European University Institute

ACADEMY OF EUROPEAN LAW

EUI WORKING PAPERS

AEL 2020/01
Academy of European Law

Thinking Inside the Box: The Promise and Boundaries of Transparency in Automated Decision-Making

Ida Koivisto

European University Institute
Academy of European Law

Thinking Inside the Box: The Promise and Boundaries of Transparency in Automated Decision-Making

Ida Koivisto

EUI Working Paper **AEL** 2020/01

ISSN 1831-4066

Published in June 2020 by the European University Institute.
Badia Fiesolana, via dei Roccettini 9
I – 50014 San Domenico di Fiesole (FI)
Italy

Views expressed in this publication reflect the opinion of individual author(s) and not those of the European University Institute.

Abstract

The normative attractivity of transparency is beyond compare. No wonder it is one of the main principles in the EU's General Data Protection Regulation. It also features in a majority of AI ethics codes. Transparency is called for because it is assumed that it will solve the so-called 'black box problem' (uncertainty about how inputs translate into outputs in algorithmic systems) and by so doing legitimize automated decision-making (computer-based decision-making without human influence; ADM). In this paper, the legitimizing effect of transparency in ADM is discussed. I argue that transparency cannot deliver in its quest to resolve the black box problem. The main claim is that transparency is inherently performative in nature and cannot but be so. This performativity goes against the promise of unmediated visibility, vested in transparency. As demonstrated, when transparency is brought into the context of ADM, its hidden functioning logic becomes visible in a new way.

Keywords

transparency; automated decision-making; the black box problem; AI ethics; data protection

Acknowledgements

I would like to thank several people for their helpful comments on the paper and inspiring discussions on the topic. In particular I would like to thank Deirdre Curtin, Joyce Davies, Markus Dubber, Jenni Hakkarainen, Riikka Koulu, Marta Maroni, Beata Mäihäniemi, Suvi Sankari, Trond Ove Tøllefsen and Sam Wrigley. I would also like to thank Allyse Cruise, Melissa Craig and Katja Repo for research assistance.

1. Introduction: The End of Human Bias in Law?

Even lawyers cannot escape their humanity. In a famous study of Israeli judges, the admittance of parole crucially depended on whether or not the judge decided over it before or after having a break.[1] Humans are prone to be affected by their moral and political preferences, different sympathies and antipathies, and even bodily sensations such as hunger or fatigue. For better or for worse, this is part of what makes us human.

Today, it is hard to maintain a romantic vision of law as a simple, formulaic solution to complex problems in the real world. Informed by legal realism and critical legal studies, we are well aware of the indeterminacy and human bias in law. Law is not a system of flawless logic but a result of political contestation. This fragility also extends to the application of law. In the history of law in action, there are countless examples of bias, favouritism and different predilections of judges and bureaucrats determining people's rights and duties. This is understandable yet depressing. We have learned to accept this state of affairs, given that a viable alternative has not existed. Instead, we have focused on redress mechanisms. At least, they provide an opportunity for acquiring a second opinion, and ultimately, contestation.

Even though human bias may have been an intrinsic part of our legal system as we know it, the regulative ideal has always been there, guiding us like the flickering shadows in Plato's Cave. We would prefer to eradicate random factors when it comes to making decisions about people's lives. Indeed, in law, *ought* should not be derived from *is*. Even if a judge's rumbling stomach in fact affects their legal deliberation, in law that should not be the case. Law should strive for fulfilling its ideals of equality, justice and predictability; in other words, legal certainty. Hence, efficiency, accuracy and equality are still *modus operandi* of how to develop law, not the inherent caveats of human decision-making.

If we look at these ideals more closely, we can see that they seem better suited for machines to execute than whimsical humans. Consequently, we might think that replacing humans with machines would make the problems of human inefficiency and bias go away, and additionally, save considerable amounts of money.[2] The steady increase in computing power, the emergence of big data analysis and artificial intelligence research show much promise in developing less human, more just law application. At break-neck pace, computers seem to be gaining the ability to do things we never thought possible. Should we thus forfeit human decision-making and hand it over to computer programmes and algorithms? Especially in routine cases, automated decision-making – computer-based decision-making without human

[1] Danziger, Levav, and Avnaim-Pesso, 'Extraneous factors in judicial decisions', (17) 108 *Proceedings of the National Academy of Sciences* (*PNAS*) (2011) 6889.

[2] Sunstein, *Algorithms, Correcting Biases*, 12 December 2018, available online at https://ssrn.com/abstract=3300171 (last visited 12 May 2020). For another optimistic view, see Coglianese and Lehr, 'Transparency and Algorithmic Governance' (1) 71 *Administrative Law Review* (2019) 1.

influence (hereinafter, 'ADM') – could help us overcome our deficiencies and lead to an increased perception of fairness.[3] So, problem solved?

This seems not to be the case – even if we did not succumb to alarmist thinking and dystopias of machines taking over the world. There is growing evidence that human bias cannot be totally erased, at least for now.[4] It can linger in ADM in many ways, as I will specify later. As a result, it is not clear who is accountable for that. Are the codes involved to blame?[5] Or the creators of those codes?[6] What about machine learning and algorithms created by other algorithms? [7] Most of the time, we do not know answers to these questions.[8] This difficulty is often referred to as 'the black box problem'. We cannot be sure how the inputs transform into outputs inside the 'black box', and who is to blame if something goes wrong. As the potentially discriminatory nature of algorithmic predictions has been identified as a thorny – and I would claim from the legal perspective, *the* primary – problem in ADM, solutions to tackle that problem are actively sought.[9] In particular, law and regulation are called upon.

However, to date, legally binding regulation is mostly lacking.[10] As the standard lamentation goes, law and regulation are lagging behind technological developments.[11]So far, the EU's General Data Protection Regulation ('GDPR') carries the most promise in resolving the problem, as much of ADM is closely linked to processing personal data. However, legitimacy of algorithmic governance is a topical concern also beyond data protection. [12]

Consequently, soft law and self-regulation are increasingly resorted to. The number of different codes of conduct (AI ethics) skyrocketed in the years 2018-2019. As an independent NGO, Algorithm Watch, shows, the number is staggering. These codes of conduct are of various kinds and published by different institutions. Some of them are private (e.g. Partnership

[3] Cf. Binns, Van Kleek, Veale, Lyngs, Zhao and Shadbolt, '*It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions* (2018), available online at https://arxiv.org/pdf/1801.10408.pdf (last visited 12 May 2020).

[4] Castelluccia and Le Métayer, *Understanding algorithmic decision-making: Opportunities and challenges* (2019), available online at https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624261/EPRS_STU(2019)624261_EN.pdf (last visited 12 May 2020)

[5] Mittelstadt, Allo, Taddeo, Wachter and Floridi, 'The ethics of algorithms: Mapping the debate', *Big Data & Society* (2016) 1.

[6] Cf. eg. Bivens and Hoque, 'Programming sex, gender, and sexuality: Infrastructural failures in the "feminist" dating app Bumble', *43 Canadian Journal of Communication* (2018) 441.

[7] Barreno, Nelson, Sears, Joseph and Tygar, *Can machine learning be secure?* (2006), available online at https://dl.acm.org/doi/pdf/10.1145/1128817.1128824?download=true (last visited 12 May 2020).

[8] Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms', *Big Data & Society* (2016) 1.

[9] Zarsky, 'The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making', (1) 41 *Science, Technology, & Human Values* (2016) 118.

[10] For an overview of pertinent legal questions, see Desai and Kroll, 'Trust but Verify: A Guide to Algorithms and the Law' (1) 31 *Harvard Journal of Law & Technology* (2017).

[11] Cohen, however, argues that this assumption is dated and erroneous. It would be better to talk about dynamic interaction between law and technology. Julie Cohen, *Between Truth and Power. The Legal Constructions of Informational Capitalism* (2019) 4-5.

[12] Cf. Danaher, 'The Threat of Algocracy: Reality, Resistance and Accommodation' 29(3) *Philosophy & Technology* (2016) 245-268.

of AI – Google, Facebook, Amazon, IBM, Microsoft, Deep Mind), some public (e.g. High-level Expert Group on AI) while some are produced by different kinds of hybrid partnerships.[13]

What unites both the GDPR and a great majority of the AI ethics codes of conduct is the call for *transparency*.[14] This is hardly surprising, as the promise of transparency is overwhelmingly positive. Although transparency can be approached in a plethora of ways, as a normative metaphor, its basic idea is simple. It promises legitimacy by making an object or behaviour visible and, as such, controllable. No more black boxes, but X-rayed ones! A metaphoric solution is thus proposed to a metaphoric problem. On a more general level, the call for transparency aims at abolishing ignorance and opacity in a society by assuming active and well-informed citizens. At the same time, it presupposes an asymmetrical power structure between the one that exercises power and the one who is subject to it. To be legitimate, this unequal use of power needs to be accountable to the subjects.

As promising as it sounds, the legitimation narrative of transparency cannot really deliver in its quest to resolve the black box problem in ADM. Instead, I will argue that transparency is a more complex ideal than is portrayed in mainstream narratives. My main claim is that, contrary to what mainstream narratives suggest, transparency is inherently performative in nature, and cannot but be so. This performativity goes counter to the promise of unmediated visibility, vested in transparency.[15] Subsequently, in order to ensure the legitimacy of ADM – if we, indeed, are after its legitimacy – we need to be mindful of this hidden functioning logic of the ideal of transparency. As I will show, when transparency is brought to the context of algorithms, its peculiarities will become visible in a new way.[16] In this article, I will problematize the promise of transparency as the solution to the black box problem in ADM.

The paper is organized as follows. First, I will analyse the black box problem theoretically, discussing the logic of discovery and the logic of justification. Which one do we want to 'see' through transparency? I will also illustrate the nature of the black box problem in ADM with the help of examples from the US, Poland, and Finland. Second, I will discuss theoretically the ideal of transparency. As hinted, I argue that it is based on certain hidden functioning mechanisms, stemming from its nature as a visual metaphor, its *icono-ambivalence* and its performativity. These points of departure lead to an overall idea of transparency as an internally contradictory ideal, building on the so-called 'truth-legitimacy trade-off'. Third, I will apply this theory. I will discuss it in the context of ADM and more specifically, the GDPR. What functions and expressions do transparency have in that regulation? What are its implications? Fourth, I will draw the discussion together and conclude that, although

[13] Algorithm Watch, *AI Ethics Guidelines Global Inventory* (2019), available online at https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/ (last visited 12 May 2020).

[14] Hagendorff, *The Ethics of AI Ethics - An Evaluation of Guidelines* (2019), available online at https://arxiv.org/ftp/arxiv/papers/1903/1903.03425.pdf (last visited 12 May 2020).

[15] Cf. however Albu and Flyverbom, 'Organizational transparency: Conceptualizations, conditions, and consequences', (2) 58 *Business & Society* (2019) 268, who attribute both verifiability and performativity to transparency.

[16] Ananny and Crawford, 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability', (3) 20 *New Media and Society* (2018) 973, at 977-982. Also Bostrom, 'Strategic Implications of Openness in AI Development', *Global Policy* (2017) 1.

transparency is widely appreciated, there are weak signals indicating that its major legitimating narrative is not sustainable in the context of ADM.

2. The Black Box Problem

A. Logic of Discovery and Logic of Justification

Before we talk about the black box problem in ADM, a few words about the black box in general are needed. What is it exactly? Why are we using this particular metaphor? Why is it a problem? Or, as Taina Bucher asks, what is at stake in framing algorithms in this way, and what might such a framing possibly distract us from asking? Although the black box would well deserve critical deconstruction in the same way as the notion of transparency, that cannot be done fully here. [17] That said, we will start by busting two common myths about the black box.

First, the metaphor of a black box need not have anything to do with technology, although technology is the context in which it is most often mentioned. Instead, a black box simply refers to a condition whereby the way in which an input translates into an output is unknown or un-knowable. That is to say, a human judge makes a black box too. Indeed, the way in which human data processing works, is hardly any clearer to us than black box algorithms.[18] Second, despite its common negative connotation – rendering the unknown an epistemological problem [19] – the black box can also be seen as value-neutral. For example, it is often approached neutrally in computer science, as a feature of a system. A black box does not need to arouse protest. The lack of transparency only becomes a problem if the outputs prove to be undesirable.

A black box may be a black box to itself, too. A judge does not really know, let alone be able to express, how her neurons are shooting when she is pondering the intricacies of a case. Something happens in the brain and different connections are brought into consciousness. This process is unfathomable even when it is happening within our own brain. Similar processes may take place in computer systems, in particular in machine learning and deep learning neural networks i.e. – software which learns by itself through inferring regularities in provided training data.[20]

There is a catch, though. Some of us would be ready to argue against that: judges do need to know how they are solving the case. The same applies for ADM – reasons must be given for the decision to be acceptable. This is exactly why we call for transparency and giving reasons as a condition for judicial legitimacy. If that were not the case, anything would go. True, judges do need to be able to explain themselves. Nevertheless, we need to make an

[17] Bucher presents a critical genealogy of the metaphor. Bucher, *If… then. Algorithmic Power and Politics* (2018) 41- 65, 44.

[18] As a societal phenomenon, see Pasquale, *The Black Box Society. The Secret Algorithms That Control Money and Information* (2015).

[19] Bucher 2018 at 44.

[20] On the different transparency standards in humans and machines, see Zerilli, Knott, Maclaurin and Gavaghan, 'Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?', 32 *Philosophy & Technology* (2018) *661*. Analysis on AI opacity, see Carabantes, 'Black-box artificial intelligence: An epistemological and critical analysis', *AI and Society* (2019). See also Lipton, The Mythos of Model Interpretability. *arXiv:1606.03490* (2017)

important distinction, which is known in philosophy of science as *the logic of discovery* and *the logic of justification*.

What do these logics mean and how do they differ from each other? The logic of discovery is a description of the empirical process by which one's brain automatically finds patterns, similarities and connections between perhaps seemingly unrelated things. This logic can be hard to account for. How, indeed, can we know or explain why certain ideas and associations rush into our consciousness following certain stimulus in a given moment? We cannot. Even if we could do that, the explanation might sound random and weird. Indeed, how do we convince someone that the taste of a madeleine dipped in tea brings an entire array of memories to our own mind?

Hence, the logic of discovery does not seek to convince others. It just describes how a heuristic process goes. Not so for logic of justification, which does attempt to convince. It is no less than the basic principle that underpins legal argumentation or giving reasons for a decision. According to the logic of justification, we need to justify, step by step, why the associations we make should be accepted, why the suggested correct answer to a given question is indeed correct. This is premised on the idea of shared understanding of how logical reasoning should take place.

As a result, the logic of justification is not limited to the way in which our private associations are built. Therefore, if we want to convince others with our argument, we need to make our thinking look like it was following a predetermined, rational logic, and only that logic, even if the logic of discovery would suggest otherwise. The logic of discovery may be of interest to a psychoanalyst, but hardly a subject of a legal decision. As philosopher Karl Popper argues, "*My view of the matter…is that there is no such thing as a logical method of having new ideas, or a logical reconstruction of this process. My view may be expressed by saying that every discovery contains 'an irrational element', or 'a creative intuition' in Bergson's sense."*[21]

How do the logic of discovery and the logic of justification relate to the question of the black box? What is their explanatory power in this context? As mentioned, a black box need not be approached as a problem. However, if we do so, implying that we should attempt to get rid of it, we covertly encounter the question of the two different logics. By opening the box, or making it transparent, we want to see the way in which the inputs translate into outputs. To that end, we need to specify which logic we want to see: the logic of discovery or the logic of justification?

As explained, the logic of discovery and logic of justification are subject to different kinds of rationalities. The first is purely descriptive or empirical while the second is normative and somewhat formulaic. Logic of discovery is the process of emerging ideas – however irrational or haphazard this process would be – whereas the logic of justification aims for general acceptance, following certain rules.

Would we want to know how the black box actually operates, regardless of whether we can understand the process? Alternatively, do we want the box to explain and justify itself, to convince us of why it follows the exact steps it does, and why we should accept its outputs?

[21] Popper, *Logic of Scientific Discovery* (2nd ed., 2002) 7-8.

This distinction is helpful although, to my knowledge, it has not been applied in this context before. I will come back to this theme when discussing the transparency requirements laid down in the GDPR.

B. Examples of the Black Box Problem in ADM

To summarize, the black box need not be a problem *per se*. However, we speak of 'a problem' in the context of algorithms and ADM for very good reasons – biases and other harms do happen.[22] As Frank Pasquale states, black boxes must be exposed to counteract any wrongdoings, discrimination, or bias these systems may contain: "*algorithms should be open for inspection—if not by the public at large, at least by some trusted auditor.*" [23]

Where do these potential wrongdoings come from? At least from two directions.[24] First, the code on which the ADM system is based can be poorly designed. That is to say, the coders may, deliberately or unbeknownst to themselves, favour choices that advantage some people over others.[25] This kind of bias is similar to those of the described judges: they are also affected by attitudes, preferences and bodily sensations – not to mention certain background variables such as gender, religion, ethnicity or culture.[26] These things may further affect the code, resulting in outputs which may be biased or otherwise unanticipated.

Second, particularly when it comes to machine learning and deep learning neural networks and big data, the bias shifts its shape. The human bias may fossilize in the very data. As learning algorithms need large amounts of data to recognize patterns in it, these patterns may prove discriminatory, crucially, because we humans are the source of those data. It reflects who we are and how we tend to behave – not how we should behave. It thus derives *ought* from *is*. When these outputs based on the skewed inputs are used as a basis for future predictions, they may actually reproduce the bias in it, and thus create self-fulfilling prophesies ('garbage in, garbage out').[27] Even if we have a neutral process, we do not necessarily end up with a neutral outcome.

Let us approach these questions through examples. The best-known example comes from the US: An article by *Pro Publica* created a scandal in 2016.[28] The article discusses the software used for assessing the recidivism potential of captive perpetrators in several states of

[22] For an illustrative inventory of potential algorithmic harms, see Future of Privacy Forum, *Unfairness by Algorithm: Distilling the Harms of Automated Decision-Making* (2017), available online at https://fpf.org/wp-content/uploads/2017/12/FPF-Automated-Decision-Making-Harms-and-Mitigation-Charts.pdf (last visited 12 May 2020).

[23] Pasquale 2015 at 141.

[24] Cf. Yeung, 'Why Worry about Decision-Making by Machine?' In Yeung and Lodge (eds), *Algorithmic Regulation* (2019) 21-48.

[25] O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (2017); Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (2018).

[26] There has even been discussion of racist soap dispensers, which allegedly do not recognize dark skin to work.

[27] Cf. Kerr, 'Prediction, pre-emption, presumption: the path of law after computational turn', in Hildebrandt & de Vries (eds), *Privacy, Due Process and the Computational Turn - The Philosophy of Law Meets the Philosophy of Technology* (2013) 91.

[28] Angwin, Larson, Mattu and Kirchner, *Machine Bias* (2016), available online at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (last visited 12 May 2020). See also Chouldechova, 'Fair prediction with disparate impact: a study of bias in recidivism prediction instruments', (2) 5 *Big Data* (2016) 153.

the US. The idea was to give a person a numeric score, ranging from 1 (low risk) to 10 (high risk), reflecting the likelihood of re-offending. This score was further used e.g. for assessing whether or not the person could be granted parole or be released prior to trial. The idea behind this is understandable: by achieving accuracy by ADM, it would lower crime and state costs, by separating low-risk and high-risk prisoners, and releasing those considered low-risk.

Nevertheless, the reality was less rosy, as was noticed by *Pro Publica* in its investigation. The algorithm systematically discriminated against blacks, giving them significantly higher risk scores than whites on average. NorthPointe, the enterprise which had created it, claimed the algorithm was a trade secret. Thus, it was impossible to know in what way it concluded that blacks were more like to reoffend than whites, and how the scores were actually calculated.

Certainly, it was not explainable by the previous criminal history of the people processed; rather, the history and the score clearly did not match. The set of questions, on which the score was at least partially based, did not include race. However, it did include questions mapping the potential rehabilitation needs of the person, such as questions of drug abuse and incarcerated friends, which nevertheless seemed to correlate with race.[29] Some backlash against the scoring system has emerged. There was even a court case against the use of the algorithm (State v. Loomis) as an alleged violation of Mr. Loomis' due process rights.[30] However, the court concluded that the use of the software was possible so long as the decisions were not solely based on it. [31]

Let us take another example from Poland. The Polish Ministry of Labour and Social Policy introduced a new system of granting unemployment benefits in 2014. It was based on a survey and an interview, which functioned as input of a score. The unemployed needed to fill in a form with a set of questions, indicating, for example, the reason for unemployment. Although there was blank space left for answering seemingly open-ended questions, in reality there were 22 predefined answers to those questions. The questionnaire also did not recognize certain reasons, such as homelessness or ethnic origin or being a convicted felon, as a valid reason, although in practice these were major employability impediments in the Polish labour market.

According to the acquired score, the applicants were sorted into three different categories. The first category of people was considered the most employable, having a high educational level and unemployment stemming from some haphazard personal or market

[29] A similar case was found in the USA health care system, in which a commercial algorithm concluded on the basis of medical costs data that black patients need less medical care than whites. The reason for this turned out to be that blacks were previously granted less treatment than whites, not that they were healthier. By inductive reasoning, the algorithm started to reproduce that pattern. Obermeyer, Powers, Vogeli and Mullainathan, 'Dissecting racial bias in an algorithm used to manage the health of populations', (6464) 366 *Science* (2019) 447.

[30] See further, Liu, Lin and Chen, 'Beyond State v Loomis: artificial intelligence, government algorithmization and accountability', (2) 27 *International Journal of Law and Information Technology* (2019)122.

[31] From the Freedom of Information Act point of view in the USA, see Fink, 'Opening the Government's Black Boxes: Freedom of Information and Algorithmic Accountability', (10) 21 *Information, Communication & Society* (2018) 1453. Later on, the discourse has been diversified, and the "anti-discrimination" discourse is considered sometimes too simplistic, overlooking, for example, questions of intersectionality. See Hoffmann, 'Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse', (7) 22 *Information, Communication & Society* (2019) 900.

reason. They were only 2 % of the applicants. The second category of applicants was somewhat worse off, although still having some important skills (65 %). They were considered potentially in need of some additional education, skills and support. The last category was the most problematic. It consisted of people on whom more of life's adversities seemed to accumulate: illness, drug abuse, lack of education, marginalization (33 %).[32]

Each of these categories were entitled to a different menu of benefits according to their needs. However, also in this case, there were hidden problems. Namely, there was virtually no possibility of contesting one's categorization; no information was available on the scoring rules. In addition, the array of different benefits and other supporting services were unevenly distributed so that they were the least available to the third category of people, who obviously needed them the most. In other words, the system was largely considered discriminatory, lacking transparency and infringing data protection rights. This system of organizing unemployment governance caused resistance, most prominently by a civil society organization. In the end, Poland's constitutional court found that the system breached the constitution, although mostly due to reasons of legislative form. As a result, the scoring system was abolished in 2019. [33]

The third example comes from Finland. Unlike the two previous examples, this case took place in a commercial context. It concerned internet commerce and different financing options while purchasing building materials online. The applicant was a man, living in rural Finland. His mother tongue was Finnish, as is the case with the vast majority of Finns. He had no prior record of disruptions of payment, or any problems in his credit history. These facts proved relevant, as he was denied the option of a partial payment arrangement. The decision was reached by using statistical methods in the ADM of the bank which was cooperating with the construction materials company.

According to the statistics used as the basis to create the algorithm, Swedish speakers and women were more likely to pay back their loans than Finnish speakers and men. The algorithm was found to favour Swedish-speaking women over Finnish-speaking men. In other words, the applicant was denied a financing option because of his gender, age, place of residence and mother tongue, and their cumulative effect. The rejection of the loan application was thus caused by profiling, not an individual assessment of creditworthiness. The case was considered both by the anti-discrimination ombudsman and, due to her initiative, the National Non-Discrimination and Equality Tribunal of Finland. The tribunal found the firm guilty of multi-reason discrimination. It was given a fine and ordered to discontinue the discriminatory practice.[34]

[32] See Jędrzej, Sztandar-Sztanderska and Szymielewicz, *Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic decision making* (2015), available online at https://panoptykon.org/sites/default/files/leadimage-biblioteka/panoptykon_profiling_report_final.pdf (last visited 12 May 2020).

[33] Jedrzej, *Poland: Government to scrap controversial unemployment scoring system* (2019), available online at https://algorithmwatch.org/en/story/poland-government-to-scrap-controversial-unemployment-scoring-system/ (last visited 12 May 2020).

[34] Register number: 216/2017, Date of issue: 21 March 2018. Available online at https://www.yvtltk.fi/en/index/opinionsanddecisions/decisions.html (last visited 12 May 2020).

These examples are by no means exhaustive. On the contrary, with the increasing use of ADM, more similar cases are emerging. Although the three examples presented above are quite different in context and consequence, they also share some important similarities. First, the examples represent the larger development of the emergence of a "scored society", a new way of quantifying and ranking people.[35] These resulting profiles are made of stereotypes of individuals based on certain characteristics, such as wealth, gender, habits, education, etc. Profiling requires individual information. This information, however, leads to simplification and generalization – to the treatment of people as representatives of a certain category rather than unique individuals.[36] Some of these profiles have proven discriminatory, as illustrated.

This brings us to the second similarity: there is a lack of information about the scoring rules. Accordingly, there were only limited possibilities to react to the breach of individual rights. It can even be unclear whether any rights have indeed been violated. Third, they all have caused backlash and protest. With varying success, we can see that there are remedies available. It is debatable, though, whether they are well suited to the legal problems of a scored society. How many problems like the described examples go completely unnoticed?

In the context of people's rights, possibilities, and equal treatment, the black box indeed looks like a problem rather than just a neutral feature. The algorithms in use are not available, visible or understandable to the people who, however, are subjects to their silent and seemingly unerring power. This may lead to potential approval of unjustified categorizations and treatment, if the black box just produces a score or the loss of an opportunity without giving reasons why that is so. Neither logic of discovery nor logic of justification can be seen: there is no transparency. In the following, I analyse the ideal of transparency more closely.

3. Transparency and its Covert Human-Faced Logic

As mentioned, transparency carries much promise in solving the black box problem. In Ananny and Crawford's words, "*The more that is known about a system's inner workings, the more defensibly it can be governed and held accountable*."[37] Depending on the context, transparency can mean different things in ADM. It can be associated to source code publicity, auditing and impact assessment, to mention but a few.[38]

On a more fundamental level, however, transparency is a major socio-legal ideal, which seldom encounters resistance or questioning. It assumes that when we see by ourselves, we can understand what is happening. By virtue of this eye witnessing, we can further fix what needs to be fixed. In public policy context, this general promise has been inherent in this justificatory

[35] Citron and Pasquale, 'The Scored Society: Due Process for Automated Predictions', (1) 89 *Washington Law Review* (2014) 1.

[36] For an overview of profiling, see Hildebrandt and Gutwirth (eds), *Profiling the European Citizen. Cross-Disciplinary Perspectives* (2008).

[37] Ananny and Crawford 2016 at 2. See also Laat, 'Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?' (4) 31 *Philosophy & Technology* (2018) 525.

[38] Felzmann, Fosch-Villaronga, Lutz and Tamò-Larrieux, 'Robots and Transparency: The Multiple Dimensions of Transparency in the Context of Robot Technologies', (2) 26 *IEEE Robotics & Automation Magazine* (2019) *71*.

narrative of transparency from its very inception – from the era of the Enlightenment, that is.[39] Similarly, algorithmic transparency follows a simple logic: if we could only open up the algorithmic black boxes and see their inner workings we could make sure they are fair.[40] The unknown—including the black box—is thus considered problematic because it obscures vision. Like that, it undermines the Enlightenment imperative of *sapere aude*, 'dare to know,' 'have the courage, the audacity, to know.'[41] Thus, transparency is regarded as an apt cure for this, as it specifically promises clear vision.

Nevertheless, transparency has proven more complex than its basic promise suggests. In order to delve into the potential of transparency for solving the black box problem, we need to discuss the hidden functioning logic of transparency on a more fundamental level. In the following, I will approach this logic from three angles: transparency as a visual metaphor, transparency as an icono-ambivalent ideal, and the latent conjunction between transparency and intentionality. This all will lead to an overall idea, which I call *the human-faced logic of transparency*.[42] This logic has quite dramatic consequences concerning the general promise vested in transparency, and consequently, the specific promise in ADM.

First, let us start with the metaphor. We can notice that as a concept, transparency appeals specifically to our vision, our ability to see things with our own eyes. We cannot hear transparency, nor smell or taste it. Therefore, it could be called a visual or an ocular-centric arrangement. Perhaps, we can better grasp this idea when we think of transparency as looking through a window. Something is made directly and intentionally visible to the viewer, which otherwise would stay hidden. Without transparency, we cannot see, but with transparency, we can. [43]

This promise underpins transparency as a metaphor. It requires that transparency is approached as a figurative placeholder for different practices, which provide information about its object. The visual undertow of transparency makes it understandable and attractive to us even in cases when we are talking about abstractions such as governance. So long as we can witness the reality with our own eyes, we do not need verbal explanations, which are, by virtue of transparency, indirectly considered less reliable than direct visual observation. This is the very core, I argue, why it is so appealing to us. Seeing by oneself seems to be self-authenticating: seeing is understanding, understanding is seeing.

However, transparency is not only a metaphor. Additionally, the literal meaning of transparency belongs to its functioning mechanisms. Sometimes transparency is organized as

[39] Hood, 'Transparency in Historical Perspective' in Hood and Heald (eds), *Transparency – Key to Better Governance?* (2006), 3, 6-7; Baume and Papadopoulos, 'Transparency: from Bentham's inventory of virtuous effects to contemporary evidence-based skepticism' (2) 21 *Critical Review of International Social and Political Philosophy* (2018) 169.

[40] Lepri, Oliver, Letouzé, Pentland and Vinck, 'Fair Transparent, and Accountable Algorithmic Decision-making processes', (4) 31 *Philosophy & Technology* (2018) 611.

[41] Bucher 2018 at 44. "Sapere aude" is particularly known from Immanuel Kant's thinking.

[42] For the outline of the theory, see Koivisto, *The Anatomy of Transparency: The Concept and its Multifarious Implications, EUI Working Paper MWP* 2016/09, available online at

https://cadmus.eui.eu/bitstream/handle/1814/41166/MWP_2016_09.pdf?sequence=1&isAllowed=y (last visited 12 May 2020).

[43] Christensen and Cornelissen, 'Organizational Transparency as Myth and Metaphor', (2) 18 *European Journal of Social Theory* (2015) 132, 133.

direct see-ability. This can be exemplified by glass walls or roofs of public buildings.[44] They allow people to see what is happening in the chambers of powers. Transparency can thus work, in Jeremy Bentham's sense, as inspective architectures. As a result, transparency oscillates between two functionalities: *transparency as actual, visual see-ability allowed by an optical arrangement* and *transparency as a metaphor for verbal practices of self-reporting*.

When referring to a governance ideal, transparency can thus mean both *actual, material transparency* and *metaphorical, as if transparency*. The *as if* aspect of transparency becomes understandable when we refer to see-ability or knowability of abstract things, which typically lack physical appearance. What is there to see when we talk about abstractions such as governance or decision-making? Indeed, nothing. Social constructs such as governance only exist in our collective imagination and can only be understood through symbols and hints. For example, how could we use transparency to see something as abstract as a person's recidivism risk?

This oscillation between literal and metaphorical meaning[45] brings us to the second, not obvious aspect of transparency: its icono-ambivalence. What does that monstrous neologism mean? It refers to another, internal duality of transparency: on the one hand, transparency is ideologically *iconoclastic*. It is suspicious of images, explanations and mediation – ultimately: humans. It attempts to strip governance from all kinds of obfuscating veils: secrecy, appearances and concealment. It promises to allow governance itself to emerge in its pure essence before the eyes of the viewer. Following that reasoning, the transcendence of governance, if you will, would take care of its own representation so long as the impediments blocking its visibility for the viewer were removed.

On the other hand, I argue, transparency is also *iconophilic* and necessarily so. If iconoclasm is the ideological aspect of transparency, iconophily is its unescapable practicality. In many cases, there is nothing to show, to emerge, without conscious efforts and constructs. Therefore, transparency needs to rely on images, metonymically understood: illustrations, statistics, reports, memoranda etc. – conscious, constructed appearances, mostly falling into the category of documents.

In this sense, transparency needs to rely on people and their mimetic abilities, their capabilities to 'capture' the essence of governance and to communicate it to the public. The iconophilic aspect of transparency thus refers to the accessibility of those created illustrations of intangible abstractions. For example, a score of one's employability is an iconophilic expression of a social construct, which does not exist naturally in the world. The icono-ambivalence of transparency leads to a paradox: transparency means, in Emmanuel Alloa's words, mediated immediacy.[46] It both is, and it needs to be, constructed.

[44] Cf. Rowe and Slutzky, 'Transparency: Literal and Phenomenal', Vol. 8 *Perspecta* (1963) 45-54; Fisher, 'Exploring the legal architecture of transparency'. In Ala'i and Vaughn (eds) *Research Handbook of Transparency* (2014) 59-79

[45]Alloa, 'Transparency: A Magic Concept of Morality', in Alloa and Thomä (eds), *Transparency, Society and Subjectivity: Critical Perspectives* (2018) 21, 31–32. Also Flyverbom, 'Transparency: Mediation and the Management of Visibilities', 10 *International Journal of Communication* (2016) 110, 113.

[46] Alloa 2018 at 21–55.

The complexity of transparency does not end there, however. As mentioned, transparency is associated with legitimacy: it is generally considered something desirable. Transparency is good, whereas the lack of transparency is bad. How does this legitimating effect work?[47] To answer that, we need to address the third aspect of the hidden functioning logic of transparency, namely that of intentionality. We can detect the significance of intentionality with careful analysis of language. Although in public discourse transparency is almost entirely treated as a positive thing – regardless of whether it is seen as iconophilic or iconoclastic – it also entails a negative connotation.

It is important to notice that transparency is not only a virtue, but under certain circumstances, a sign of failure. This contention has its roots in a linguistic observation, available for anyone to test: "You are so transparent! I can see through you!" we might say, when we notice someone's failure to come across in a certain, predetermined way. In that case, the attempt is implausible to the extent that we cannot but see the "truer truth" behind that leaking appearance, or at least we think we do. Perhaps counter-intuitively, we resent this revelation. We prefer hidden motives to be hidden, and value transparency only when it is intentional. This negative if not pejorative connotation of transparency is completely unnoticed and consequently untheorized in current academic literature on transparency.

Hence, transparency is regarded as a value when it is consciously created or allowed but frowned upon when it is a sign of involuntary revelation, signifying the incapability to keep hidden things hidden. Unintentional transparency refers to the lack of control, which we tend to abhor. This intentionality is the key to the paradoxical, and as was mentioned, the human faced nature of the ideal of transparency. This dynamic of transparency has largely remained unexplored in the academic literature on transparency. However, it has huge implications when we think about the promise and beliefs vested in transparency.

This is to say, transparency, both referring to social life and as a governance ideal, is closely linked to prestige, appearance, favourable impressions, and in case of failure, loss of strategy, or the emergence of shame. Involuntary transparency makes one appear in an unplanned way. It is about mediating of what can be seen. In other words, it is about managing visibilities.[48] The key word that captures this dynamic is *impression management*. It is a term coined by social psychologist Erving Goffman in his seminal work 'Presentation of Self in Everyday Life' (1959). In it, he explains how social life is, and cannot but be, performative in nature: we carefully plan how we want to appear to others, and what part of our lives we want to keep to ourselves, in turn. This enables us to have a face, a social persona. In that way, transparency is a narcissistic ideal.

I argue that a similar mechanism is characteristic of institutions. They, too, have an interest to uphold a certain image, a certain face, to control what information they release. If that were not the case, information leaks, for example, could not lead to such scandals as they

[47] Curtin and Meijer, 'Does Transparency Strengthen Legitimacy?', (2) 11 *Information Polity* (2006) 109. See also de Fine Licht, *Magic Wand or Pandora's Box? How Transparency in Decision Making affects Public Perceptions of Legitimacy* (2014).

[48] Cf. Flyverbom, 2016 at 110, also Flyverbom, *The Digital Prism: Transparency and Managed Visibilities in a Datafied World* (2019)

often do. As a result, it is possible to hypothesize that the use of different transparency practices – whether physical or metaphorical, iconoclastic or iconophilic – are motivated by this very goal: to appear in a favourable light.

If we take this idea to the extreme, we reach a rather radical conclusion. The ultimate logic of transparency can be called the truth-legitimacy trade-off. It means that by intentional transparency more legitimacy is achieved, but most probably, it is based on a carefully curated picture of reality. If, on the other hand, there is no such curation, there will be more extensive access to information, but most probably, less legitimacy, because the less flattering elements of reality would also be subject to external gaze. This is premised on the idea that only intentional transparency is capable of creating legitimacy. The image created by transparency is designed to be seen, it delivers managed visibilities.

In the context of this paper, it is not possible to delve into this human-faced logic of transparency more deeply. That said, the most important implication of the logic needs to be highlighted: transparency as an ideal is not neutral visibility or an undistorted flow of information. When something is framed as "transparency", it is also planned to deliver a particular kind of message, to enable its deliverer to uphold a persona, a face. This message may be constructed or allowed to emerge, depending on the context. In any case, the release is controlled. In other words, we do not only see through transparency, we also see the created transparency, which makes the medium the message.

In the context of ADM, the human agent caring about her appearance to others may be distant if not, in some cases, completely absent. Regardless, the main feature of planned visibilities remains, as I will argue in the following. If a scoring algorithm for credit-worthiness software was deliberately revealed in the name of transparency, that could increase the legitimacy of the releasing institution. If it were leaked, instead, we would be equally informed but most likely less impressed; we would assume they have something to hide.[49] However, in ADM, the particular object of transparency – an algorithm – complicates the issue further. It proves the insufficiency of transparency as a cure-all concept. This is will be discussed in the following.

4. The EU's GDPR: Law Coupling Transparency and ADM

A. *Transparency Portrayed in the GDPR*

I have now presented some of the key factors of my general theory of transparency as a socio-legal ideal. These factors function as tools for analysis in assessing the potential of transparency to solve the black box problem in ADM. What happens to the human-faced logic of transparency, if, at least seemingly, humans are no longer always the gatekeepers of information and the managers of impression? Is there anyone to reveal or conceal? Alternatively, does human mediation govern transparency also in ADM; if so, what follows from that? What is the role of law in this? To analyse this field of questions, we need to move to a somewhat more

[49] Cf. Gibbs, 'Sigmund Freud as a theorist of government secrecy' *Research in Social Problems and Public Policy,* 19 (2011) 5-22, 15: "The underlying principle in law, psychology and historical research is the same: When people make declarations that go against their interest, such declarations have high credibility."

concrete level of discussion. As mentioned, the EU's General Data Protection Regulation (GDPR) is so far the most sophisticated legal attempt to solve the black box problem with the help of transparency. This is why it is worth a closer look. How is transparency portrayed in it and the ADM regulations it includes? [50]

GDPR has been applied since May 2018. As is widely known, it has changed the data protection regime in the EU, not least due to the increasing use of ADM.[51] The aim of the GDPR is to "harmonize data privacy laws across Europe, to protect and empower all EU citizen's data privacy, and to reshape the way organizations across the region approach data privacy".[52] Most crucially, it has created new rights for data subjects and new duties for data controllers. It is also built on a risk-based approach, which obliges the data controllers to assess the effects of processing personal data. The regulation is extensive, complex and, I would maintain, somewhat difficult to decipher. Consequently, it includes many interesting research themes, and indeed, the amount of academic writing on the topic is on rise.

In ADM, the questions of privacy and data protection go hand in hand with the call for transparency. Transparency is expected from the data controllers and from ADM, and data protection and privacy are demanded for the data subjects. This is because algorithmic models – such as different scoring and profiling tools employed in ADM – typically feed on huge amounts of personal data. That is necessary for them to form accurate outputs, as was illustrated through the examples above. Those personal data further originate from data subjects, and they are valuable raw material for a data-driven economy. Therefore, we should be keenly interested in how these data are gathered and handled. How can we know whether there is an illegal or unethical bias involved?[53] As presented, we often cannot. This ignorance is a growing legal concern, to which the call for transparency is closely connected. Would it help, then, if the data processing were made transparent to the data subjects? It is believed so.

This belief in transparency has strong institutional support in the GDPR. Transparency is one of the key principles of the entire regulation along with fairness and lawfulness.[54] ADM, in turn, is regulated specifically in article 22 ("Automated individual decision-making, including profiling"). [55]

[50] For a detailed analysis on transparency in GDPR, see Felzmann, Fosch-Villaronga, Lutz and Tamò-Larrieux, 'Transparency You Can Trust: Transparency Requirements for Artificial Intelligence between Legal Norms and Contextual Concerns', (1) 6 *Big Data & Society* (2019).

[51] The EU's AI Strategy has also begun to take form. See White Paper On Artificial Intelligence - A European approach to excellence and trust, Brussels, 19.2.2020 COM(2020) 65 final, available online at https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (last visited 12 May 2020)

[52] Available online at https://www.europarl.europa.eu/committees/en/libe/events-nationalparl.html?id=20180419MNP00301 (last visited 12 May 2020).

[53] See further e.g. Hacker, 'Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law', (4) 55 *Common Market Law Review* (2018) 1143.

[54] "Personal data shall be -- processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency')". Council and Parliament Regulation 2016/679, OJ 2016 L119/1 ("GDPR") 5(1) (a).

[55] Cf. Temme, 'Algorithms and Transparency in View of the New General Data Protection Regulation',(4) 3 *European Data Protection Law Review* (2017) 473.

> Art. 22 GDPR Automated individual decision-making, including profiling
>
> (1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
>
> (2) Paragraph 1 shall not apply if the decision
>
> a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
>
> b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
>
> c) is based on the data subject's explicit consent.
>
> (3) In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
>
> (4) Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

Although as a main rule, the article defines a right not to be subject to ADM alone when there are legal or similar kinds of effects on the individual, it also defines a number of exceptions when it is, in fact, allowed. Some writers even argue that this hollows out the entire right not to be subject to ADM, making the exceptions the main rule.[56] However, when ADM is applied by virtue of the exceptions laid down in article 22, it does not mean that data controllers can forget about the related data protection issues, including the call for transparency. Indeed, it can be argued that these very exceptions make transparency relevant in ADM.

As mentioned before, the background assumption in the call for transparency is an asymmetrical power structure. Here, that structure emerges between the data controller and the data subject. Therefore, accountability mechanisms are needed. To that end, article 22 needs to be read together with articles 13-15, which regulate the rights of the data subject to information and access to personal data.[57] The idea is that a data subject should be sufficiently informed on how her data are being handled, also when ADM is in question.

When it comes to the black box problem and transparency as its potential solution, there is a specific formulation in articles 13-15 which is worthy of closer analysis. That is to say, those articles require "meaningful information" as a right of the data subject to ensure fair and transparent processing. The formulation is virtually identical in all of the articles 13-15. In article 13(2)(f), for example, it says that:

[56] Brkan, 'Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond', (2) 27 *International Journal of Law and Information Technology* (2019) 91, 119-120.

[57] Information to be provided where personal data are collected from the data subject, 13(2)(f) GDPR; Information to be provided where personal data have not been obtained from the data subject, 14(2)(g) GDPR; Right of access by the data subject, 15(1)(h) GDPR.

> "In addition to the information referred to in paragraph 1, the controller shall, at the time when personal data are obtained, provide the data subject with the following further *information necessary to ensure fair and transparent processing*:" - -
>
> (f) "the *existence of automated decision-making,* including profiling, referred to in Article 22(1) and (4) and, at least in those cases, *meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject*. [Italics mine.]

We can notice that information is required both of the *existence* of ADM, and in that case, at least, *meaningful information of the logic involved* and *the envisaged consequences of such processing for the data subject*. In other words, the data controller needs to consider the entire lifespan of the ADM and provide information extensively, although some of it might be speculative. Additionally, article 12 specifically defines *transparent information*: "The controller shall [provide information] in a *concise, transparent, intelligible and easily accessible form, using clear and plain language*, in particular for any information addressed specifically to a child" [italics mine].

Regardless of these many paragraphs, the extent and the quality of information furnishing obligations has proven somewhat unclear. In academic literature, the enigmatic formulation of "meaningful information" has caused much debate: do these mentioned paragraphs together create *a right to explanation* when ADM is being used? Some authors argue that no such right exists based on the wording of the regulation.[58] Some writers, in turn, state that a systemic reading is necessary instead, in particular, when the articles 22 and 13-15 are read together with the recitals 71-72.[59] As Brkan summarizes, "*the basic dilemma that the overview of the literature reveals is the quest whether the so called 'right to explanation' would be a right that is read into another existing GDPR right, such as the right to information or access, or whether such a 'right to explanation' could potentially be created in addition to other existing rights from the binding provisions of the GDPR.*"[60]

I am not delving into the debate on "right to explanation" more deeply. However, the mere emergence of it is symptomatic. Transparency and "meaningful information" should constitute the general ethos of the regulation, and yet their formulations are so vague that there is uncertainty about the very existence of the right to explanation.[61] The deeper question is, therefore, whether the transparency formulations laid down in the GDPR are serving their purpose or not. Some of the confusion may stem from the fact that the regulation does not only

[58] Wachter, Mittelstadt and Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2) 7 *International Data Protection Law* (2017) 76.

[59] Malgieri and Comandé, 'Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation' (4) 7 *International Data Protection Law* (2017) 243; Goodman and Flaxman, 'European Union Regulation on Algorithmic Decision-Making and a "Right to Explanation"' (3) 38 *AI Magazine* (2017) 50; Selbst and Powles, 'Meaningful information and the right to explanation' (4) 7 *International Data Protection Law* (2017) 233; Edwards and Veale, 'Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'', (3) 16 *IEEE Security & Privacy* (2018) 46.

[60] Brkan 2019 at 111.

[61] Wachter, Mittelstadt and Russell, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR', (2) 31 *Harvard Journal of Law & Technology* (2018) 841: "Meaningful information about the logic involved" is said to require only "clarifying of the categories of data used to create a profile; the source of the data; and why this data is considered relevant" as opposed to a "detailed technical description about how an algorithm or machine learning works."

define the quantity of information but also the quality of it. This is particularly visible in the context of the right to explanation. Not only does information need to be at hand, it needs to be *meaningful*[62] and, in the light of article 12 "*concise, transparent, intelligible and easily accessible form, using clear and plain language".*

It seems that in the GDPR, transparency is regarded both as an umbrella concept, under which the access to information rights may be gathered, and an interpretative principle, which should inform all personal data processing, and the quality of provided information, which the access to information rights concretize.[63] Is "meaningful information", thus, an expression of the general principle of transparency (cf. articles 13-15: "- - to ensure transparent - - processing"), or is it ultimately something else? The entire question leads us back to the question of what kind of information we are after.[64]

B. *Transparency as Showing or Explaining its Object?*

To better understand the functioning mechanism of transparency in the GDPR and the black box problem, we need to return to the questions of *the logic of discovery* and *the logic of justification,* which were already briefly discussed. The distinction becomes important in assessing the way in which information can be transparent or meaningful. Namely, what is pursued through the call for transparency is often, in fact, some kind of conceivable message.[65] How do the operations in a black box affect my legal standing? Interestingly, the wording in articles 13-15 specifically mandates expressing the "meaningful information about logic involved" in ADM. Is it the logic of discovery or logic of justification, or a logic of some other kind? Do we need the black box to reveal or to justify itself?

Against the described backdrop, we need to analyse what the right to explanation – or whatever it is called – actually signifies. What does a data subject want or need to know?[66] To answer that, we need to approach the question from the data subject's point of view. Assumedly, she would not be primarily interested in the ADM *per se,* out of sheer human interest. Instead, she would probably be more interested in why it was applied to her, how the result of it was achieved, and how it affects her. In a similar vein, the WP29 guidelines on the use on ADM state that: *"The GDPR requires the controller to provide meaningful information about the logic involved, not necessarily a complex explanation of the algorithms used or disclosure of*

[62] Malgieri and Comandé discuss the question of "meaningfulness" of information from different angles. Malgieri and Comandé 2017 at 256-258.

[63] Also WP29 Guidelines on Transparency under Regulation 2016/679 (wp260rev.01) clarify the way in which transparency can and should be understood in the context of the GDPR., available at https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227 (last visited 28 May 2020)

[64] How these ideas have been conceived in different EU member states, see Malgieri, 'Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations', (5) 35 *Computer Law & Security Review* (2019) 1.

[65] WP29 Guidelines on Transparency under Regulation 2016/679 (wp260rev.01), at 5, para 3, also recognize this problem indirectly: "Where changes or additions are made to such information [provided to data subjects to fulfill transparency obligations], controllers should make it clear to data subjects that these changes have been effected in order to comply with the GDPR." Controllers should, "at a minimum," make this information available on their websites, but, "if the changes or additions are material or substantive, then … such changes should be actively brought to the attention of the data subject."

[66] Cf. Brkan 2019.

the full algorithm. The information provided should, however, be sufficiently comprehensive for the data subject to understand the reasons for the decision."[67] With the help of the provided information, the data subject could then assess whether she approves the automated decision or not, whether she wishes to contest it, and whether she wants to have a human intervention. In this form, we encounter the standard legitimating narrative of transparency: when we see, we can control, and possibly, change whatever is found unsatisfying.

As Bucher argues, when algorithms are conceptualized as black boxes, they are simultaneously understood as a problem of the unknown. This does not simply mean the lack of knowledge or information; rather, the black box points to a more specific type of unknown. The dominant discourses of transparency and accountability suggest that in fact, algorithms are *knowable known unknowns*. That is to say, they are knowable if the right resources are provided. This is further done, as the mainstream narrative goes, by opening the black boxes. [68]

However, it is possible to argue that in the context of ADM, the standard narrative of transparency is not necessarily entirely valid. For example, Ananny and Crawford argue that to "look inside the black box" may be too limited a demand. The metaphor is unsuitable, when we are talking about something as complex as algorithmic systems. It suggests falsely easy certainty, which would follow from looking, and ignores the ideological and material complexities involved in ADM. Furthermore, the promise of "seeing is understanding" may fail in the call for accountability; its object is hard to decipher and be held accountable. [69]

Similarly, Wachter, Mittelstatdt and Russel remain sceptical about the "look inside the black box" approach. However, instead of the danger of "easy certainty" and simplification, they think that opening the black box would lead to unnecessary complexity and leave the data subject confused about what is going on. They state that although interpretability is desirable, explanations can be, in principle, offered without opening the black box. In their view, less weight should be put on the data subject's ability to understand – whether through looking inside the box or being provided with an explanation – and more on how explanations are to empower data subjects to act towards their specific goals.[70]

These criticisms point in the same direction: seeing inside the black box does not necessarily lead to understanding, and understanding does not necessarily lead to control or other type of action.[71] Consequently, understanding has become a concern in its own right. The aforementioned debate on the right to explanation can be considered one aspect of this. Also conceptually, understanding has partially started to diverge from the vocabulary of transparency. Provided information needs to be *meaningful* and *transparent*, as it is

[67] WP29, *Guidelines on Automated individual decision-making,* at 25, available at https://iapp.org/media/pdf/resource_center/W29-auto-decision_profiling_02-2018.pdf (last visited 12 May 2020)

[68] Bucher 2018 at 43.

[69] Ananny and Crawford 2016 at 10.

[70] Wachter, Mittelstadt and Russel 2018 at 843: "We propose three aims for explanations to assist data subjects: (1) to inform and help the subject under-stand why a particular decision was reached, (2) to provide grounds to contest adverse decisions, and (3) to understand what could be changed to receive a desired result in the future, based on the current decision-making model."

[71] See also Macgregor, Daragh and Ng, 'International Human Rights Law as a Framework for Algorithmic Accountability', (2) 68 *The International and Comparative Law Quarterly* (2019) 309.

conceptualized in the GDPR. Additionally, other similar conceptualizations have lately emerged into the discourse: *explainability, interpretability, intelligibility, explicability, understandability* and *comprehensibility*. These concepts imply that transparency is not enough to guarantee understandability. [72]

Hence, more than transparency is needed, but that "more" can be malleable to many purposes.[73] This is readily explainable with the help of my theory, presented above. If transparency is conceived as a purely iconoclastic ideal which, by virtue of this, lets its object merely "shine through" transparency, it would not do much work. It would not be able to deliver and fulfil its promise. The revelation through the iconoclastic mechanism – removing the obstacles of visibility (cf. opening the box approach) – would not necessarily communicate anything to an average data subject. The revealed information would be highly esoteric, comprehensible only to experts. The underlying promise of transparency, "understanding is seeing", would be thus jeopardized. Indeed, in the context of ADM, seeing seldom constitutes immediate understanding to a layperson but instead leaves one puzzled by confusing information, perhaps comparable to a text in a foreign language. As Ananny and Crawford argue, transparency may privilege seeing over understanding.[74]

Thus, whether we talk about transparency or the right to explanation, or meaningful information about the logic involved, we need to consider the question of iconophily and the necessary human involvement it entails (cf. the right to human intervention). We can also assume that ADM is not naturally attuned to consider meaningfulness of information from the point of view of human comprehension.[75] This human intervention may shift transparency again towards understandability. This may be problematic, because ideologically, transparency specifically privileges immediate seeing over more mediated verbal explanations. If transparency becomes a synonym of explanation, it inevitably loses something from its legitimating power. The core promise of transparency "do not believe what I say, see for yourself" would thus be transformed to "do not believe what you see, let me explain instead".

However, as presented, the iconophily – transparency requiring constructs in order to create a visible appearance – together with the intentionality of transparency, enables considering human understanding and its limitations. On the one hand, it may produce information which is meaningful from an average data subject's point of view and creates legitimacy like that. On the other hand, however, it is potentially also a forum of impression management logic. The more human mediation there is, resulting in carefully managed visibilities, the more legitimacy may be produced. At the same time, this may also mean less "truth", when the intricacies of the black box cannot, by being exposed, necessarily communicate anything (the truth-legitimacy trade-off).

[72] See e.g. Olsen, Livingston Slosser, Hildebrandt and Wiesener, 'What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration', 162 *iCourts Working Paper Series* (2019)

[73] Cf. Buhmann, Paßmann and Fieseler, 'Managing Algorithmic Accountability: Balancing Reputational Concerns, Engagement Strategies, and the Potential of Rational Discourse' *Journal of Business Ethics* (2019) 1.

[74] Ananny and Crawford 2016 at 8-9.

[75] Cf. de Fine Licht and de Fine Licht, 'Artificial intelligence, transparency, and public decision-making' *AI & Soc* (2020)

5. Conclusions: Has ADM Broken the Promise of Transparency?

In this paper, I have discussed the ideal of transparency as a suggested solution to the black box problem in ADM. According to its promise, transparency would open or X-ray black boxes. This would enable data subjects to look at what is inside the boxes and perhaps question and change their inner workings. As I have explained, the main narrative of transparency has been adopted from the discourses of public law and governance into the discourses of ADM and algorithmic governance. Despite its well-institutionalized and seldom questioned promise, transparency is, I argued, more complex an ideal than the mainstream narratives acknowledge. I claimed that transparency is covertly a human-faced ideal, due to its basis in a visual metaphor, icono-ambivalence and the connection between intentionality and legitimacy.

As I have argued, even as an institutional value, transparency is underpinned by attempts of impression management and the avoidance of losing face, even if that face is that of an institution. I argue that the deep structure of transparency is ultimately control. Only controlled information release can create the promised and desired legitimacy. If that is not the case, the agent releasing the information could not influence the impression it gives and would thus be unable to govern its image (the truth-legitimacy trade-off of transparency). This functioning logic is further premised on transparency being an ocular-centric metaphor, and icono-ambivalent as a governance ideal.

This complexity of transparency is surreptitiously surfacing in the discourses and even regulation on ADM. As argued, transparency seems to be seen increasingly insufficient in addressing the core issues of the black box problem. Additionally, there are attempts to complement and/or to replace the concept of transparency with some other more fitting terminology such as a right to explanation, explicability or understandability, which would better consider the recipient of the information. On a theoretical level, the conceptual plurality does better work in differentiating the logic of discovery from the logic of justification.

Nonetheless, the term transparency still seems to carry a justificatory promise that other terms do not. This is visible e.g. in the vocabulary of the GDPR, in which transparency specifically is one of the key principles, trickling down to more concrete information release practices. Its history is longer than the other similar concepts, and it is closely linked to democracy and citizen participation. Transparency has the potential to empower to action – after all, it is a mechanism of control – because it assumes that everyone has the potential to understand by seeing and then taking necessary action. Understandability, in turn, has the potential to make people passive recipients of simplified information, being increasingly dependent on translating intermediaries. Additionally, the idea of immediate visibility inherent in transparency has emancipatory potential different from the mentioned neighbouring concepts. Explanation includes more human influence than sheer transparency, however illusory.

The extent to which the performative logic of transparency fuels the attempts to replace the term needs to be further analysed. How well is it recognized in those newer terms? It is important to notice that (human?) mediation is needed in the process of “translating” the inner workings of the black boxes into a form understandable to a layperson. Bucher explains how critics of the Enlightenment vision have been suspicious of the notion of revealing or decoding inner workings. The assumption is that a kernel of truth would just be waiting to be revealed by a mature a rational mind. In the Kantian tradition, she continues, the audacity to know is not only directly linked to rationalism but also to the quest for the condition under which true knowledge is possible. In this way, black boxes threaten the very possibility of knowing the truth.[76]

To what extent is this translation a matrix of impression management logic? How much is lost in translation, and how much should one anticipate those potential explanations serve the interest of the data controller? These questions require more work to be answered. Maybe black boxes even represent, in Elena Esposito's words, *divinatory rationality*. In pre-modern times, the mystery of the oracle was the guarantee of the rationality of the procedure. It was convincing and reliable precisely *because* humans lack the ability to understand the logic of the world, not despite that.[77] In a similar vein, in Socratic tradition the unknown was considered the prerequisite for wisdom, not a hindrance to it.[78]

An important feature of transparency's problems in the context of ADM stem from the fact that impression management logic cannot take place effortlessly. It requires assessing the effects of the release. How would they influence the desired impression? ADM lacks the sense of common decency and the understanding of when to interpret things to the letter and when more liberally. It lacks the human capability to steer through varying contexts with a compass such as the law or, indeed, transparency; in other words, it lacks practical wisdom. That feature would make it hard to create transparency by design, transparency which would not include this kind of *ex-post* evaluation (cf. meaningful information about the envisaged consequences in the GDPR).

In the end, the entire binary distinction between humans and machines may prove problematic. To the extent that transparency is seen as human-faced, it presupposes people who are concerned about their impression.[79] If transparency is seen as a tool for representation, whether in terms of sincere mimicking, impression management or full-fledged distortion, it still relies on the idea of the reality principle: that there is a ground truth to be represented, and that truth can be delivered and understood. What would it imply from the perspective of transparency's legitimating promise if human were removed from the equation? Are we left with governance, which no longer needs humans as its agents? Would such governance promise acceptability precisely because of the lack of ever so dubious and self-interested humans?

These questions are tricky for several reasons. Namely, the basic functioning logic of transparency may change if governance starts running through algorithmic modelling and deep machine learning. It may well soon be the case that algorithms, once created by human beings with human desires, become increasingly independent, and along with this, humans may lose their monopoly to control them. Algorithms can be independent to the extent that they themselves create new algorithms or even audit other algorithms.

[76] Bucher 2018 at 44.

[77] Esposito, 'Digital Prophesies and Web Intelligence' in Hildebrandt and de Vries (eds) 2013, 121-142, 129-132.

[78] Bucher 2018 at 44.

[79] Cf. Albu and Flyverbom 2019.

There is no reason to assume that algorithms in ADM would necessarily "think" like humans.[80] It would be hard to imagine that algorithms would desire other algorithms' approval, would want to be in contact with them and belong to the community of other algorithms. There is neither a reason to assume that they would want to be seen in a favourable light by other algorithms, to have high status in the algorithmic community, and avoid being shamed in front of other algorithms. Just by this little thought experiment, our own humanity, having a core of a social animal, becomes sufficiently clear. It is easy to see how transparency practices work through our human way of thinking and acting.

Maybe we should question the entire human-machine distinction and test what would happen to transparency.[81] As Ananny and Crawford state, "*We suggest here that rather than privileging a type of accountability that needs to look inside systems, that we instead hold systems accountable by looking across them—seeing them as sociotechnical systems that do not contain complexity but enact complexity by connecting to and intertwining with assemblages of humans and non-humans*."[82]

That said, it seems impossible to permanently eradicate black boxes in decision-making. Whether we are talking about hungry judges or algorithms which covertly privilege certain people over others, or even some kind of hybrid transcending the human-machine distinction, the complexity of decision-making cannot be reduced to simple steps of reasoning without something being lost. Following Bucher, mythologizing the inner workings of machines is not helpful. Neither should we think that algorithmic logics were somehow more hidden and black-boxed than the human mind, which is, as explained, a black box too.[83]

The best we can achieve, in the end, are descriptions of logics to justification. Logic of discovery may remain unfathomable to us, and may even be increasingly so, as machine learning models proliferate. This, in turn, may bifurcate the two realms of what happens in reality and what is conceivable to us. It seems that we want to both go beyond human understanding and to keep it as the guiding principle of ADM. In consequence, there may be less and less use for the ideal of transparency, or it will be reduced to its iconophilic aspect.

[80] Cf. Burrell 2016.

[81] Hansen, 'Numerical operations, transparency illusions and the datafication of governance', (2) 18 *European Journal of Social Theory* (2015) 203.

[82] Ananny and Crawford 2016, at 2.

[83] Bucher 2018, at 60.

With the support of the Erasmus+ Programme of the European Union

The European Commission supports the EUI through the European Union budget. This publication reflects the views only of the author(s), and the Commission cannot be held responsible for any use which may be made of the information contained therein.