The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called “right to explanation”

Emre Bayamlioglu

KU Leuven Centre for IT & IP Law (CITIP), Leuven, Belgium

Abstract

The right to contest automated decisions as provided by Article 22 of the General Data Protection Regulation (GDPR) is a due process provision with concrete transparency implications. Based on this, the paper in hand aims, *first*, to provide an interpretation of Art 22 and the right to contest (as the key provision in determining the contours of transparency in relation to automated decisions under the GDPR); *second*, to provide a systematic account of possible administrative, procedural, and technical mechanisms (transparency measures) that could be deployed for the purpose contesting automated decisions; and *third*, to examine the compatibility of these mechanisms with the GDPR. Following the introduction, *Part II* starts with an analysis of the newly enacted right to contest solely automated decisions as provided under Article 22 of the GDPR. This part identifies the right to contest in Article 22 as the core remedy, with inherent transparency requirements which are foundational for due process. Setting the right to contest as the backbone of protection against the adverse effects of solely automated decisions, *Part III* focuses on certain key points and provisions under the GDPR, which are described as the *1st layer* (human-intelligible) transparency. This part explores to what extent “information and access” rights (Articles 13, 14, and 15) could satisfy the transparency requirements for the purposes of contestation as explained in Part II. Next, *Part IV* briefly identifies the limits of *1st layer* transparency – explaining how technical complexity together with competition and integrity-related concerns render human-level transparency either infeasible or legally impossible. In what follows, *Part V* conceptualizes a *2nd layer* of transparency which consists of further administrative, procedural, and technical measures (i.e., *design choices facilitating interpretability, institutional oversight, and algorithmic scrutiny*). Finally, *Part VI* identifies four regulatory options, combining *1st and 2nd layer* transparency measures to implement Article 22. The primary aim of the paper is to provide a systematic interpretation of Article 22 and examine how “the right to contest solely automated decisions” could help give meaning to the overall transparency provisions of the GDPR. With a view to transcend the current debates about the existence of a so-called right to an explanation, the paper develops an interdisciplinary approach, focusing on the specific transparency implications of the “right to contest” as a remedy of procedural nature.

Keywords: algorithmic transparency, algorithmic regulation, automated decisions, GDPR.

[...]and there is some temptation to obey the computer. After all, if you follow the computer you are a little less responsible than if you made up your own mind. (Bateson 1987, p. 482)

1. Introduction and outline

Increasing automation has been an important topic of concern even at the earliest stage of the debates about the legal, political, and economic impact of data practices in the digital realm. It was clear by the early 1970s that resentment engendered by the systems such as *computerized billing* would soon spill over onto more delicate domains of life. Cautions were expressed that automated data processing would impair the system operators’ capacity to provide explanations about the results produced by the system and thus, contribute to the “dehumanizing” image of computerization.

Correspondence: Emre Bayamlioglu, Sint-Michielsstraat 6, Box 3443, 3000 Leuven, Belgium. Email: emre.bayamlioglu@kuleuven.be

Conflict of interest: The author declares that there exists no conflict interest and no relevant data to be made available.

Accepted for publication 6 February 2021.

Based on these concerns, the notion of transparency has long been regarded as a means to limit the risks and mitigate the harms arising from the opaque nature of data processing. Since the enactment of the Data Protection Directive (DPD) in 1996, the foundational idea underlying the EU data protection regime has been that the adverse effects of data processing may be best addressed by permitting individuals to learn about the data operations concerning them. Today, with the General Data Protection Regulation (GDPR), the European data protection regime may now be considered as the most extensive body of law aiming to regulate the activities involving personal data. It not only maintains well-defined individual rights fleshing out the principle of transparency but also accommodates various tools and mechanisms for the implementation and enforcement of these rights.

With data-driven practices based on machine learning (ML) being the primary *foci* of the data protection reform which resulted in the GDPR, one of the novelties of the Regulation is the enhanced transparency scheme provided for solely automated decisions – in particular, the introduction, the right to human intervention, and right to contest in Article 22.¹ Accordingly, the paper in hand deals with this specific type of transparency, namely “transparency” in the sense of *interpretability* for the purpose of contesting automated decisions. The aim is to determine to what extent the GDPR accommodates the practical implications of “right to contest” and the ensuing transparency requirements.

Taking right to contest as a due process provision, Part II starts with a systematic interpretation of Article 22, examining how the concepts of *contestability*, *obtaining human intervention*, and *expressing one's view* should be understood and interrelated. Rather than a prolongation of the initial provision (Article 15 of the DPD), the right to contest is regarded as the backbone provision with a key role in determining the scope of algorithmic transparency under the GDPR. To fully lay out the transparency implications of the right to contest, this Part also addresses the question: *what should be made transparent or known in order to render automated decisions interpretable and thus contestable on a normative basis?* (Bayamlioğlu 2018). The analysis inquires what interpreting the “algorithm” could mean for the purpose of contesting automated decisions – confirming that the transparency implications of right to contest are too complex to be dealt with merely by addressing certain opacities or invisibilities.

Overall, Part II lays out the theoretical basis of the paper approaching *data processing* and *automated decision-making* (ADM) as regulatory technologies, which enable a form of “algorithmic regulation” (Yeung 2018).² Such techno-regulatory approach allows for a conceptualization of ADM and the surrounding transparency debate as a *procedural*, put in other words, as a *due process problem*.³ Therefore, instead of handling automated decisions from the narrow lens of *discrimination*, *bias*, or *unfairness*, this paper regards ADM systems as “procedural mechanisms” which produce legally challengeable consequences. The concepts like fairness, equality, or nondiscrimination – as being mainly contextual and domain-dependent – can only address a fragment of the problem and thus, cannot serve as a theoretical basis for the intended analysis. Moreover, misuse of these quasi-legal concepts (to give meaning to the statistical results) runs the risk of *technical solutionism*, which will misinform policy-makers about the ease of incorporating transparency and accountability *desiderata* into ML-based systems (Cath 2018, p. 3, 4).

Having laid out the transparency implications of Article 22 as a general provision of due process, Part III, IV, and V inquires to what extent the GDPR can accommodate different conceptions of transparency inherent in the right to contest.⁴ The analysis is based on a twofold approach. That is, the “information and access” rights (Article 13–15) and the safeguards (Article 22) are treated as complementary but distinct sets of remedies (as *1st* and *2nd layer* transparency).⁵ This twofold methodology is guided by the understanding that recognizing distinct forms of opacity inherent to ADM systems is vital in developing (technical and nontechnical) solutions to address the risks arising due to the impenetrable nature of ML (Burrell 2016, p. 2).

In what follows, Part III focuses on certain key principles and provisions under the GDPR, which we describe as the *1st layer* (human-intelligible) transparency. It explores to what extent “information and access” rights in Articles 13, 14, and 15 of the GDPR could facilitate or improve the contestability of automated decisions as explained in Part II.

Part IV briefly identifies the limits of *1st layer* transparency, explaining how technical constraints together with the competition and integrity-related concerns (of the system developers and operators) render human-level transparency infeasible or legally impossible. Reflecting on both technical and economic limits, this Part offers an account of why the transparency requirements for contesting automated decisions could not be limited to access, notification, or explanation in the conventional sense.

Having seen the limits of directly human-intelligible models based on disclosure and openness in the previous Part, Part V inquires what further solutions the GDPR could accommodate in terms of implementing different conceptions of transparency aiming for contestability. As the *2nd layer transparency*, this part systemizes various regulatory instruments and techniques under a threefold structure: (i) the design choices facilitating interpretability; (ii) the procedural and administrative measures; and (iii) the software-based tools for algorithmic scrutiny.

Given that the problem lays with the framing of the optimum extent of transparency and the appropriate mode of implementation, Part VI offers regulatory options (implementation modalities) combining *1st and 2nd layer transparency* with a view to implement Article 22 without prejudice to the integrity of the systems or the legitimate interests of the stakeholders.

The final Part concludes that despite the normative, organizational, and technical affordances explained throughout the paper, between the right to contest as provided in the GDPR and its practical application, there are many gaps to be bridged to achieve the desired level of protection without hindering data-driven businesses and services. Accordingly, the conclusion points out the relevant research domains where further progress is required to construct a compliance scheme capable of balancing competing interests. Hence, the paper also serves as a conceptual framework for future research aiming to unravel sector or domain-specific barriers in relation to implementation of the right to contest,

With a view to transcend the current debates surrounding the so-called right to an explanation, the paper conceptualizes ADM as a regulatory technology and focuses on the specific transparency implications of the “right to contest” as a remedy of procedural nature. Building on the former writings of the author about the transparency implications of ADM (Bayamlioğlu 2018), the main contribution of the paper lies in this procedural perspective – which enables an interpretation of Article 22 as a due process provision – followed by a systematic analysis of the possible implementation tools and modalities under the GDPR.

2. Article 22 of the GDPR and the right to contest automated decisions

The principle laid out in Article 22, requiring that the automated data-driven assessments cannot be the sole basis of the decisions about the data subjects, is unique to the EU data protection regime. Such provision is not generally included among the US fair information practices or in the OECD guidelines preceding the 1996 DPD (Edwards & Veale 2017). Article 22 does not directly target personal data processing but a certain type of outcome, that is, the decisions that are fully automated and that substantially affect individuals.⁶

Since the enactment of the DPD in 1996, the practical application and proper implementation of Article 15 (the precursor to Article 22 of the GDPR) has not been of concern neither to the supervisory nor to the judicial authorities (Korff 2010). Although the provision was found intriguing and forward-looking, due to its complex nature – which makes individual enforcement difficult – it has been mostly overlooked and underused (Mendoza & Bygrave 2017). In practice, the compliance standards of the provision have remained at *de minimis* level, reducing compliance to a mere formality. According to Zarsky: a rule which is rarely applied (Zarsky 2017, p. 1016).

At first glance, Article 22 of the GDPR may be seen not to have brought much change in terms of wording. In this regard, the initial formulation of the provision in 1996 seems to have made it somehow future-proof. However, as will be explained below, with the newly introduced safeguards (the right to human intervention and contestation), the provision now has an essential role in determining the scope of transparency for solely automated decisions under the GDPR.

2.1. Decisions based solely on automated processing, with legal or similarly significant effects

The key provision of the GDPR on ADM, Article 22, applies to processes, which are fully automated, and which bring about legal or similarly significant effects for the data subject. Automated decisions, which fail to comply with the definition provided in Article 22(1), shall not be bound with the provision.

The application of Article 22 initially requires the existence of a “decision,” though neither the former DPD nor the GDPR provides any guidance as to what amounts to a decision. Bygrave suggests that the term “decision”

should include similar concepts such as *plans, suggestions, proposals, advice, or mapping of options*, which somehow have an effect on its maker such that she/he is likely to act upon it (Bygrave 2001).

Article 22(1) further requires that the decision must also be fully automated, allegedly involving no human engagement. Because the level of human intervention to render the decision not fully automated is not clarified, many data controllers interpret the provision narrowly. As a result, significant amount of data-driven practices may be kept out of the reach of EU data protection regime simply by the nominal involvement of a human in the decision-making process.⁷ This requirement which also existed in the DPD has been widely used as a loophole by the data controllers to derogate from the provisions on automated decisions. This has been despite the preparatory work of the DPD, which explicitly stated that one of the rationales behind Article 15 of the DPD was that human decision-makers might attach too much weight to the seemingly objective and incontrovertible character of sophisticated decision-making software – abdicating their own responsibilities.⁸

The scope of Article 22 is limited to the decisions that produce legal or similarly significant effects. Legal effects may be described as all qualifications established by a legal norm either in the form of obligations, permissions, rights, powers; or in relation to one's status such as citizen, parent, spouse, debtor; or relating to categories of things (e.g. moveable, negotiable instrument, public domain). The inclusion of the term *similarly significant effects*, expands the scope of the provision to cover certain adverse decisions even if the outcome does not straightforwardly affect the data subjects' legal status or rights.

Regarding the implementation of Article 22 in the EU, some member states have adopted a wider approach, such as Hungary, which includes all automated decisions prejudicial to the data subject or France, where the specific legislation covers ADM producing any significant effect (Malgieri 2019).

2.2. Derogations: Consent, contractual necessity, and mandatory laws

As Article 22(1) grants data subjects the right not to be subject to solely ADM, the provision also contains certain exceptions (derogations) to the rule – subject to Art. 22/4 on special categories of personal data.

One of the most important changes brought by the GDPR as compared to Article 15 of the DPD is the introduction of “*explicit consent*” in Article 22(2)(c) as one of the grounds, which may be relied upon by the controllers to carry out fully automated decisions. According to Mendoza and Bygrave (2017), the introduction of consent comes as an impairment to the essence of the provision, lowering the *de facto* level of protection (p. 96). Considering that *consent may practically be used to deprive the data subjects of the control of their data*, the concerns about this new derogation – as a swift mechanism to carry on with automated decisions – are not all without merit. Nonetheless, rather than serving as a backdoor to circumvent data protection rules, consent may equally be construed as a leverage for transparency (Kaminski 2019). This is particularly the case where explicit and informed consent is taken as the initial step of the safeguards to render automated decisions contestable under Article 22(3). Furthermore, consent does not relieve the data controller from the duty of compliance with the general data protection principles such as fairness and proportionality provided in Article 6. Taking into account the complexity and subtlety of the current ADM systems, the requirement of explicit consent inevitably entails some “*explanation*” to allow data subjects to make informed choices.⁹ The extent of communication necessary to render the data subject's consent explicit and thus valid, may also be taken as a benchmark to determine the minimum content of the notifications under Articles 13 and 14. Consent implemented as a leverage to individualized transparency – but not as a *carte blanche* for ADM without encumbrance – could play a critical role in reinforcing the right to obtain human intervention and right to contest.

Formation and performance of a contract (contractual necessity) is another derogation provided in Article 22(2)(a). The prohibition on automated decisions shall not be applicable where an automated decision is necessary for entering into or for the performance of a contract between the data subject and the data controller. The derogation based on contractual necessity provides a broad field of play which is tempting for abuse and *creative compliance*. The extent of automated decisions, which would be necessary in a contractual context is an issue that requires the consideration of the mutual benefits and expectations of the parties. For instance, increasing the efficiency of the system – as a general argument – cannot be regarded as necessity for this is simply what makes data processing more invasive (Guinchard 2017, p. 12).

Article 22/(2)(b) lays out another derogation providing that data subjects may be deprived of the safeguards in Article 22(3) where processing is mandated by the Union or Member State Law. Despite the reference to *suitable measures to safeguard the data subject's rights and freedoms and legitimate interests*, this exclusion is likely to create discrepancies in terms of contestability standards between the administrative decisions based on public law and ADM relying upon consent or contractual necessity (Massé & Lemoine 2019, p. 9). Confirming this, some member states have already implemented the derogation in a way similar to a blanket-exemption, permitting ADM as default practice for public institutions (Malgieri 2019).

2.3. Safeguards against automated decisions and the right to contest

2.3.1. Safeguards in Article 22(3) in general

Under Article 22(3), where the exemptions based on contractual necessity (22 (2)(a)) or consent (22 (2)(c)) take effect, the data controller is obliged to implement measures to safeguard the data subject's rights, freedoms, and legitimate interests. In principle, these measures should at minimum contain a fair amount of human intervention so that the data subjects may express their view and effectively contest automated decisions. Before the GDPR, the DPD only spoke of arrangements allowing the data subjects to put forward their point of view. The Regulation has improved this position by formulating safeguards providing *for human intervention and contestation*.

Article 22(3) reads as:

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.

Although frequently commented and explored in the scholarly writing, there seems not to be much attention to a coherent and systematic interpretation of the provision, in particular how the right to obtain *human intervention*, *expressing one's views*, and *contesting* the decision could practically be implemented. They are usually treated as if rights or remedies of equal footing (alternatives to each other), without clarity whether they are complementary, gradual, or distinct rights, or they should be treated as a unity.¹⁰ Very few seem to pay attention to the inevitable difficulties arising from the provision's open-ended and clumsy syntax – dressed in dense and contorted “legalese” – muddying its interpretation (Yeung & Bygrave 2020, p. 10). Wachter and others rightly point out that whether these remedies are “interpreted as a unit that must be invoked together, or as individual rights that can be invoked separately, or in any possible combination” would determine how a decision could be contested. After acknowledging the possibility of several interpretations, they conclude that treating each Article 22 safeguard as “individually enforceable” would be the most sensible option (Wachter *et al.* 2017).

More importantly, Wachter *et al.* (2017) also argue that – depending on the costs and the “likelihood of success” – expressing views or obtaining human intervention should not necessarily be followed by further legal means to challenge the decision. However, a systematic and teleological reading of the provision reveals that the right to contest is the backbone of the safeguards provided under the GDPR and cannot be ignored on the mere grounds of low likelihood of success as Wachter *et al.* suggest. The wording makes it clear that the right to obtain human intervention is the minimum of the remedies that data controllers are obliged to implement to satisfy the ultimate aim of the provision. This necessarily implies that human intervention may not always be the best option to address the adverse effects of automated decisions. As Part V will explain, there may be further options (other than or in addition to human intervention), which provide technical and procedural means to challenge the ML outcome. Specific inclusion of the “right to express one's point of view” confirms that data subjects not only have a right of appeal to obtain a new decision but they can also provide information that might be relevant for reconsidering the initial result (Malgieri 2019, p. 11). There may also be dignitary concerns or possible mutual benefits in allowing the data subjects to express their views even if they are not willing to challenge the decision. For instance, through an online forum, data subjects can provide feedback and voice their complaints about the extenuating circumstances that they believe the system fails to consider.

Due to its procedural character, Article 22/3 is inevitably silent on the substantial grounds which could be relied upon to challenge the reasoning or the criteria underlying the automated decisions (*see* Part 2.4). That is,

whether or when certain ML outcome could be regarded as unfair or unlawful is a conclusion, which requires resorting to normative propositions provided in the relevant legal domains, for example, labor law, consumer law, insurance law (Macmillan 2018, p. 51). Data protection law is not where we could seek for standards or rules that can be used to test the legality or the justifiability of the undesirable outcomes of ADM (Hoffmann-Riem 2020, p. 14).¹¹

2.3.2. *The right to contest*

The newly adopted wording of the GDPR using the term “contest” connotes more than mere opposing. It rather points at a “right of recourse” as a version of algorithmic due process (Kaminski 2019a) or, at least, an obligation to hear the merits of the appeal and to provide a justification for the decision. Although the essence of the provision has not changed with the introduction of “human intervention,” the inclusion of an appeal (contestation) process against automated decisions stretches the legal boundaries of Article 22 to the broadest extent possible. It obliges the data controller either to render automated decisions contestable or to cease ADM at all. What is required by Article 22(3) is not about informing or disclosing but rendering the decision contestable at least against a human arbiter. In principle, the data controller has to “explain” the decision in such a way that enables the data subject to assess whether the reasons that led to a particular outcome were legitimate and lawful (Goodman & Flaxman). Confirming this, German law – long before the GDPR – has taken the view that the data subject needs to have an understanding of a specific decision in order to evaluate and be persuaded about its accuracy and legal compatibility (Korff 2010, p. 83).

Borrowing from the Contract Law of civilian (continental) legal systems, the distinction between the access and information rights (Art. 13–15 GDPR), and the due process rights in Article 22(3) may be seen similar to that of “obligations of conduct” and “obligations of result,” respectively. An *obligation of conduct* requires the dedication of an adequate amount of resources or the use of reasonable endeavors for a certain end, without any guarantees as to the outcome (a.k.a. input-based obligation) (Economides 2010). From this perspective, duties pertinent to access rights may be found more akin to an obligation of conduct since “informing,” “explaining,” or “disclosing” are acts which refer to a certain behavior, rather than assuring a specific outcome. The right to contest under Article 22, on the other hand, is similar to an *obligation of result*. It mandates a mechanism that will enable data subjects to have their objections heard and decided. This theoretical distinction helps draw a coherent picture of the contestation scheme provided by the GDPR for automated decisions. That is, contestability is a goal-oriented concept, unless the desired outcome is achieved, what is disclosed or communicated is irrelevant. In line with this, Kroll rejects the view that computer system behaviors should be held to a “best-effort” standard because they are unforeseeable. He asserts that treating these systems as though they are uncontrollable would ignore the fact that they are human artifacts – built to a purpose by some human agency that must be accountable for their behaviors (Kroll 2018, p. 7).

Despite this central role of the right to contest in Article 22(3) – which somehow presupposes a form of explanation of the decision – some scholars, nevertheless, argue that the GDPR does not contain an *ex-post* “explanation right” for individual automated decisions (Wachter *et al.* 2017). To support their argument, the proponents point out the discrepancy between the wordings of Article 22(3) and Recital 71. That is, while the latter specifically mentions obtaining an “explanation of the decision,” the former does not. Accordingly, it is contended that the GDPR provided no right of explanation for individual automated decisions (Wachter *et al.* 2017). Leaving aside the objections raised to this rigid and contained interpretation, the approach taken in this paper renders such argument and the surrounding debate mostly irrelevant. It is inherent in the formulation of the right to obtain *human intervention* and *contestation* that Article 22(3) requires much more than a mere explanation of the decision. As will be explained in Part III below, access and information rights relating to automated decisions can only be effectively enforced if they contribute to the due process rights provided in Article 22. The omission of the phrase “an explanation of the decision” from Article 22(3) may be understood as the intention of the legislature to keep an open view as to the safeguards and their possible implementation. Considering that the *travaux préparatoires* of the Regulation do not provide much clarity about the rationale of the provisions relating to automated decisions, the legislature seems to acknowledge that such novel legal obligation, expressed in abstract terms, could not be properly contained or effectively addressed by the concept of “explanation.” Therefore, the provision may be regarded as giving leeway for different methods of implementing safeguards. Article 22 of the

GDPR not only subsumes the much-discussed “right to explanation” but also allows for different regulatory options or modalities to ultimately provide the data subjects with the means to contest automated decisions. As Selbst and Barocas (2017) put it: “...focusing on explanation as an end in itself, rather than a means to a particular end, critics risk demanding the wrong thing” (p. 1).

These above-explained controversies and ambiguities are also reflected in the implementation of Article 22(3) within the EU. Member states formulate provisions on ADM in various ways without much coherence, and few expressly refer to explanation. French Law, as providing the most generous framework, permits data subjects to obtain information about the rules of the data processing and the features relating to the practical implementation of the algorithm together with the source code. Hungarian law, with a rather innovative formulation, requires information *about the methods and criteria used in the decision-making mechanism* (Malgieri 2019).

2.4. The transparency implications of the right to contest (contestability requirements)

Having explored the possible interpretation of the right to contest as provided in Article 22, this Part – building on the author’s earlier work – provides a brief account of the possible transparency(contestability) requirements to effectively challenge automated decisions. The below analysis reflects on a dynamic and instrumental conception of transparency, which do not aim to analyze the system by the semantic route of explanation but rather by defining requirements enabling scrutiny on a normative basis (Bayamlioğlu 2018; Rader *et al.* 2018). Put in other words, contesting ML-based decisions is not about reading off the computer code but rather relates to the question of how these systems make up the regulatory realm we are subjected to.

Since machines are built for a purpose, they are expected to exhibit certain behaviors associated with their function (De Ridder 2006). That is, every decision-making system contains some inherent “normativity”¹² as the system’s output is directed to achieve some preset goals or to serve certain ends (Castañeda 1970; Krist 2006; Binns 2017). As mentioned above, ADM systems may also be seen as techno-regulatory assemblages, which select and reinforce certain values at the expense of others (Bayamlioğlu & Leenes 2018; Eyert *et al.* 2021). Accordingly, challenging an automated decision initially requires a conceptualization of the outcome as a process where certain input leads to certain results. This may be seen akin to the decisions in a legal system, based on “facts,” “norms,” and the ensuing “legal effects.” Such conceptualization aligns with the approach, which portrays regulation as a cybernetic process involving three core components that form the basis of a control system, that is, ways of gathering information, setting standards, and ways of changing behavior (Yeung 2015). In the context of automated decisions, this would imply how and why a person is classified/profiled in a certain way, and what consequences would follow from that classification. Such modeling, which maps input/data with the effects within a contemplated normativity, provides us with a rule-based reconstruction of the decision-making process. (Bayamlioğlu 2018). The contestation of a decision relates both to the interpretation of the input and the normative basis relied upon to reach that decision (Hoepman 2018). How transparency is effected will depend on what it is intended to accomplish. The concrete transparency requirements of such a model entail numerous types of differently purposed, but complementary, information flows along multiple axes (Malgieri & Comandé 2017; Kaminski 2019a). In the below paragraphs, we briefly provide the essentials of a model, which aims to open up and systemize what interpreting the “algorithm” could mean for the purpose of contesting automated decisions.¹³

For computers to solve a problem, it is necessary that a computation manipulates a representation of the world, and the meaning of a computation depends on the meaning of the representation it transforms. Therefore, what ML systems do may be regarded as the creation of *internal models* of the pertinent environments (Eyert *et al.* 2021). Based on this, as the *initial* step of contestation, we need the knowledge of *what the system learns about persons, places or events, and how people are represented as inputs to the algorithm*. In a ML process, data instances exist as values of feature variables where each feature (e.g. age, height, weight) is an individually measurable dimension of the problem in question. Determining which data features to consider is a part of the regulatory process as it reduces the complexity of the environment to a specific segment of “reality” (Eyert *et al.* 2021). As the necessary interface between the underlying attributes and the decisions that depend on them, data features are open normative challenges. Accordingly, decisions may be contested based on the selection of the relevant data features and the ensuing inferences that are relied upon. Or as Jasanoff and Simmet (2017) put it: “...the choice of which realities one takes as consequential and therefore which facts one sees as important or

controlling, is normative” (p. 752).¹⁴ Having said that, it should be noted, ML-based systems are less and less programmed with a predefined feature space. Deep learning techniques using neural networks can define features autonomously by analyzing the data directly coming from the input layer. This severely impedes the capacity to scrutinize the factual or inferential basis of any decision or outcome.

When ML tools are used to make decisions, it is possible to contemplate a “decision rule” such as: *do this if the estimated probability of “z” is larger than “x” or smaller than “y,” and so on* (Baer 2019, p. 88). Thus, the *second* type of insights required for contestation is the decision rules (normative basis of the decision), which describe how certain ML findings are translated into concrete results in a wider decision-making context. For instance, speech analysis – which can detect one’s dialect or accent – may be used as “factual” input for the selection of the suitable content in political microtargeting. Accordingly, Spanish voters identified as having a Catalan accent may be delivered political messages supporting Catalunya’s secession from Spain. In this example, the delivery of the relevant content is the result of applying “decision rules” to the outcome of the classifier. Decision rules (normative choices) are shaped by the hypotheses and assumptions about the root cause of the targeted problem. They are the formalizations of the general goals (objectives) of the system such as winning elections, avoiding customer churn, or better distribution of insurance risks. ML-based systems allow for various combinations of general goals and the ensuing decisional rules “nested inside one another” (Eyert *et al.* 2021). Since this normativity does not necessarily rely upon legal or moral grounds but has a computational and data-driven basis, the normative orientations of the system are not always as straightforward as the relation between having Catalan accent and supporting independence. It is often the case in ML practices that the objectives and the underlying assumptions may not be deterministically configured but rather adaptively and dynamically adjusted (Yeung 2018).

Third, the “impact” of the decision could also be an essential ground for contestation. Speaking of the impact of the decision, a simple credit score does not only determine whether one would get a loan, but it may also, fully or partially, determine the loan pricing, type of loan monitoring, the amount of credit, and how the credit risk would be managed. As such, for the purposes of contestation, the “context” determines the actual consequences (impact) of the decision. A decision may be “good” in a particular context but less “good” in others (Zeng *et al.* 2018). In case of contextual uncertainty, there could be several explanations which may seem equally plausible (Gollnick 2018).

Fourth, information about the “accountable actors” behind the ADM process is also an essential piece of information necessary for effective contestation. Although, this may seem like a procedural requirement – not primarily of relevance to interpreting the decision in substance – it is nevertheless vital to fully understand the context and the purposes underlying the decision-making process. In case of automated decisions, there may be several legal grounds of contestation relating to different actors (e.g. errors in data collection, flaws in analysis, illegitimacy of the purposes, or the ensuing decision rules). As such, challenging an automated decision not only requires the disclosure of the entities involved but also the organizational and contractual set-up under which these entities operate and conduct business. The GDPR contains various provisions, which may accommodate disclosures reaching beyond the conventional actors of “data controller” and “data possessor.” The reference made to “the recipients or categories of recipient of personal data” in Articles 13, 14, and 15, and further definitions provided for “representatives,” “group of undertakings,” and “enterprises” are clear indications that the GDPR was drafted in consideration of the complex network of actors behind the current data-driven practices.

3. A general overview of “Access and Information Rights” (1st layer transparency)

This part will explore how the individual information and access rights of the GDPR (*1st layer* of transparency) could substantially contribute to the objectives of Article 22. Put in other words, the below analysis explores: to what extent the relevant provisions in Articles 13 to 15 could be interpreted in the direction of “contestability” as defined above.

3.1. The intended purposes and the legal basis of processing

Articles 13(1)(c) and 14(1)(c) of the GDPR provide that data subjects will be given information about the purposes of the processing for which the personal data are intended together with the legal basis for processing. As a

reflection of the principle of purpose limitation (purpose specification) in Article 5(1)(b), information about the “intended purposes” is a key element, which helps reveal the business strategy and the objectives pursued by the data controller as well as the other related parties. Such information enables the “reverse-engineering” of the decision-making process with a view to understand the underlying normative setup. The purposes pursued by the system are also of direct relevance to the context of the decision.

In case the data controller relies on Article 6(1)(f), which lays the general grounds for lawful processing, the obligation to notify the data subject of the intended purposes is further reinforced by the requirement to provide *information about the legitimate interests pursued by the controller* (Articles 13(1)(d) and 14(2)(b)). In addition, the reference made to the *legal basis of processing* (statutory or contractual) in Articles 13(1)(c) and 14(1)(c) makes it clear that the GDPR envisages a link between the *intended purposes*, *pursued interests*, and the *legal basis* of data processing. These altogether may be implemented as effective transparency mandates against the data controllers that carry out solely automated decisions. In this respect, Hildebrandt draws attention that the purpose of processing that is to be defined by the data controller is not the same as the tasks to be defined by the system designers to achieve that purpose. While the former is related to the commercial, institutional, political, or moral aims of those who deploy the system, the latter deals with the objectives and/or targets that the learning algorithms have been programmed to follow. Purpose limitation does not primarily aim for *the methodological integrity of data science*, but it is rather a specific reflection of the principles of legality and due process. As such, “it relates to the justification of such decision-making rather than its explanation in the sense of its heuristics.” (Hildebrandt 2019, p. 113).

The principle of “purpose limitation and specification” also plays a key role in determining the extent of liability with regard to relevant stakeholders. In various cases, the CJEU has held that those who exert influence over the processing of personal data and participate in the determination of the purposes of that processing, may be regarded as a *controller* within the meaning of the EU data protection regime.¹⁵ Accordingly, in *Fashion ID* case,¹⁶ it was made clear by the Advocate General that the power to decide and specify for which purposes the data will be processed is a crucial factor in the apportionment liability among the involved parties.

Since the enactment of the DPD, data controllers have had trouble in deciding how to adequately specify the purpose in a certain data processing operation. So far, many data controllers have chosen to phrase their purposes as vague and abstract as possible. This is both to have the maximum leeway for further use of the data and also to avoid the disclosure of any commercially valuable information regarding their data operations. Taking that into account, in its 2018 *Guidelines on Automated individual decision-making and Profiling*¹⁷ (hereafter *WP29 Guidelines on automated decisions*), Working Party 29 (WP29) made it clear that the purposes defined such as “improving users’ experience,” “IT-security,” or “future research” would not suffice in the absence of further clarification. For instance, processing of data for online advertising may not be compatible if the initial notification only contained a mere reference to “marketing purposes.”

In terms of exercising the right to contest, *purpose limitation*, *compatible use*, and *notification of the intended purposes* are important leverages, the implementation of which are dependent on certain enacted and applied transparency requirements.

3.2. “Meaningful information about the logic involved” and “the envisaged consequences”

Articles 13(2)(f), and 14(2)(g) GDPR provide that the controller shall inform the data subjects of: (i) the existence of ADM as defined in Article 22; (ii) meaningful information about the logic involved; and (iii) the significance and the envisaged consequences of the decisions. As seen, these provisions directly correspond to the essential constituents of the contestability requirements defined above, namely the “facts” (data input in the form of features) that are relied upon and the *decision rules* informing us about the goals pursued by the system.

As an initial step, the relevant provisions in Articles 13 and 14 require that the data controller provides information about the existence of a decision based on solely automated processes as defined in Article 22(1). Next, the data controller is obliged to provide *meaningful information about the logic involved* in the processing together with the *significance and the envisaged consequences* of the decision. Although emerging big data practices have been one of the major driving forces behind the GDPR, this crucial provision has remained similar to its DPD counterpart, which was mostly neglected or underused during the lifetime of the Directive. In more than

20 years that the DPD had been in force, the scope, requirements, and the possible limitations regarding the right of access as applied to automated decisions was not tested before the European courts. There is thus hardly any practical guidance on the interpretation of this enduring provision. So far, the obligation to provide information about the logic of the ADM has had varying implementations in the EU member states (Korff 2010, p. 85).

As the GDPR adds the term “meaningful” to the original provision in the DPD, it is now generally accepted that the controller should convey information about the rationale and the criteria relied upon in reaching the decision. The quality of being “meaningful” must be evaluated from the perspective of the data subject, treating accessibility and comprehensibility as the primary components. In parallel with the contestability requirements in Part II, “meaningful information” may also be understood as a functional description, which connects the decisional cues (data as input) with the consequences in a normative contemplation. Selbst and Barocas (2017) assert that “[t]he GDPR’s demand for meaningful information requires either that systems be designed so that the algorithm is simple enough to understand, or can provide enough functional information about the logic of the system that it can be tested” (p. 31). Lipton (2016), drawing attention to Article 22, more elaborately states that the information to be conveyed must “(i) present clear reasoning based on falsifiable propositions and (ii) offer some natural way of contesting these propositions and modifying the decisions appropriately if they are falsified” (p. 4).

Further explicit reference in the provision to the “significance” and the “envisaged consequences” of processing resonates with the above-defined contestability requirement about the impact of the decision (Part 2.4). For the purposes of contestation, it is essential to fully understand the concrete results and the risks emanating from the contextual use of the data. For instance, in credit scoring, envisaged consequences may include whether the result of the analysis will be used for subsequent evaluations, the period during which the evaluation will be held valid or the third-parties who might have access to the results. The information about the “envisaged consequences” should elicit the real-life impact of the automated decisions to enable the data subjects to oversee the process and evaluate the consequences. The envisaged consequences should be assessed in tandem with the *intended purposes* of data processing.

In line with above, WP29 *Guidelines on automated decisions* clarify that “[t]he data subject will only be able to challenge a decision or express their view if they fully understand how it has been made and on what basis.” (p. 27). The Guidelines recommend (to data controllers carrying out automated decisions) to provide information about “why a certain profile is relevant to the automated decision-making process and how the profile is used for a decision concerning the data subject” (Annex 1).

4. Limits of and impediments to human-interpretable models

The analysis thus far reveals that the GDPR provides several individual rights, which accommodate a fair amount of information to facilitate the exercise of the right to contest at a human-intelligible level (*1st layer transparency*). That being said, satisfying contestability requirements through openness, disclosure, and notification is neither desirable nor necessarily feasible for the purposes of challenging automated decisions (Desai & Kroll 2017, p. 39).

This Part further systemizes and examines where and why the *1st layer transparency* (information and access rights) fails – necessitating that the *2nd layer transparency* (Part V) come into play. The main impediments to *1st layer transparency* are: (i) the technical (complexity-related) intransparencies; and (ii) the secrecy demands of the businesses and institutions primarily arising from the integrity or competition-related concerns (Also see Mantelero 2019, p. 11).

4.1. The technical limits: Computational complexity and unpredictability

Human-level transparency (*1st layer*) for the purpose of contestation will mean that given enough time and resources, the computational processes producing the result should be human intelligible for the purposes of challenging the decision (Hildebrandt 2018). However, in practice, computational complexity gives rise to inscrutable models as a technical matter.¹⁸ Exceedingly complex models are often very difficult or even impossible for humans to parse.

First, even in simple models, the rules that govern decision-making process may be so numerous and interdependent that they defy practical inspection and thus, resist comprehension (Selbst & Barocas 2018, p. 1094). Generally, the more factors are incorporated into the model as input, the more rules will be required to explain all possibly valid relations between the input and the output. In consequence, the system may end up with too many predictors, each having a weak relationship to the result.

A *second* type of opaqueness arises from the fact that the value of ML lies largely in its capacity to find patterns that go beyond human intuition. This results in ML models, which make it impossible to weave a sensible story to account for the statistical relationships that seem to weigh in. The assumed causality that the decision relies upon may be obscure and thus, may defy our intuitive expectations about the relevance of the criteria. Correlative relationships can be sufficiently complex and nonintuitive especially when dealing with human behavior. For instance, ML tools deployed for recruitment can decide about the best prospective employees according to the applicant's place of birth, music taste or, peculiarly, whether the applicant has any numerical characters in her social media account name. When the features used do not bear a comprehensible relationship to the outcome, the model will resist an assessment whether the decision is reliable – both as a matter of validity and as a normative matter (Selbst & Barocas 2018, pp. 1098, 1129).

Third, adaptive and dynamic data-driven systems are capable of modifying their responses according to the changes in the environment. Accordingly, in adaptive decision-making systems, the decision rule is no longer predetermined, but constantly adjusted (Yeung 2018). An adaptive or nondeterministic algorithm may produce different results for each instance of its execution (each case it handles). Therefore, while complexity can be seen as a barrier to overall understanding, adaptive algorithms seriously impair the capacity to predict the results for a particular set of input (Felten 2017).

4.2. Business-related barriers

Many scholars draw attention that it is not “the black box issue” which makes the production of knowledge about ML-based systems (or AI in general) a difficult task for regulators. Even in cases where the system in question is simple enough to allow for a proper explanation of the decision, it is rather the legitimate business interests or other institutional concerns, which make individual access to relevant information a delicate balancing act (Wischmeyer 2020, p. 79). The disclosures made for the purposes of contestability may reveal information jeopardizing the integrity of the system or may impair the competitive advantages of the system operator/designer. Thus, it goes without saying that those who deploy ADM systems have a strong interest in the deliberate establishment and maintenance of opacity.¹⁹

4.2.1. System's integrity

Concealment, nondisclosure, and controlled access are the strategies that data controllers may resort to protect the integrity of their systems by preventing users from gaming or circumventing the decision-making process. Individuals who manipulate the inputs of the system (based on their intimate knowledge of the system's behavior) not only gain advantage for themselves but also impair the predictive capacity of the system. Gaming may be seen as the rational behavior of the users where the cost of manipulating the input is lower than the expected benefits or the eliminated risks.

Gaming of the system – also referred to as adversarial learning in ML context – may involve strategies in the form of *avoidance*, *altered conduct*, *altered input*, and *obfuscation* (Bambauer & Zarsky 2018). Depending on the context and the values that a system designer/operator wants to prioritize, each type of gaming or “gameability” may affect the individual and the society differently, and not necessarily negatively. In cases of *altered conduct*, where the individual changes his/her course of action to avoid adverse effects, the end-result may simply amount to lawful or desired behavior – accomplishing the very objective of the ADM system. This may be seen as a form “nudging,” which alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives (Thaler & Sunstein 2008, p. 6).

The gaming behavior may also take the form of *altered input*, which aims to improve some proxy features without actually improving the underlying attributes that the system aims to reinforce. For instance, while a loan applicant may choose to pay her bills on time to increase her credit score, she can also invest in efforts to discover the proxy features and heuristics that she could manipulate to present herself as if she was creditworthy.

(Kleinberg & Raghavan 2018).²⁰ This is especially the case for ADM systems, which target unobservable or hard-to-measure characteristics and therefore need to use *proxy features* which are assumed to be representative of the actual attributes.

As seen, the policy implications of gaming and its countermoves are a “mixed bag” in that ADM systems may incentivize both productive and unproductive forms of effort. As such, the diversity of the concept resists any overarching theory about how the gaming behavior must be weighed against competing values. Hence, there is no uniform or straightforward justification for secrecy practices aiming to prevent gaming (Bambauer & Zarsky 2018, pp. 22, 33).

Deploying more complex models, making frequent changes in the parameters of the system, and using differently sourced proxies are the strategies employed to make gaming more difficult or less rewarding. Predictably, the systems, which use more immutable characteristics or observe rather nonvolitional behavior are more resistant to gaming. From the legal perspective, data controllers’ integrity claims may generally be based on the right to conduct business since the system’s accuracy and efficiency diminish due to the “false” data fed by the gaming behavior. In many cases, gaming behavior amounts to a tortious act or a breach of contract. Where appropriate, gaming may also be opposed and counteracted on the basis of public health, privacy, or security risks.

4.2.2. Economic rivalry

Integrity claims are usually conflated with competition-related arguments. Industry players may be reluctant to disclose the coding of the system or the training data, or they may simply refuse to provide an explanation of the ML model as this may weaken their competitive advantage. Intellectual property (IP) rights, and in particular trade secrets,²¹ is the main legal framework that businesses rely on to prevent competitors from gaining access to commercially valuable information.

Many informational elements in a ML process such as individual data, databases, algorithms, profiles, data features, or ML models fully or partially fall within the ambit of IP protection. However, speaking of disclosure of and access to ML systems (*1st layer* transparency), only trade secrets may fully be relied upon for the purpose of secrecy. Other types of IP protection do not in principle provide secrecy on the content but focus on the reproduction, dissemination, adaptation, or other specific uses of the protected subject-matter. Hence, copyright (including software protection²²), *sui generis* database right or patent rights could be relevant in case of *2nd layer* transparency measures (Part V). The procedural, technical, or administrative mechanisms deployed to scrutinize ADM systems or to contest specific decisions may require the copying, reverse engineering, or otherwise modification of the ML elements.²³

Not much substantial work exists to offer guidance about how different transparency or contestability needs could be reconciled with data controllers’ and other industry players’ legitimate interests. The issue is also poorly addressed in the GDPR. In connection with the technologies enabling data subjects’ remote access to their personal data, Recital 63 of the Regulation reads: “[t]hat right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software.” Although special reference to software may suggest that the statement is confined to remote access systems, an interpretation of the Recital to include Article 22 would contradict neither with the spirit of the GDPR nor with the general principles of law. Having said that, there is also no compelling reason to read too much into this single reference. As with other fundamental rights, the right to property should also be respected in the application of the EU *acquis communautaire* notwithstanding whether there exists a specific reference in the GDPR. Moreover, the wording “adversely affect” in Recital 63 may be seen as no more than a mere reminder because the Recital also assures that “the result of those considerations should not be a refusal to provide all information to the data subject.” The WP29 *Guidelines on transparency*²⁴ further make it clear that businesses cannot rely on trade secret protection as an excuse to categorically deny access or refuse to provide information and recommends a case-by-case approach when dealing with conflicting values and interests.

Irrespective of the affordances of the IP rights, in practice, industry players mostly rely on their physical control over the systems to keep the ADM systems and the data in the dark. This physical control may also be complemented with contractual terms to prohibit any testing or reverse-engineering of the decision-making process.

5. Beyond impediments: The 2nd layer transparency

Where the technical limits and/or the business-related concerns prevent human-intelligible contestation, compliance with Article 22 requires the deployment of further tools and methodologies to render ADM systems contestable before a human arbiter and/or by use of software. This Part defines these tools and methodologies as the *2nd layer transparency* measures and explores their compatibility with the GDPR.

5.1. DPbD and implementing transparency under the GDPR

With the explicit reference in the GDPR, many of the *2nd layer transparency* measures (tools and methodologies) fall under the banner of *Data Protection by Design and by Default* (DPbD) as provided in Article 25. DPbD is a generic concept based on the idea that privacy intrusive or other harmful features of a product or service must be limited to what is necessary for the simple use of it. Under Article 25, data controllers are obliged to implement measures *in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects*. Combined with Article 28(1), this requires data controllers and processors to “hard wire” data protection norms into their systems’ architecture and *modus operandi* (Yeung & Bygrave 2020). As a “design-based” regulatory technique, the provision targets all personal data processing, irrespective of the technology used to construct and operate the systems.

DPbD entails a series of regulatory, technical, and organizational measures that should be actively followed through the entire life-cycle of data-driven practices. As Article 25 makes a general reference to the “requirements” under the GDPR, it is clear that the DPbD is not limited to implementation of certain data-protection principles but may be extended to a notion of “*Contestability by Design*” (CbD) as a sub-species (Almada 2019). In a similar vein, Hildebrandt and Koops (2010) suggest the concept of “smart transparency,” which refers to the designing of the socio-technical infrastructures carrying out automated decisions in such a way that individuals can anticipate and respond to how they are observed or profiled (p. 450).

Inclusion of DPbD as a legal obligation in the GDPR clarifies that the liability as to the design choices and the operational decisions of the ADM system lies with the data controller as the addressee of the norm and thus, may not be shifted to the contractors or third-parties. A further implication of the DPbD is that the design choices should address the rights and obligations outlined in the GDPR (e.g. right to contest), rather than referring to less definable and general notions such as privacy or accountability (Hildebrandt & Tielemans 2013, p. 517). Taking a broad view of “design,” Article 25 speaks of “appropriate technical and organizational measures and procedures,” which may extend as far as the integration of the relevant methodologies into the business models of data controllers.

5.2. Design choices to reduce complexity for more interpretable results

A variety of design choices might be introduced to enhance or facilitate the contestability of automated decisions as a form of *design-based regulation* aiming to prevent or inhibit certain conduct or social outcomes (Yeung 2015). By interfering with the design and construction stages of the system – rather than addressing its usage or consequences – the idea is to orchestrate the learning process so that the resulting ML model is more amenable to interpretation. *First*, as a rule, systems may be allowed to operate only on a limited set of possible features. By doing so, the total number of relationships handled by the algorithm may be reduced to a human-intelligible level. *Second*, the chosen learning method may allow for models that can be more easily parsed (e.g. decision tree algorithms) in comparison to, for instance, deep learning or neural network type of algorithms. A *third* method could be setting general parameters for the learning process to bring a threshold to complexity so that the resulting model would not defy human comprehension.

In general, *regularization* methods allow for model complexity to be taken into account during the learning process by assigning a cost for excess complexity (Selbst & Barocas 2018, p. 1112). In addition, linear models – with sufficiently small set of features – are regarded to be more concise for humans to grasp the relevant statistical relationships and to simulate different scenarios. Also, systems with *monotonicity* offer simpler models because in monotonic relationships, an increase in an input variable can only result in either an increase or a decrease in the output.

It is usually believed that there is a trade-off between the interpretability and the accuracy of the model. Models considering larger number of variables and more diverse relationships between these variables are assumed to be more accurate. Although this belief is largely relied upon by the system designers when modeling various kinds of problems, it is in fact questionable whether such comparison rests on a rigorous definition of interpretability or conforms to the findings from empirical studies. Hildebrandt notes, the developers may be inclined toward the “low hanging fruit,” meaning that they go after the data, which is easily available but not necessarily the most relevant or complete. Yet, more data do not always result in a better target function to define the problem in hand. A detailed model also carries the risk of overfitting and thus weakening its capacity to generalize new data (Hildebrandt 2018, pp. 102, 104). For instance, a model may assign significance to too many features and thus may learn patterns that are peculiar to the training data or not intuitively representative of the phenomena under analysis. Therefore, removing unnecessary features and accordingly reducing the complexity of the model to improve its interpretability does not necessarily decrease the performance of the system (Hand 2006).

Nevertheless, simpler models may not always be expressive enough for proxying sophisticated human behavior. Since ML is best suited to detect subtle patterns and intricate dependencies, it is possible that a complex phenomenon may require a complex model to better account for “reality” (Selbst & Barocas 2018, p. 1129). Article 25 of the GDPR acknowledges this need for “requisite complexity” by setting technical and economic feasibility as the two criteria informing the scope of the obligation of DPbD.²⁵ This confirms, as a principle, both technical difficulties and economic downsides may set a limit in terms of adopting plainly intelligible decision-making models.

5.3. Procedural, administrative, and institutional measures

As simpler models may not be feasible due to “requisite complexity,” there is need for further *institutional*, *administrative*, and *procedural* measures, which may facilitate the monitoring, review, and the contestation of automated decisions. These measures do not aim to impose limits on the ML process (e.g. capping the number of data features) but rather offer ways to improve the accountability and/or interpretability of the system.

Part of these measures fall under the concept of *procedural regularity* which, as a design principle, ensures that ML systems are actually doing what they are declared to be doing by their designers and operators (Kroll *et al.* 2017). In addition to procedural regularity, there are other design features and technical “add-ons” that may be implemented into the system during the development stage. These solutions aim to render ADM systems and their output intelligible to human reason or auditable through algorithmic means (algorithmic scrutiny). For instance, ADM systems may be designed to register processes leading to their actions, identify possible sources of uncertainty, and disclose any assumptions relied upon. In this regard, even a simple internal log kept by the system could enhance transparency for the purpose of contestation. Such records may be arranged to indicate the state of the model at the time of the decision or to provide information about the decisional input together with the rules actually employed for a specific outcome.

Apart from the procedural requirements, various administrative measures may be put to use to improve the accountability and interpretability of the ADM systems. In this respect, the *2nd layer* transparency/contestability measures also accommodate the idea of an institutional setup to carry out the necessary inspection and supervision tasks on behalf of data subjects. Where legitimate secrecy claims of the system designers/operators prevail, institutional oversight may be a way both for ex-ante inspection of the systems and for ex-post challenging of specific decisions (Sandvig *et al.* 2014). Such institutional review allows for “selective transparency,” assuring that critical information is not disclosed to the public but kept limited to the legally designated entities representing data subjects (Desai & Kroll 2017).

A number of GDPR provisions either explicitly provide for the above *institutional*, *administrative*, and *procedural* measures or leave space for them. The Recitals of the Regulation – together with several *guidelines* and *recommendations* published by the EU bodies – also articulate requirements and elaborate on the rights and obligations to this effect. In line with this, WP29 *Guidelines on automated decision-making* refers to *code of conduct* (Article 40, Recitals 77 and 98), *certification* (Article. 42), *agreed standards*, and *ethical review boards* as

formal mechanisms for scrutinizing ADM. Yeung and Bygrave (2020) regard these *self*-, *meta*-, and *coregulatory* instruments and techniques as a *cooperative problem-solving approach* between the regulator and the regulatee.

Though not mandatory under the GDPR, certification – which verifies that a product, process, or service adheres to a given set of standards and/or criteria – may require the disclosure of extensive technical information. This may include the source code, the hardware/software environments in which the systems has been developed, and the performance of the system in the testing environments (Hoffmann-Riem 2020, p. 13). As an integral part of the certification process, the International Organization for Standardization (ISO) has so far published several standards on *big data*, and further developing more on AI. *Institute of Electrical and Electronics Engineers Standards Association* (IEEE-SA) is also working on ML standards such as P7001, which focuses on transparent operation of autonomous systems. According to the 2019 report (*Ethically Aligned Design*) of the IEEE, the aim is to describe measurable and testable levels of transparency so that autonomous systems can be objectively assessed to determine their levels of compliance. Regarding these efforts, Matus and Veale (2020) note that, rather than prescribing concrete formulations, standard-setting initiatives have so far been limited to terminology issues and analytical frameworks laying out “meta-standards.”

Data Protection Impact Assessment (DPIA), as provided in Article 35 of the GDPR, could also serve as an important transparency mechanism which could aid the scrutiny of ML-based systems in various ways and dimensions. It is argued that a version of an Algorithmic Impact Assessment (AIA) might be derived from the DPIA to obtain an external review of the system. WP29 *Guidelines on automated decision-making* confirms this by mandating DPIA for any ADM subject to Article 22. In a similar vein, Kaminski and Malgieri (2020) approach DPIA as a “collaborative governance mechanism,” which may help determine the optimum extent of transparency and the appropriate mode of implementation of the right to contest in a specific ADM context (p. 72).

Despite these regulatory and institutional affordances provided by the GDPR together with various soft-law documents, it is arguable whether the current data protection authorities both at the member state and the Union level can handle such a wide range of regulatory, monitoring, and auditing tasks. As this will require a significant expansion of their powers and personnel, there are also views in favor of specialized institutions to monitor AI-based applications and develop performance standards (Hoffmann-Riem 2020, pp. 14–15).

5.4. Algorithmic scrutiny

There exist a variety of algorithmic scrutiny tools that enable both ex-ante and ex-post testing and verification of ADM processes. The deployment of these tools for the practical purpose of scrutinizing automated decisions present a spectrum ranging from modules integrated into the systems to stand-alone external audit tools for “black-box testing” (Pedreschi *et al.* 2018). Any combination of these may also, in varying degrees, involve humans in the decision loop to adjust the specifications and interpret the results.

Technologies for algorithmic scrutiny may be used both to approximate the ML model in general and also to discover the features that are most relevant for an individual decision. The former, a.k.a. “global interpretability,” aims to understand the underlying logic and the mode of reasoning of the system in its entirety (Selbst & Barocas 2018, p. 1113). Global interpretability can be regarded to generate a *model of the model* to simulate possible outcomes. The idea is to reconstruct the model on the basis of *interpretable rules* that describe the input–output relationships. “Local interpretability” on the other hand, is an ex-post method looking for the reasons for a specific decision. It is the “review of a software-driven action after the fact of the action” (Desai & Kroll 2017, p. 39). The local interpretability tools generally focus on importance-measuring methods aiming at explaining the most important variables for a specific result. It is a user-centric approach, where the importance of any feature to a particular decision is detected by iteratively varying the value of that feature, while holding the others constant. The idea is to develop an interpretable model taking on the predictions of a supposedly uninterpretable (black-box) model. This enables to determine the relative contribution of different features and identify the values that need to be altered to bring about a certain outcome (Ribeiro *et al.* 2016). Although local interpretability seems to work well to explain a specific decision, solely employing these limited techniques without a model-centric inspection or verification may be misleading. This is mainly due to the fact that an explanation that accounts

for a certain decision does not apply in the same way to other decisions (Doshi-Velez & Kortz 2017). That is, the reasons for a specific decision do not illustrate a general rule with regard to the system's behavior and thus, may be insufficient for the purposes of contesting another decision. Especially in terms of understanding the context of the decision, a proper scrutiny of automated decisions requires both the use of system-centric and user-centric approaches simultaneously.

Although ex-post techniques may be used as stand-alone scrutiny tools, their success depends on the extent to which they are reinforced by the necessary administrative, technical, and organizational measures in an overarching DPbD framework. As Article 25(1) clearly states that DPbD should be pursued through adequate technical and organizational means (*both at the time of the determination of the means for processing and at the time of the processing itself*), the provision may be interpreted to include the construction of software-based tools enabling contestation.

In addition, Recital 71 of the GDPR on automated decisions is also found to be supportive of software tools for the purpose of review and oversight. The Recital states that the data controllers are under the duty to implement technical and organizational measures against inaccuracies in data and to minimize the error against the risks involved for the rights and interests of the data subject. The WP29 *Guidelines on automated decisions* also recommends “algorithmic auditing” as a safeguard under Article 22. However, rather than the scrutiny of individual decisions, these references envisage algorithmic audit as a general model-centric tool to assess data controllers' compliance. The strongest support for the algorithmic contestation of specific decisions may be found in the wording of Article 22 itself. The provision defines the right to obtain human intervention as the least of the measures that the data controller could implement – implying that further solutions such as software-based tools could also be necessary for the purposes of contestation. Lastly, Article 21(5) on the right to object to personal data processing, including profiling, provides that “the data subject may exercise his or her right to object by automated means using technical specifications.”

6. Options for implementation

Having seen the possible legal, technical, and organizational measures to overcome the opacities and the legal barriers that may stand in the way of contestability of the automated decisions, it becomes clear that there is no one-size-fits-all solution (Kaminski 2019a). The problem lays with the framing of the optimum extent of transparency and the appropriate mode of implementation without prejudice to the integrity of the systems or to the legitimate interests of the stakeholders involved.

Following the analysis in the previous Parts, what remains yet to be answered is the question of *which uses of ADM should be subject to right to contest, and through which (combination) of the above tools and measures?* As this paper is primarily focused on a conceptual analysis of the right to contest and the relevant affordances provided by the GDPR, further enforcement-related particularities are beyond the scope. Therefore, this Part is limited to an outline of the possible regulatory options regarding the implementation of the right to contest.

6.0.1. Not permissible

Where satisfactory measures for the exercise of the right to contest at human-intelligible level are not possible or feasible, taking into account the potential risks to the rights of the data subjects or to the societal interests, solely automated decisions may be banned. In case the benefits from the ADM do not justify the risks that it creates, system developers and operators should look for hybrid (human-machine symbiotic) approaches.

In their implementation of the GDPR, many member states have introduced prohibitions or restrictions for certain types of decisions. Among them, French Law adopts a regime based on traditional state functions and accordingly prohibits fully or semiautomated *judicial decisions* aiming to evaluate aspects of personality, while permitting automated *administrative decisions* subject to strict conditions (Malgieri 2019, p. 13–14). The exclusion of certain types of automated decisions also means that data protection principles shall be respected at all stages of ADM, including the initial decision on whether or not to carry out the processing.

6.0.2. Permissible – Subject to ex-ante design and procedural requirements and/or ex-post algorithmic scrutiny

Despite the methodological distinction we have made in Part IV, transparency issues are never of purely technical, legal, or institutional nature. The complexity-related problems – that are deeply intertwined with both

deliberate and unintentional design features – often result in unintelligible ML models. Therefore, determining the right combination of the *2nd layer* transparency measures requires a context-specific case-by-case approach, taking into account the technical limits, possible gaming strategies, and the competition-related issues. While in some cases only an ex-post analysis (black-box testing) may suffice, the *institutional, administrative, or procedural* measures are more effective when accompanied with ex-post tools and methodologies.

6.0.3. *Permissible – Only subject to ex-post black-box testing*

This regulatory option deals with the situations, where it is not possible or permissible to impose constraints on the model or apply other ex-ante measures at the design stage. In such cases, data controllers may still be allowed to carry on with ADM subject to ex-post algorithmic scrutiny measures. Since black-box testing alone may not reveal sufficient insights about the decision-making process, this type of scrutiny may remain limited to testing of the outcome against some minimum (e.g. fairness, antidiscrimination, or due process) requirements, without involving a full-fledged contestation. Considering the risks involved, there may be need for further procedural safeguards such as the immediate suspension of the automated decision upon challenge or the reversed burden of proof in the contestation proceedings.

6.0.4. *Permissible without restrictions – Only subject to 1st layer access and notification requirements*

This is where the individual access rights and notification duties under the GDPR suffice to provide functional and systemic information about the input, the decision rules, and the underlying causal relations necessary for contestation.

7. Conclusion

The implementation of the above options (modalities) as a practically meaningful transparency scheme requires deeper interdisciplinary research on two parallel but interacting tracks. *First*, there is need for an elaboration of the technical limits and other impediments to human-intelligible models. That is, where there are genuine technical and legal barriers and where complexity and/or legal claims are used as a pretext for unsubstantiated or unlawful secrecy practices. This line of inquiry, briefly touched upon in Part IV, will involve the consideration of heterogeneous composition of legal matters including fundamental rights, legislative initiatives, case-law, data protocols, regulatory tools, contracts, and so on. Without understanding the true nature and the cause of intransparencies in a given system, it will not be possible to calibrate the measures to be implemented. It is hoped that the analysis provided in this paper could serve as a “launching pad” for further legal research to draw an entire picture of legal impediments in relation to the implementation of the right to contest. As we identify what counterbalancing rights and interest are at stake on the side of the data controllers and other industry actors, the *second* line of research should then inquire how to define and treat the risks to the rights of the data subjects. Risk assessment for the practical application of the above regulatory options should initially identify the conditions where a less interpretable model – for the sake of efficiency gains or other alleged benefits – could be justified. This would require an overall consideration of the type and source of the data, the reliability of the ML model, the specific conditions of the data subject, and most importantly, the materiality of the output to the individuals and third parties concerned.

Despite the flux of academic papers in the recent years about the *transparency, explainability, interpretability, legibility*, and the discriminative and unfair effects of automated decisions, we are still far from establishing any practical way of exercising rights under Article 22. To bridge the gaps between the GDPR and its practical application, the affordances laid out in this paper need to be operationalized as a comprehensive compliance scheme. Failure to do this uniformly at the EU level may result in a fragmented implementation, rendering legal safeguards to a great extent ineffective (Massé & Lemoine 2019). That is, despite the proactive approach of some member states such as Hungary and France, for many member states, Article 22 may remain an ancillary provision as it has been during the time of the DPD (Malgieri 2019). Considering the current political incoherence among the EU member states, it would not be unrealistic to expect further disarray with regard to the implementation of Article 22 if Member States are left to legislate on their own discretion.²⁶

Whether or not the GDPR provisions on automated decisions will turn out to be a toothless mechanism depends on whether the EDPB and other EU authorities take prompt action. In this respect, there is need for an

agenda for the development of a dynamic and scalable compliance regime with concrete practical targets. The progress on this front will determine whether Bygrave (2001) is still right in his conclusion about Article 15 of the DPD (GDPR, Art. 22), which he had phrased 20 years ago as: “all dressed up but nowhere to go.”

Acknowledgments

The author received funding from Tilburg Institute for Law, Technology, and Society (LTMS-TILT), Tilburg University, The Netherlands, and the paper was finalized during a research stay at the Institute of Computing and Information Sciences (iCIS), the Science Faculty of Radboud University, The Netherlands, funded by the Privacy and Identity Lab.

Endnotes

- ¹ Similar transparency obligations (requirements) are emerging in various legal domains and regulatory frameworks, for example, *Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services*. The Regulation foresees disclosure obligations for the providers of online intermediation services which use algorithms to rank goods and services. See Wischmeyer 2020, p. 77.
- ² For a literature review shedding light on the diversity of the concept of *algorithmic regulation*, see Eyert *et al.* 2021.
- ³ Approaching transparency in ADM as a procedural problem also aligns with the process-based nature of the EU data protection regime, which has long focused on the regulation of digital interactions while maintaining regulatory flexibility in the face of complex and ever-changing technology (Yeung & Bygrave 2020).
- ⁴ For different conceptions of transparency under the GDPR, see Felzmann *et al.* 2019. Authors propose to understand transparency relationally, where information provision is conceptualized as communication between technology providers and users, and where assessments of trustworthiness based on contextual factors mediate the value of transparency communications.
- ⁵ For this “layered” approach, see Bayamlioğlu “Transparency of Automated Decisions in the GDPR: An Attempt for Systemization” (PLSC 2018 Discussion Paper, Brussels). Kaminski also speaks of “tiers” of transparency under the GDPR, albeit in a different conception (Kaminski 2019a). Diakopoulos and Koliska (2017) use the “layers” concept to define various aspects of disclosable information in relation to a “transparency model”. Also see note 15.
- ⁶ The origin of these provisions is attributed to French data protection legislation enacted in 1978, which prohibited behavioral assessment through automated processing in legal matters (Bygrave 2001, p. 21; Korff 2010, p. 83).
- ⁷ Hildebrandt notes, albeit critically, “[t]he fact that usually some form of routine human intervention is involved means that art. 15 is not applicable, even if such routine decisions may have the same result as entirely automated decision making” (Hildebrandt 2008, p. 28, fn. 22).
- ⁸ EC Commission’s amended proposal of DPD 1992. COM(92) 422 final – SYN 287, 15.10.1992, 26.
- ⁹ Here, the question arises whether *consent* as used in Article 22 has the same requirements as *consent* defined in Articles 3 and further regulated in Art. 7 of the GDPR.
- ¹⁰ For instance, WP29 *Guidelines on automated decisions* is also silent on this matter. The Guidelines refer to the safeguards cumulatively as “a further layer of protection for data subjects” (p. 15).
- ¹¹ This procedural approach to Article 22, treating DP law as meta-regulation (a body of rules, regulating a regulatory technology), partly addresses the arguments that DP regime slowly becomes the “law of everything,” penetrating every sector where digital technologies are involved, see Purtova 2018.
- ¹² “Normativity” is hereby used in the sense of not being explicable in purely factual terms. Such attribution does not make sense for ordinary physical objects, such as rocks, geological systems, or oil molecules—leaving aside living organisms.
- ¹³ The transparency implications (contestability requirements) in this Section partially parallels with Diakopoulos and Koliska’s (2017) “transparency model,” which enumerates *information factors* that might be disclosed about algorithms. Their model provides a set of pragmatic dimensions of information (across layers such as data, model, inference, and interface) that are essential for algorithmic transparency efforts.
- ¹⁴ See WP29 2018 *Guidelines on automated decisions*, raising the questions about the “categories of data used in the profiling or decision-making process” and “why these categories are considered pertinent” in relation to the implementation of Article 22. In addition, Article 14 of the GDPR clearly counts “categories of personal data” among the information to be communicated to the data subject, though without guidance about what these categories might be other than sensitive/

nonsensitive data (also see Article 28 and 30). On this regulatory gap regarding the categories of personal data, see Wachter *et al.* (2018).

- ¹⁵ *Jehovan todistajat*, C-25/17, EU:C:2018:551.
- ¹⁶ The case discussed the joint controllership between the website owner, *Fashion ID*, and *Facebook Ireland*, where the former has embedded in its website the Facebook “Like” button. Case C-40/17, *Fashion ID GmbH & Co. KG v Verbraucherzentrale NRW eV*. ECLI:EU:C:2019:629.
- ¹⁷ WP29 *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (last revised and adopted on 6 February 2018). As of 25 May 2018, the Article 29 Working Party (WP29) ceased to exist and is replaced by the EDPB.
- ¹⁸ The “technical limits,” as explained here, partly corresponds with the concept of “epistemic opacity” that Evert and others define as “inherent methodological intransparency of ML” (Eyert *et al.* 2021). Also see Wischmeyer referring to “epistemic constraints” (Wischmeyer 2020, p. 80).
- ¹⁹ Evert and others conceptualize these informational asymmetries as “sociomaterial opacity” arising due to the concentration of massive data sets in the hands of a few private companies as well as the inaccessibility of closed-source algorithms (Eyert *et al.* 2021).
- ²⁰ On the question of when algorithms need to be kept secret due to the risk of gaming and when disclosure is permissible, see Cofone & Strandburg 2019.
- ²¹ *European Parliament and the Council, Directive (EU) 2016/943 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure* (OJ L 157/1, 15.6.2016), 8 June 2016.
- ²² *Directive 2009/24/EC on the legal protection of computer programs* (Computer Programs Directive), (OJ L 111/16, 5.5.2009).
- ²³ On this matter, see Benjamin L. W. Sobel, “Artificial Intelligence’s Fair Use Crisis” 41 COLUM. J.L. & ARTS 45 (2017); Banterle (2018); Mattioli (2014).
- ²⁴ WP29 ‘Guidelines on transparency under Regulation 2016/679 (WP260)’ (EC, 24 January 2018)
- ²⁵ It should also be noted that “whatever seemed technically and/or economically infeasible during the design of the data processing system, will again be considered once the processing is in operation” (Hildebrandt & Tieleman 2013, p. 517).
- ²⁶ Malgieri provides a detailed account of the current implementation efforts of the EU member states. His analysis reveals that member states have so far taken various approaches mostly without a clear or concrete methodology. Apart from some good examples, member state laws generally rephrase the wording of the official documents or refer to “explanation” in an abstract and descriptive way (Malgieri 2019).

References

- Almada M (2019) Human Intervention in Automated Decision-Making: Toward the Construction of Contestable Systems. ICAIL '19: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, June 2019 pp. 2–11. <https://doi.org/10.1145/3322640.3326699>
- Baer T (2019) *Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists*. Apress Imprint, Berkeley, CA.
- Bambauer J, Zarsky T (2018) The Algorithm Game. *Notre Dame Law Review* 94(1), 1–48.
- Banterle F (2018) The Interface between Data Protection and IP Law. In: Bakhoun M *et al.* (eds) *Personal Data in Competition, Consumer Protection and Intellectual Property Law*, pp. 412–440. Springer-Verlag GmbH, Germany.
- Bateson G (1987) *Steps to an Ecology of Mind*, (1972 1st edition). Jason Aronson Inc, Northvale, NJ, and London.
- Bayamlioğlu E (2018) Contesting Automated Decisions. *European Data Protection Law Review* 4(2018), 433–446.
- Bayamlioğlu E, Leenes R (2018) The Rule of Law Implications of Data-Driven Decision-Making: A Techno-Regulatory Perspective. *Law, Innovation and Technology* 10, 295–313.
- Binns R (2017) Algorithmic Accountability and Public Reason. *Philosophy & Technology* 31, 543–556.
- Burrell J (2016) How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3(1), 1–12.
- Bygrave L (2001) Automated Profiling: Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling. *Computer Law & Security Review: The International Journal of Technology Law and Practice* 17, 17–24.
- Castañeda HN (1970) On the Semantics of the Ought-to-Do. *Synthese* 21, 449–468.
- Cath C (2018) Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges. *Philosophical Transactions of the Royal Society A* 376, 20180080. <https://doi.org/10.1098/rsta.2018.0080>.
- Cofone I, Strandburg K (2019) Strategic Games and Algorithmic Secrecy, 64 McGill L.J.
- Commission, E., Directorate General Justice, F. & Security (2010). Comparative Study on Different Approaches to New Privacy Challenges in Particular in the Light of Technological Developments. Working Paper No. 2. In: D. Korff (ed.), *Data*

- Protection Laws in the EU: The Difficulties in Meeting the Challenges Posed by Global Social and Technical Developments, London Metropolitan University, London.
- De Ridder J (2006) The Inherent Normativity of Technological Explanations. *Techné: Research in Philosophy and Technology* 10(1), 79–94.
- Desai D, Kroll J (2017) Trust but Verify: A Guide to Algorithms and the Law. *Harvard Journal of Law & Technology* 31, 2–64.
- Diakopoulos N, Koliska M (2017) Algorithmic Transparency in the News Media. *Digital Journalism* 5(7), 809–828. <https://doi.org/10.1080/21670811.2016.1208053>.
- Doshi-Velez F, Kortz M (2017) Accountability of AI under the Law: The Role of Explanation. Harvard University Berkman Klein Center Working Group on Explanation & the Law, Working Paper No. 18-07. <http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584>
- Economides C (2010) Content of the Obligation: Obligations of Means and Obligations of Result. In: Crawford J, Pellet A, Olleson S, Parlett K (eds) *The Law of International Responsibility*. Oxford University Press, New York.
- Edwards L, Veale M (2017) Slave to the Algorithm? Why a “Right to an Explanation” Is Probably Not the Remedy You Are Looking for. *Duke Law & Technology Review* 16, 18–84.
- Eyert F, Irgmaier F, Ulbricht L (2021) Extending the Framework of Algorithmic Regulation. The Uber Case. *Regulation & Governance* (in this issue).
- Felten E (2017) What Does It Mean to Ask for an “Explainable” Algorithm? *Freedom to Tinker*. Available from URL: <https://freedom-to-tinker.com/2017/05/31/what-does-it-mean-to-ask-for-an-explainable-algorithm/>
- Felzmann H, Villarronga EF, Lutz C, Tamò-Larrieux A (2019) Transparency Requirements for Artificial Intelligence between Legal Norms and Contextual Concerns. *Big Data & Society* 6, 1–14. <https://doi.org/10.1177/2053951719860542>.
- Gollnick C (2018) Induction Is Not Robust to Search. In: Bayamlioğlu E, Baraluic I, Janssens L, Hildebrandt M (eds) *Being Profiled Cogitas Ergo Sum: 10 Years of Profiling the European Citizen*, pp. 106–111. Amsterdam University Press, Amsterdam.
- Guinchard A (2017) Contextual Integrity and EU Data Protection Law: Towards a More Informed and Transparent Analysis. *SSRN Electronic Journal*, 2017. <https://dx.doi.org/10.2139/ssrn.2946772>.
- Hand D (2006) Classifier Technology and the Illusion of Progress. *Statistical Science* 21, 1–14.
- Hildebrandt M (2008) Defining Profiling: A New Type of Knowledge. In: Hildebrandt M, Gutwirth S (eds) *Profiling the European Citizen Cross-Disciplinary Perspectives*, pp. 17–45. Springer, Dordrecht.
- Hildebrandt M (2018) Preregistration of Machine Learning Research Design. Against P-Hacking. In: Bayamlioğlu E, Baraluic I, Janssens L, Hildebrandt M (eds) *Being Profiled Cogitas Ergo Sum: 10 Years of Profiling the European Citizen*, pp. 102–105. Amsterdam University Press, Amsterdam.
- Hildebrandt M (2019) Privacy as Protection of the Incomputable Self: From Agonistic to Agnostic Machine Learning. *Theoretical Inquiries of Law* 19(1), 83–121.
- Hildebrandt M, Koops J (2010) The Challenges of Ambient Law and Legal Protection in the Profiling Era. *Modern Law Review* 73(3), 428–460.
- Hildebrandt M, Tielemans L (2013) Data Protection by Design and Technology Neutral Law. *Computer Law & Security Review* 29(5), 509–521.
- Hoepman J (2018) Transparency as Translation in Data Protection. In: Bayamlioğlu E, Baraluic I, Janssens L, Hildebrandt M (eds) *Being Profiled Cogitas Ergo Sum: 10 Years of Profiling the European Citizen*, pp. 102–105. Amsterdam University Press, Amsterdam.
- Hoffmann-Riem W (2020) Artificial Intelligence as a Challenge for Law and Regulation. In: Wischmeyer T, Rademacher T (eds) *Regulating Artificial Intelligence*, pp. 1–32. Springer, Cham.
- Jasanoff S, Simmet H (2017) No Funeral Bells: Public Reason in a “Post-Truth” Age. *Social Studies of Science* 47(5), 751–770.
- Kaminski M (2019) The Right to Explanation, Explained. *Berkeley Technology Law Journal* 34(1), 189. <https://scholar.law.colorado.edu/articles/1227>.
- Kaminski M (2019a) Binary Governance: Lessons from the GDPR’s Approach to Algorithmic Accountability. *Southern California Law Review* 92(6), 1–77.
- Kaminski M, Malgieri G (2020) Multi-Layered Explanations from Algorithmic Impact Assessments in the GDPR. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 68–79. <https://doi.org/10.1145/3351095.3372875>.
- Kleinberg J, Raghavan M (2018) How Do Classifiers Induce Agents To Invest Effort Strategically? arXiv: 1807.05307v5
- Krist V (2006) How Norms in Technology Ought to Be Interpreted. *Techné: Research in Philosophy and Technology* 10(1), 95–108. <https://doi.org/10.5840/techne200610144>.
- Kroll JA (2018) The Fallacy of Inscrutability. *Philosophical Transactions of the Royal Society A* 376(2133). <http://doi.org/10.1098/rsta.2018.0084>.
- Kroll JA, Huey J, Baroas S et al. (2017) Accountable Algorithms. *University of Pennsylvania Law Review* 165(3), 633–705.
- Lipton Z (2016) The Mythos of Model Interpretability. *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York. [Last accessed 25 Jun 2019.] Available from URL: <https://arxiv.org/pdf/1606.03490.pdf>
- Macmillan R (2018) *Big Data, Machine Learning, Consumer Protection and Privacy*. Geneva, Switzerland: International Telecommunication Union (ITU) Security, Infrastructure and Trust Working Group.
- Malgieri G (2019) Automated Decision-Making in the EU Member States: The Right to Explanation and Other “Suitable Safeguards” in the National Legislations. *Computer Law & Security Review* 35(5), 105327.
- Malgieri G, Comandé G (2017) Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation. *International Data Privacy Law* 7(4), 243–265.

- Mantelero A (2019) Artificial Intelligence and Data Protection: Challenges and Possible Remedies. The Council of Europe's Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. <https://rm.coe.int/artificial-intelligence-and-data-protection-challenges-and-possible-re/168091f8a6>
- Massé E, Lemoine L (2019) One Year Under the GDPR, *Access Now Publication*. Available from URL: <https://www.accessnow.org/cms/assets/uploads/2019/06/One-Year-Under-GDPR.pdf>
- Matus KJM, Veale M (2020) Certification Systems for Machine Learning: Lessons from Sustainability. *Regulation & Governance* (in this issue).
- Mendoza I, Bygrave L (2017) The Right Not to Be Subject to Automated Decisions Based on Profiling. In: Synodinou T, Jougoux P, Markou C, Prastitou T (eds) *EU Internet Law*, pp. 78–98. Springer, Cham.
- Pedreschi D, Giannotti F, Guidotti R, Monreale A, Pappalardo L, Ruggieri S, Turini F (2018) Open the Black Box Data-Driven Explanation of Black Box Decision Systems. arXiv: 806.09936v1.
- Purtova N (2018) The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law. *Law, Innovation and Technology* 10, 74–75.
- Rader E, Cotter K, Cho J (2018) Explanations as Mechanisms for Supporting Algorithmic Transparency. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pp. 1–13. Montréal, QC, Canada: ACM Press. <https://doi.org/10.1145/3173574.3173677>.
- Ribeiro MT, Singh S, Guestrin C (2016) “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD '16*, pp. 1135–1144. San Francisco, CA: ACM.
- Sandvig C, Hamilton K, Karahalios K, Langbort C. (2014) Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. Presented at *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. 22 May, Seattle, WA.
- Selbst A, Barocas S (2017) Regulating Inscrutable Systems Available from URL: <http://www.werobot2017.com/wp-content/uploads/2017/03/Selbst-and-Barocas-Regulating-Inscrutable-Systems-1.pdf>
- Selbst A, Barocas S (2018) The Intuitive Appeal of Explainable Machines. *Fordham Law Review* 87, 1085–1139.
- Thaler RH, Sunstein CR (2008) *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.
- Wachter S, Mittelstadt B, Floridi L (2017) Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law* 7(2), 76–99.
- Wachter S, Mittelstadt B, Russell C (2018) Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31(2), 841–887.
- Wischmeyer T (2020) Artificial Intelligence and Transparency: Opening the Black Box. In: Wischmeyer T, Rademacher T (eds) *Regulating Artificial Intelligence*, pp. 75–101. Springer, Cham.
- Yeung K (2015) Design for Regulation. In: van den Hoven J, van de Poel I, Vermaas PE (eds) *Handbook of Ethics, Values and Technological Design*, pp. 447–472. Springer, Dordrecht.
- Yeung K (2018) Algorithmic Regulation: A Critical Interrogation. *Regulation & Governance* 12(4), 505–523.
- Yeung K, Bygrave L (2020) A Critical Examination of the Legitimacy of the Modernised European Data Protection Regime Through a “Decentred” Regulatory Lens. *Regulation & Governance*, forthcoming.
- Zarsky T (2017) Incompatible: The GDPR in the Age of Big Data. *Seton Hall Law Review* 47, 995–1020.
- Zeng Z, Fan X, Miao C, Wu Q, Cyril L (2018) Context- Based and Explainable Decision Making with Argumentation. *AAMAS '18 Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 10–15 Jul, Stockholm, pp. 1114–1122. Available from URL: <http://ifaamas.org/Proceedings/aamas2018/pdfs/p1114.pdf>