kind of digital representation of an individual. Combined with data representations from other sources they constitute someone's 'digital *persona*', which renders a real-world subject identifiable. This digital portrait provides a fragmented representation of an individual based on distributed, partial data sets. Information technologies endlessly *divide* people in different data representations and reshuffle them to create 'recombinant identities'. Recombinations can happen in several ways based on criteria set by 'data controllers', often large ICT organizations. The digital person is here 'intended for use as a *proxy* for the individual' (Clarke 1996). This digital 'mask' allows the individual to be acted upon in the digital world, for specific purposes such as service provision.

Drawing contrasts

The potential entry of electronic persons in the Hall of faces sparked an exploration of various profiles of personhood. These profiles have been juxtaposed and can be put into contrast to 'find' differentiating patterns between salient attributes. First, whereas the public person and the average person both share their composite nature, they contrast in their *means of composition*. The public person of the Leviathan is composed through unification of a multitude through consent of each person in a social contract, by which the sovereign represents this assembled public community. The average person of statistics to the contrary, becomes assembled based on statistical grouping of entire populations, or certain communities and classes. Such communities 'were united by fate, not choice' (Gamboni 2005, 182), when ordered along a mean. The visualizations make this apparent. In Leviathan's composite image all the people composing the body of the public person remain individualized, their wills juxtaposed. In the composite portraits by Galton and Google, the separate individuals become superimposed and lose their individuality, only to merge in the new reality of an average human type. The digital person moves back to the level of the individual and is premised on division and recombination of data representations from multiple data sources. Divisibility hinges on criteria of someone's identifiability for service provision. This contrasts with the juristic person whose divisibility hinges on a legal entity or relation regarding a set of rights and obligations.

Secondly, there are significant differences regarding the actors bearing the masks (*representers*), and what they can do with these representations (*affordances*). In the production of average human types, statistical knowledge could be used to set out normative coordinates for new 'public goods' ('healthy' type) and 'public bads' ('criminal' type). This can form the basis for governmental policies aimed at controlling and improving the population and its relevant classes of people. In profiled human types, the clustering of people is even more virtualized, not necessarily given by pre-established criteria. The data controller can utilize resulting 'interested', 'interesting' and 'risky' types, for decisions on whether to grant a service. This digital mask is mainly operated by the data controller, not primarily on the (data) subject's behalf, but based rather on their organizational, often market-based interests. The juristic mask to the contrary, is worn during a legal process by a lawyer with the duty to legally represent the subject and act in their interest, with the goal of letting certain rights be imputed to them.

Lastly, we can focus on the *representative relation*. Quetelet and Galton conceived of the 'average man' and 'correlations' between human traits as statistical realities, i.e. real socio-biological qualities of populations that could be acted upon in policy-making. This contrasts with the juristic person as a double for the individual, to which social or biological qualities of humans should not be attributed. The juristic person can only produce its effects as denaturalizing device when human nature is kept at bay, and a fictive point is set up in legal space for attributing rights and duties. It is an empty legal form, the most anonymous of masks, which can be distributed to everyone in the multitude (or even to this multitude itself) precisely because it abstracts from traits that make each specific.

When we judge the entry of the electronic person in the hall of faces as a new type of legal mask, one should not to get carried away by symbolic discourses on artificial agency, fuelled by science fiction, speculative AI philosophy and overambitious promises by roboticists. Keeping symbolic and legal levels apart allows us to study the problem constellations around robotics and their economic and political dimensions, and conceive of juristic personhood as one possible technical solution among others. Attention should not be diverted from how a new type of person can upset relations between already existing persons, especially when it affects imputations of fundamental rights to people, or the equilibration of power relations in society.

References

Cicero, Marcus Tullius. 1967. De Oratore. Translated by E.W. Sutton. Cambridge, MA: Harvard University Press: 337.

Clarke, Roger. 1996. "The digital persona and its application to data surveillance." Information Society 10(2): 77-92.

Foucault, Michel. 1994. Dits et Écrits 1954-1988. Paris: Gallimard.

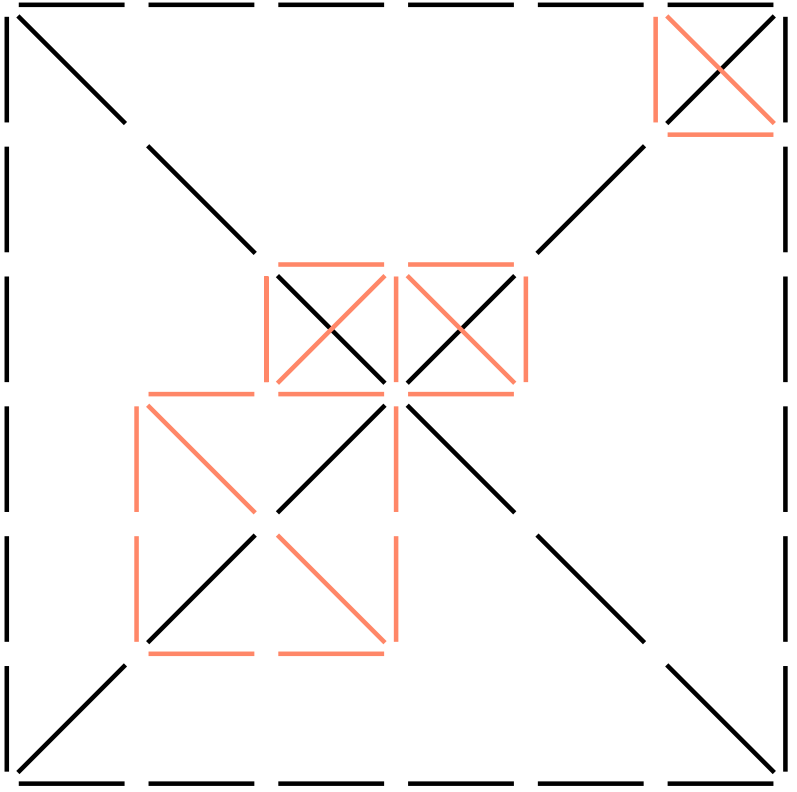
Galton, Francis. 1907. Inquiries into Human Faculty and its Developments. 2nd edition. J. M. Dent & Co.

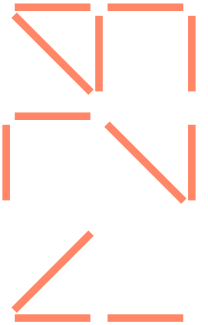
Gamboni, Dario. 2005. "Composing the Body Politic" In Making Things Public: Atmospheres of Democracy, edited by Bruno Latour and Peter Weibel, 162-96. Cambridge, MA: MIT Press.

Hobbes, Thomas, and John Charles Addison Gaskin.1998. Leviathan. Oxford: Oxford University Press: 109.

Le, Quoc V., Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, and Andrew Y. Ng. 2012. "Building high-level features using large scale unsupervised learning." Proceedings of ICML: 81–88.

Quetelet, Adolphe. 1845. "Sur l'appréciation des documents statistiques, et en particulier sur l'application des moyennes.", Bulletin de la Commission Centrale de la Statistique de Belgique: 98 (my translation).





IMAGINING DATA, BETWEEN LAPLACE'S DEMON AND THE RULE OF SUCCESSION

REUBEN BINNS

What is the probability that the sun will rise tomorrow? In 1814, Laplace posed this question and a means of answering it. By Laplace's reckoning, there were 1,826,251 recorded days in human history in which the sun had risen, and none in which it had not, giving odds of 0.9999994% that the sun will rise tomorrow. While Laplace's 'rule of succession' was a poor answer to Hume's problem of induction, it was part of a theory of probability and statistical inference which fleshed out the actuarial calculations of Bayes, and ultimately furnished the mathematical foundations of modern statistical inference and machine learning.

Perfect prediction

But Laplace is perhaps better known for a thought experiment which informed the classical definition of a deterministic universe (Laplace 1951, 4):

> *An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.*

Thus, with complete knowledge of the position and motion of every atom, and all the laws of cause and effect existing in nature, the future could be perfectly predicted like clockwork.

Laplace presents these two models of prediction; one, an impossible ideal, the other a pragmatic compromise. On the one hand, an imaginary intellect with complete knowledge (Laplace's 'demon' as it became known) encapsulates the fantasy of perfect prediction which may be metaphysically coherent but is epistemically forever out of reach. On the other, the rule of succession has a more modest aim; to put a precise figure on our inductive faith given limited observation and no background knowledge.

Modern computational systems of prediction, classification and inference are, for the most part, following the rule of succession. And this rule is expected to do more than ever; as data are generated from any corner of economic or social life, they are pressed towards the prediction or classification of some unknown, in the hope of reducing risk, increasing efficiency or exerting control. Machine learning is an exercise in fitting curves around these known data points in a multi-dimensional feature space, in such a way as to maximise the number of future data points falling on the right sides of the curves. Where Laplace plundered the historical record for observations of the sun's rising, modern data scientists mine the legacy databases of banks and welfare systems, or construct new ones out of the many digital traces we leave online.

Alongside widespread enthusiasm for data-driven decision making in the private and public sector, there are often strong concerns over its use to make consequential

decisions concerning people's lives. Such concerns about algorithmic decision-making have been articulated in terms of the threats to individual dignity, procedural justice, discrimination and fairness (see e.g. Hildebrandt and Gutwirth 2008). As a result of these concerns, data protection regulation affords various rights to subjects of data-driven-decisions; foremost, to not be subject to them, but also to have their logic meaningfully communicated, and to request a human reviewer.

One way of framing these concerns, and the philosophical motivation for such legal protections, is in terms of the damaging consequences of conflating Laplace's two models of prediction. Despite their non-causal, non-explanatory nature, the insights of machine learning are often presented and treated as if they approximate those of Laplace's omniscient demon. Patterns of geolocation, mouse movements, locations within social graphs and their statistical associations with loan repayments, retail purchases or employee productivity are taken as equivalent to the demon's knowledge of the position and forces of nature.

Laplace's sunrise problem is not a fruitful machine learning problem, but it is an extreme example which illustrates the limited nature of the probabilistic knowledge that statistical methods, following the rule of succession, can furnish us with. Relying purely on observation, it eschews theory. It does not pretend to know anything about the 'forces that set nature in motion'; it merely provides us with guidance on what to believe in the absence of such knowledge. We can explain our belief that the sun will rise not by reference to astronomical theory, but by subjective degrees of belief derived from numerical operations over observations.

Laplace readily acknowledged that estimations of likelihood are merely a set of consistent rules about how to act based on limited and subjective sets of evidence. And they cannot substitute for causal models; the rising of the sun, or human behaviours like repaying a loan, are fundamentally unlike the process of drawing coloured marbles from an urn or flipping a coin. As Ian Hacking charted, probability emerged in the 17th century as a new way of knowing, and statistical regularities became elevated to a status analogous to laws of nature (Hacking 2006). But patterns at population level are not explanations for any single individual's behaviour. From the perspective of Laplace's omniscient demon, no individual person is 60% likely to default on a loan or commit a crime; they either will or they won't, in the fullness of time, depending on the precise configurations of the position of matter and the operations of natural laws. But absent such omniscience, individual behaviours are indeterminate, and only predictable at the population level.

It is in this space of indeterminacy where we act in ways we personally identify with, and where we attribute both to ourselves and to others freedom, agency and intentionality. Regardless of how we understand the notion of free will at a metaphysical level, attributions of agency persist in the face of population-level statistical regularities. This is one reason why people may object to the use of statistical prediction to make decisions about individuals. Even if every other person with a given set of features acted in a certain way, the nth person sharing those same features might act otherwise.

This theme came to the surface in recent experimental work, where we probed people's perceptions of justice in response to a variety of hypothetical automated decisions, accompanied by a range of different explanations which aim to impart meaningful information about the system (Binns et al. 2018). One participant, reacting to a decision to deny an individual a financial loan on the basis of a machine learning model trained on data from prior borrowers, argued that: 'it's unfair to make the decision by just comparing him to other people and then looking at the statistics. He isn't the same person' (Binns et al. 2018, 7). This suggests that ML-driven decisions will always be on some level unfair, because at any point, someone might act counter to the trend. As such, we need human intervention to allow for discretion and the chance that people might act otherwise.

But there is another potential response, one more likely to be favoured by advocates of such systems; make the system better to catch the exceptions. This means finding new sources of data, building more complex models which encompass different sub-groups, or both. Any discretion that might be exercised in the case of a human reviewer treating an individual differently to the model's output, could perhaps be subsumed under the statistical model by adding more data. This strategy is compelling because it suggests that the demands of justice are ultimately in line with the goal of accuracy.

But to call only for more data is a problematic response to questions of justice. The data you might need to update the model in ways that would enable it to handle the exceptions generated by human discretion might never exist. Training data from the real world usually does not encompass the full range of possible values for a set of features. One cannot always draw samples from all logically possible populations for various societal, economic, or even biological reasons. For instance, there may not be data on the population of prisoners who were deemed 'high risk' but were released; or of those with low credit scores who were nevertheless given loans; or of pregnant males (except in rare circumstances). This is a practical problem for machine learning in any sphere, not only those in which human lives are at stake. Laplace could not experiment with the astronomical circumstances underlying the sun's motion to say 'why', beyond induction from the past, we should expect the sun to rise tomorrow, or explain the conditions under which it would not.

But non-existent data is not just a problem for machine learning. It is also, perhaps, something we need to imagine as a pre-condition for justice in decision-making. To understand why this might be so, consider recent work on 'de-biasing' machine learning (e.g. Pedreshi, Ruggieri and Turini 2008). The problem is that models may be trained on data which reflect unjust social biases, such that certain populations are more likely to be given a certain label. Both the variables used to predict an outcome, and those used to measure the outcome itself, might be biased. For instance, an educational qualification may be a decent proxy for a job applicant's knowledge, but if the awarding institutions have structural gender biases then a model for predicting applicant's future performance using such a proxy will be unfairly biased against women. Similarly, if work performance itself is measured by managerial reviews that

are also gender biased, then both the predictor variables and the outcome labels will be biased, potentially reinforcing those underlying discriminatory patterns.

Imagining data justice

While various different definitions of discrimination and fairness have been proposed for correcting such systems, taken to their logical conclusion, they ultimately require us to go beyond the data and imagine alternate states of affairs in which some discriminatory patterns do not exist. They might require us to determine what qualification the female applicant would have got in an unbiased institution, or what evaluation she would have got as an employee in a discrimination-free workplace. This requires causal models of discrimination and social injustice. But even causally understanding injustice may not be enough. We may also need to imagine what the just alternative might be, i.e. imagine the situation of the individual as they might be under a fundamentally different, non-patriarchal society. Defining fair decisions thus requires thinking about counterfactual causal scenarios in imaginary worlds (and perhaps even 'impossible' worlds, but hopefully not).

This leaves us somewhere orthogonal to Laplace's two extremes of minimal inference to subjective probabilities from incomplete data, and the 'single formula' of the all-knowing intellect. Justice is partly about the ability to imagine things that are not in fact the case; while we clearly fall short of the total predictive capacity of Laplace's demon, our human faculties of imagination give us access to an infinite variety of possible alternative worlds against which the actual world can be compared. And to imagine alternative possible worlds is as much a political act as it is an exercise in counterfactual causal reasoning.

References

Binns, Reuben, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions." Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM. Paper no. 377.

Hacking, Ian. 2006. "The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference." 2nd edition. London: Cambridge University Press.

Hildebrandt, Mireille, and Serge Gutwirth, eds. 2008. Profiling the European citizen: Cross-Disciplinary Perspectives. Dordrecht: Springer.

Laplace, Pierre Simon. 1951. "A Philosophical Essay on Probabilities." Translated by Frederick Wilson Truscott and Frederick Lincoln Emory. 6th edition. New York: Dover.

Pedreshi, Dino, Salvatore Ruggieri, and Franco Turini. 2008. "Discriminaton-aware data mining." Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 560–68.

PATRICK ALLO is a postdoctoral researcher at the Centre for Logic and Philosophy of Science at the Vrije Universiteit Brussel and a research associate of the Digital Ethics Lab at the Oxford Internet Institute, University of Oxford. From October 2015 to September 2017 he was a Marie Skłodowska-Curie Fellow at the University of Oxford.

ERO BALSA is a PhD candidate at COSIC in the Electrical Engineering department of KU Leuven, where he is a member of the privacy group. He works on privacy technologies, with a focus on obfuscation tools and data privacy. Personal website: http://homes.esat.kuleuven.be/~ebalsa/index.html.

IRINA BARALIUC As an affiliated researcher at the Vrije Universiteit Brussel - Law, Science, Technology and Society (LSTS), she focuses her research on the public and the private in the digital world in relation to the enjoyment of copyright protected works online, and on the development of the concept of intellectual privacy. Since 2016, she is the cultural programme coordinator at the Privacy Salon non-profit organization working on bringing privacy, data protection, surveillance, algorithmic awareness as well as other legal, ethical and societal issues raised by technologies into public discussion by means of artistic interventions, public discussions and workshops.

EMRE BAYAMLIOĞLU is a researcher at the Tilburg Institute for Law, Technology, and Society (TILT), the Netherlands. He is also an external fellow of the Research Group on Law Science Technology & Society (LSTS) at Vrije Universiteit Brussels. Before joining TILT in 2015, he has participated in the foundation of the Istanbul Bilgi University Information Technology Law Institute, and worked as a faculty member at Koç University Law School (2010-2015). His current research focuses on the transparency and contestability of automated decisions, and the possible legal impediments at the level of implementation. Personal website: https://www.tilburguniversity.edu/webwijs/show/emre.bayamlioglu/.

REUBEN BINNS is a researcher in Computer Science at the University of Oxford, and a research fellow in Artificial Intelligence at the UK Information Commissioner's Office. His research interests include technical, legal and ethical aspects of privacy, machine learning, and decentralised systems. He has a BA and MSc in Philosophy from the University of Cambridge, and a PhD from the Department of Electronics and Computer Science and the Faculty of Business and Law at the University of Southampton. His recent work has focused on two strands: large-scale measurement and human-computer interaction challenges of third-party tracking on the web, apps and IoT; and transparency, fairness and accountability in profiling and machine learning. Personal website: reubenbinns.com.

AUTHORS AND EDITORS

TOBIAS BLANKE

is a Professor of Social and Cultural Informatics in the Department of Digital Humanities and current Head of Department. Tobias' academic background is in philosophy and computer science. Before joining King's, he held various positions in the digital industries. His particular research expertise is big data from heterogeneous social and cultural collections as well as the social shaping of technologies. Tobias works on several international projects and committees and in particular led on the research and development work of the European Holocaust Research Infrastructure.

BART CUSTERS

PhD MSc LLM is associate professor and director of research at eLaw, the Center for Law and Digital Technologies at Leiden University. He has a background in both law and physics and is an expert in the area of law and digital technologies, including topics like profiling, big data, algorithmic decision-making, privacy, discrimination, cybercrime, technology in policing and artificial intelligence. As a researcher and project manager he has done research for the European Commission, NWO (the National Research Council in the Netherlands), the Dutch national government, local government agencies, large corporations and SMEs. Custers published three books on profiling, privacy, discrimination and big data, two books on the use of drones and one book on the use of bitcoins for money laundering cybercrime profits. On a regular basis he gives lectures on profiling, privacy and big data and related topics. He has presented his work at international conferences in the United States, Canada, China, Japan, Korea, Malaysia, Africa, the Middle East and throughout Europe. He has published his work, over a hundred publications, in scientific and professional journals and in newspapers. His list of publications and further information can be found at: https://www.universiteitleiden.nl/en/staffmembers/bart-custers#tab-1.

SYLVIE DELACROIX

focuses on the intersection between law and ethics, with a particular interest in Machine Ethics, Agency and the impact of habit on moral decisions (Habitual Ethics?, Bloomsbury, 2019). Her current research focuses on the design of both decision-support and `autonomous' systems meant for morally-loaded contexts. She also researches the effect of personalised profiling and ambient computing on ethical agency. Her work has notably been funded by the Wellcome Trust, the NHS and the Leverhulme Trust, from whom she received the Leverhulme Prize. Sylvie Delacroix has recently been appointed to the Public Policy Commission on the use of algorithms in the justice system (Law Society of England and Wales). She is also a Fellow of the Alan Turing Institute. Personal website: https://www.birmingham.ac.uk/staff/profiles/law/delacroix-sylvie.aspx.

NIELS VAN DIJK

is a lecturer in legal philosophy at the law faculties of the Vrije Universiteit (VUB) and the Saint-Louis University in Brussels, a post-doctoral researcher at the VUB Centre for Law Science Technology and Society, and director of the Brussels Laboratory for Privacy and Data Protection Impact Assessments (d.pia.lab). He has been a researcher in several national and European research projects on ICT technologies. His research focuses mainly on the challenges digital technologies pose to practices of law, especially in the fields of privacy, data protection and intellectual rights, including perspectives from legal theory, science and technology studies (STS) and ethnography of legal institutions. Niels van Dijk holds a PhD degree in law from the VUB, and LLM

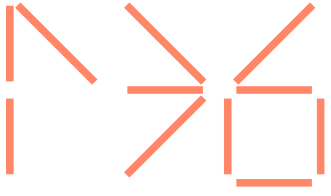
and MA degrees in law and philosophy from the University of Amsterdam. He has been a (visiting) researcher at the digital security department of Radboud University Nijmegen and the law department of the London School of Economics. Personal websites: http://www.vub.ac.be/LSTS/members/vandijk/ https://www.researchgate.net/profile/Niels_Dijk.

SABINE GLESS is professor at the Faculty of Law at the University of Basel (Switzerland) where she holds the Chair for Criminal Law and Criminal Procedure. Her research interests include international criminal law, evidence law and the impact of the digital revolution on criminal justice. Gless' work covers different issues of criminal liability and evidentiary problems connected to driving automation as well as to predictive policing and criminal trials. She is on the editorial board of several law journals and acts as an expert with the European Commission as well as with Swiss governmental authorities. She served as a member of the Swiss Research Council of the Swiss National Science Foundation and the Review Board of Legal Studies (Fachkollegium Rechtswissenschaften) of the German Research Association. Personal website: https://ius.unibas.ch/de/personen/sabine-gless/.

CLARE ANN GOLLNICK is a practicing data scientist, technologist and entrepreneur based in New York city. She specialized in intuitive modeling and predicative analytics. Previously, Clare was an experimental and computational neuroscientist. Her academic publications focus on information processing and perception within neural networks. Clare holds a PhD in biomedical engineering from Georgia Tech and a BS in bioengineering from the University of California, Berkeley.

GLORIA GONZÁLEZ FUSTER is a Research Professor at the Vrije Universiteit Brussel (VUB) Faculty of Law and Criminology. Member of the Law, Science, Technology and Society (LSTS) Research Group and of the Brussels Privacy Hub (BPH), she investigates legal issues related to privacy, personal data protection and security, and teaches 'Data Policies in the European Union' for the Data Law option of the Master of Laws in International and European Law (PILC) of VUB's Institute for European Studies (IES). Her academic background encompasses Law, Communication Sciences, and Modern Languages and Literature. Personal website: https://glgonzalezfuster.blog/.

SEDA GÜRSES is an FWO post-doctoral fellow at COSIC in the Electrical Engineering department of KU Leuven where she is a member of the privacy group. She works on privacy technologies, requirements engineering and optimization systems. Personal website: http://vous-etes-ici.net.



MIREILLE HILDEBRANDT

is a Research Professor at Vrije Universiteit Brussel (VUB), appointed by the VUB Research Council on the Chair of 'Interfacing Law and Technology'. She also holds the Chair of 'Smart Environments, Data Protection and the Rule of Law', at Radboud University, Nijmegen. In 2018 she was awarded an ERC Advanced Grant for a 5 year research project on 'Counting as a human being in the era of computational law', see www.cohubicol.com.

JAAP-HENK HOEPMAN

is associate professor at the Digital Security group of the Radboud University, Nijmegen, the Netherlands. He is also an associate professor in the IT Law section of the Transboundary Legal Studies department of the Faculty of Law of the University of Groningen. Moreover he is a principal scientist (and former scientific director and co-founder) of the Privacy & Identity Lab. He studies privacy by design and privacy friendly protocols for identity management and the Internet of Things. He speaks on these topics at national and international congresses and publishes papers in (inter) national journals. He also appears in the media as security and privacy expert, and writes about his research in the popular press. He is actively involved in the public debate concerning security and privacy in our society. In his free time he enjoys making composing music, designing graphics, and practising Okinawan Goju Ryu karate-do. Personal website: https://www.cs.ru.nl/~jhh.

LISA JANSSENS

is a scientist at the Netherlands Organisation for Applied Scientific Research (TNO) and researcher at LSTS, Vrije Universiteit Brussel (VUB) Faculty of Law and Criminology. She is working on analyses that examine Artificial Intelligence in the scope of Law, Philosophy and Cybersecurity. She obtained a Master's degree in Law at Leiden Law School (LLM) and a Master's degree in Philosophy at Leiden University (MA). She was assigned to the project on ethics and 'The Internet of Things' by the Dutch Cyber Security Council (National Coordinator for Security and Counterterrorism of the Dutch Ministry of Security and Justice). During her studies she was a trainee at SOLV lawyers and she was a trainee at the Dutch Scientific Council for Government Policy (WRR).

ORLA LYNSKEY

is an Associate Professor of Law at the LSE. She teaches and conducts research in the areas of data protection and digital rights, technology regulation, and EU law. Orla read law at Trinity College Dublin; the College of Europe, Bruges and the University of Cambridge. Her doctoral research and monograph (The Foundations of EU Data Protection Law, OUP 2015) focused on the dual dignitary and economic nature of personal data and the normative limits of individual control over such data. She is currently working on inter-related projects on the fundamental rights implications of 'data power' in digital markets and the privatization of personal data. Orla is a general editor of International Data Privacy Law and a case-note editor of the Modern Law Review. She is also a member of the EU Commission multi-stakeholder expert group on the GDPR and a member of several advisory boards on data protection matters.

REBEKAH OVERDORF is a post-doctoral fellow at COSIC in the Electrical Engineering department at KU Leuven. Her work focuses on the application of machine learning to privacy and studying the challenges that are specific to these privacy problems. Personal website: https://www.esat.kuleuven.be/cosic/people/rebekah-overdorf/.

LUCIA M. SOMMERER is a binational (German-American) legal scholar with focus on the intersection of criminal law and emerging technology. She is a Fellow at Yale Law School's Information Society Project, a Master of Laws (LL.M.) candidate at Yale Law School, and a doctoral candidate at the chair for Criminal Law and Criminology at Göttingen University, Germany. In her doctoral work she explores the use of predictive algorithms in the criminal justice system of Germany and the U.S. She has studied law at Munich and Oxford University. During her time in Munich she has focused inter alia on the legal regulation of climate engineering technologies such as carbon capture and storage. Further, she has taught sociology of law and criminal law as a teaching fellow at the law faculty of Göttingen University. Her past activities include research assistant with Hogan Lovells LLP and the Sino-German Institute for Legal Studies in Nanjing, China.

FELIX STALDER is a professor for Digital Culture at the Zurich University of the Arts, a senior researcher at the World Information Institute in Vienna and a moderator of <nettime>. His work focuses on the intersection of cultural, political and technological dynamics, in particular on new modes of commons-based production, control society, copyright and transformation of subjectivity. Among his recent publications are *Digital Solidarity* (PML & Mute 2014) and *The Digital Condition* Polity Press, 2018). Personal website: felix.openflows.com.

LINNET TAYLOR is Assistant Professor of Data Ethics, Law and Policy at the Tilburg Institute for Law, Technology, and Society (TILT). Her research focuses on global data justice – the development of a conceptual framework for the ethical and beneficial governance of data technologies based on insights from technology users and providers around the world.

ANTON VEDDER is a professor of IT Law at KU Leuven, board member of the Centre for IT and IP Law (CiTiP) in Leuven, and program director of the master's program of IP and ICT Law of KU Leuven in Brussels. In his research and teaching, he focuses on the interaction of technological developments, normative outlooks and regulation. Recent publications include books and articles on innovative technology, health care and enhancement, privacy, justice, fairness, and profiling, privacy and security, autonomous technology, quality of information and credibility of experts, legitimacy and the acceptance of innovative technology and technology regulation. Anton is an active member of international networks. He was involved in the EU research project BIOMED I, during the early 1990s. He was visiting researcher at Georgetown University Law Center, the Kennedy Institute of Ethics (Washington, DC) and the University of Maryland. He was also involved in several European FP7 and FP8 projects. Recently, he acquired national funding for research projects on ethical and legal aspects of e-coaching, e-health, and unmanned aerial vehicles (drones), and European funding (Horizon 2020) for

projects on, amongst others, public security and protection of critical infrastructures, and ehealth and enhancement. Personal webpage: https://www.law.kuleuven.be/citip/en/staff-members/staff/00097097.

ARJEN P. DE VRIES

is professor of information retrieval at Radboud University Nijmegen in the Netherlands. His research aims to resolve the question how users and systems may cooperate to improve information access, with a specific focus on the value of a combination of structured and unstructured information representations. Homepage: http://www.cs.ru.nl/~arjen/.

ADRIAN WELLER

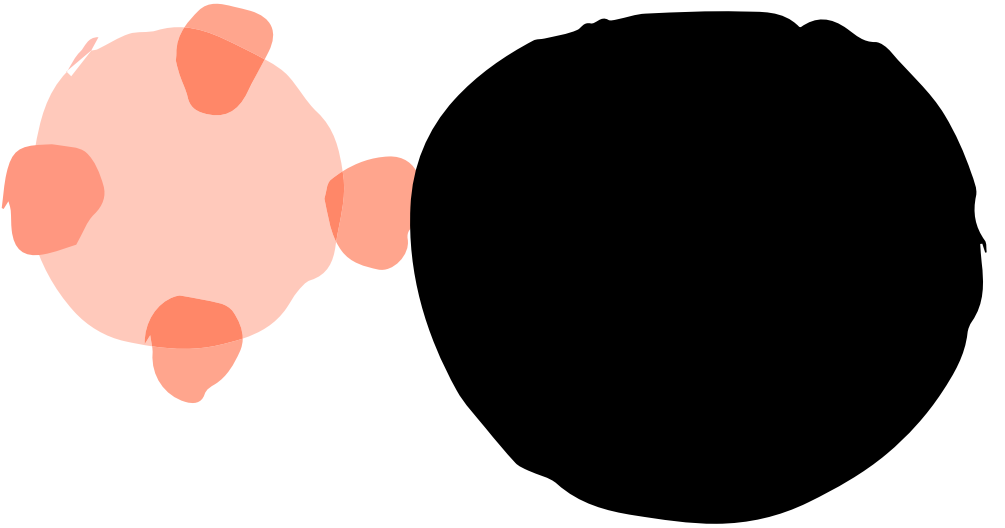
is Programme Director for AI at The Alan Turing Institute, the UK national institute for data science and AI. He is a Senior Research Fellow in Machine Learning at the University of Cambridge, and at the Leverhulme Centre for the Future of Intelligence where he leads a project on Trust and Transparency. He is very interested in all aspects of AI, its commercial applications and how it may be used to benefit society. He advises several companies and charities. Personal website: http://mlg.eng.cam.ac.uk/adrian/.

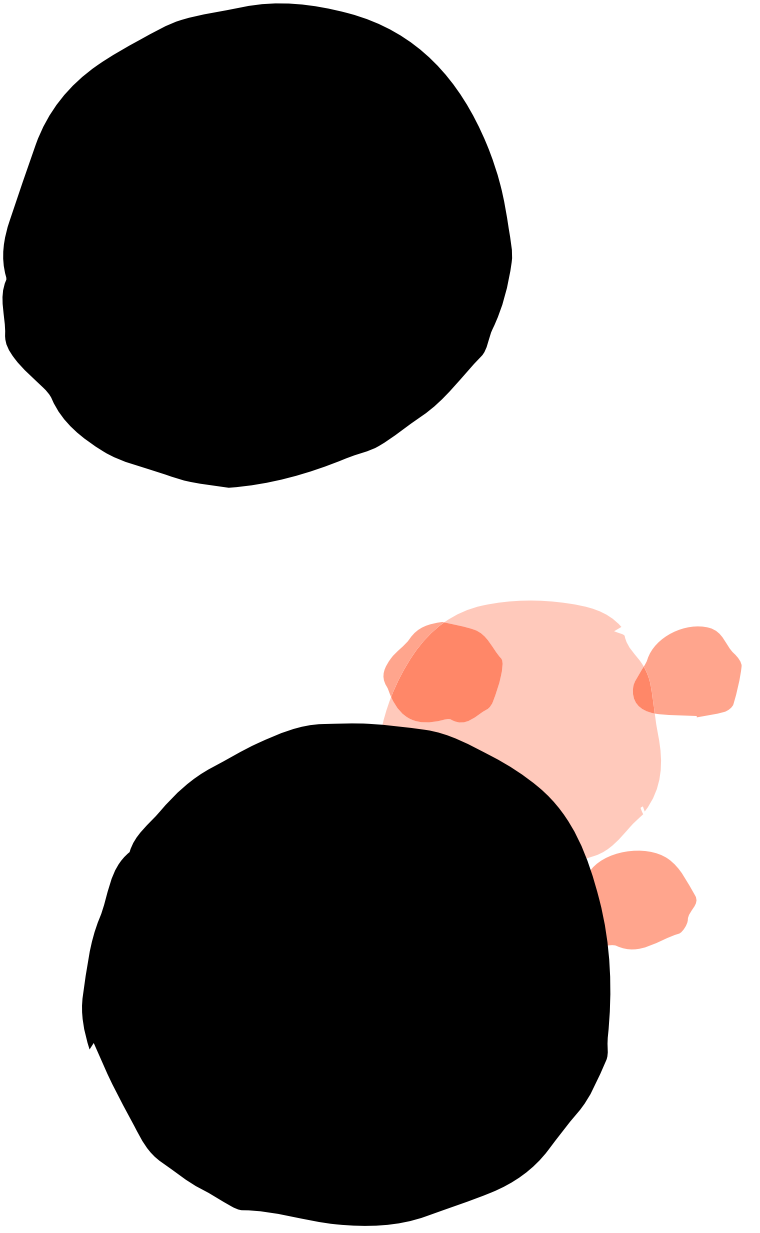
BEN WAGNER

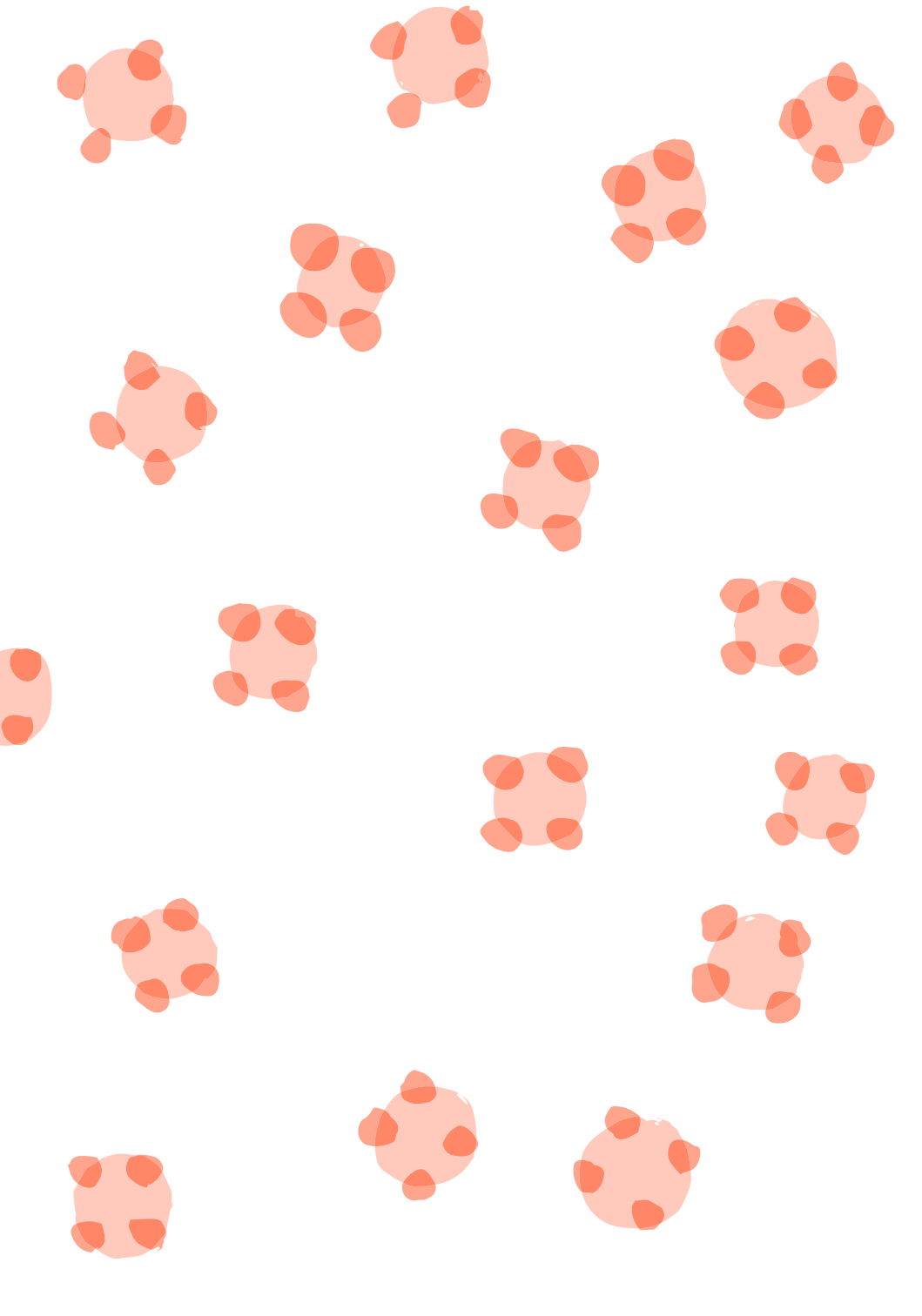
is an Assistant Professor and Director of the Privacy & Sustainable Computing Lab at Vienna University of Economics and Business. In 2014 he founded the Centre for the Internet and Human Rights (CIHR) at European University Viadrina and served as CIHR Director from 2014 to 2016. His research focuses on communications technology at the intersection between rights, ethics and governance. Ben holds a PhD in Political and Social Sciences from European University Institute in Florence. He was previously worked at the German Institute for International and Security Affairs, the University of Pennsylvania, Human Rights Watch and the European Council on Foreign Relations.

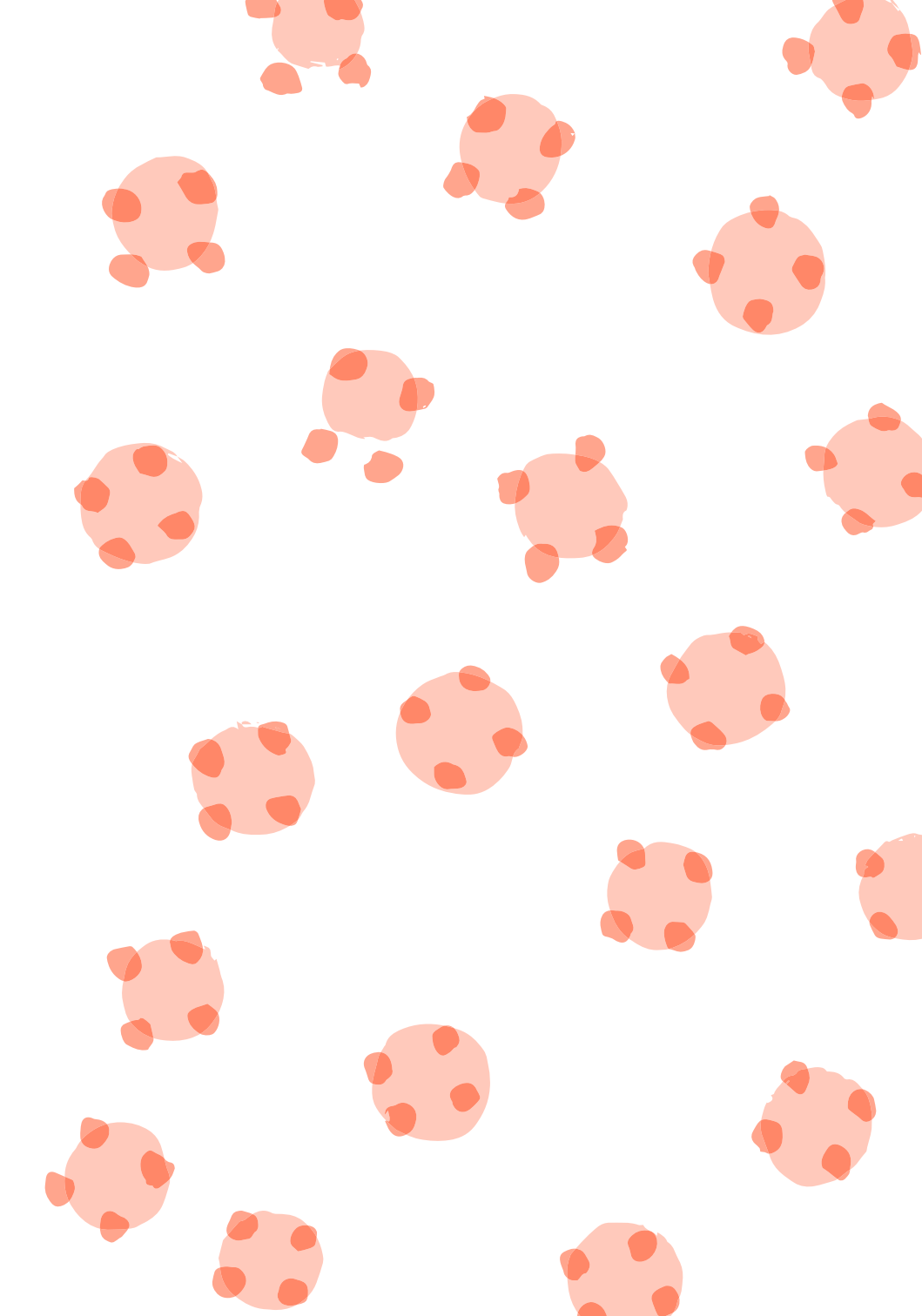
KAREN YEUNG

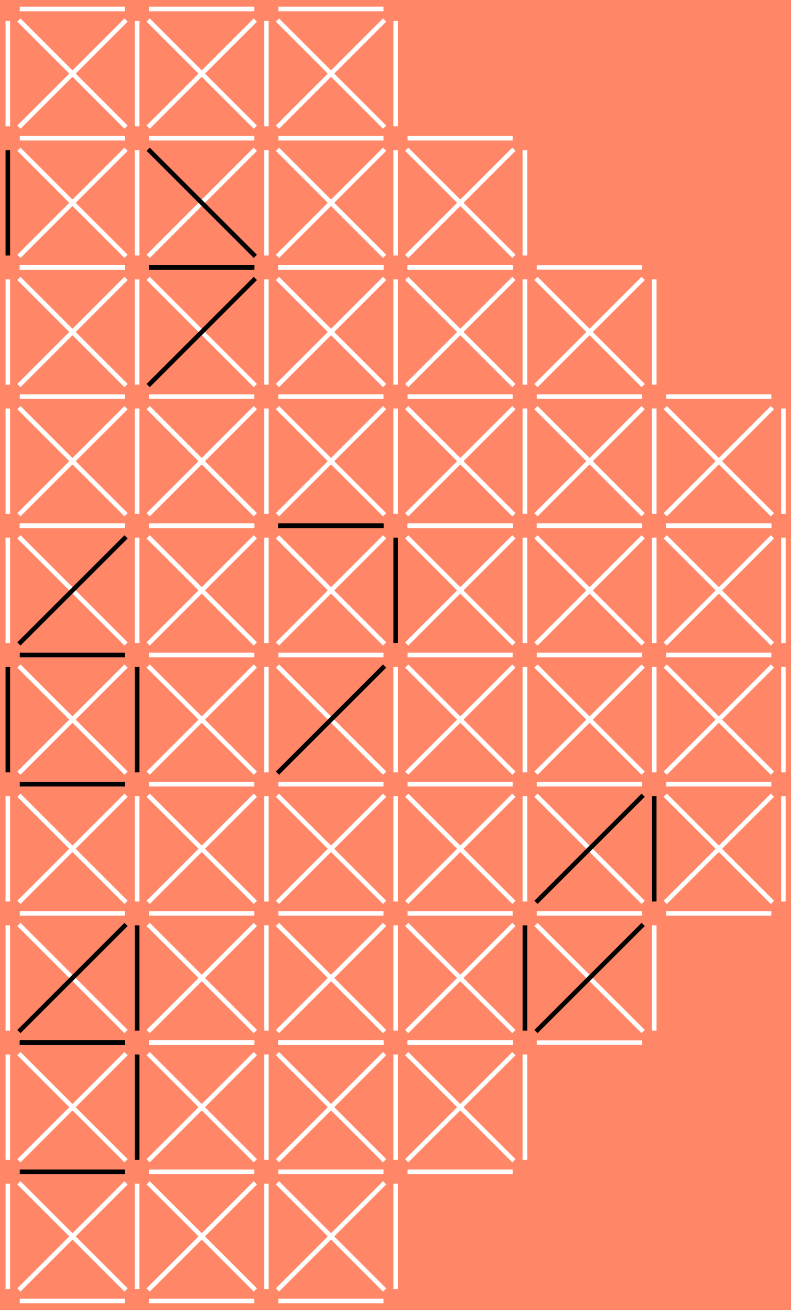
is Interdisciplinary Professorial Fellow in Law, Ethics and Informatics at the University of Birmingham in the School of Law and the School of Computer Science at the University of Birmingham. Her research interests include the governance of, and through, AI and other computational technologies. She is actively involved in several technology policy and related initiatives in the UK and worldwide, including membership of the EU's High Level Expert Group on Artificial Intelligence and she is a member and rapporteur for the Council of Europe's Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT). See https://www.birmingham.ac.uk/staff/profiles/law/yeung-karen.aspx.

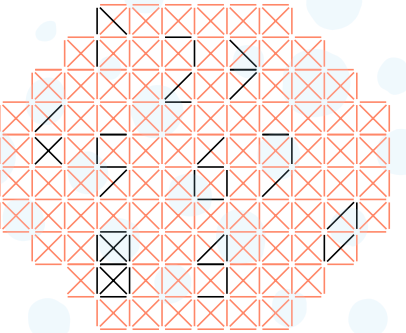












EMRE BAYAMLIOĞLU,
IRINA BARALIUC,
LIISA JANSSENS,
MIREILLE HILDEBRANDT
(EDS)