



## Algorithmic explainability and legal reasoning

Zsolt Zódi

To cite this article: Zsolt Zódi (2022): Algorithmic explainability and legal reasoning, The Theory and Practice of Legislation, DOI: [10.1080/20508840.2022.2033945](https://doi.org/10.1080/20508840.2022.2033945)

To link to this article: <https://doi.org/10.1080/20508840.2022.2033945>



Published online: 21 Feb 2022.



Submit your article to this journal [↗](#)




View related articles [↗](#)



View Crossmark data [↗](#)



## Algorithmic explainability and legal reasoning

Zsolt Zódi 

Institute of the Information Society, University of Public Service, Budapest,  
Hungary

### ABSTRACT

Algorithmic explainability has become one of the key topics of the last decade of the discourse about automated decision making (AMD, machine-made decisions). Within this discourse, an important subfield deals with the explainability of machine-made decisions or outputs that affect a person's legal position or have legal implications in general – in short, the algorithmic legal decisions. These could be decisions or recommendations taken or given by software which support judges, governmental agencies, or private actors. These could involve, for example, the automatic refusal of an online credit application or e-recruiting practices without any human intervention, or a prediction about one's likelihood of recidivism. This article is a contribution to this discourse, and it claims, that as explainability has become a prominent issue in hundreds of ethical codes, policy papers and scholarly writings, so it has become a 'semantically overloaded' concept. It has acquired such a broad meaning, which overlaps with so many other ethical issues and values, that it is worth narrowing down and clarifying its meaning. This study suggests that this concept should be used only for individual automated decisions, especially when made by software based on machine learning, i.e. 'black box-like' systems. If the term explainability is only applied to this area, it allows us to draw parallels between legal decisions and machine decisions, thus recognising the subject as a problem of legal reasoning, and, in part, linguistics. The second claim of this article is, that algorithmic legal decisions should follow the pattern of legal reasoning, translating the machine outputs to a form, where the decision is explained as applications of norms to a factual situation. Therefore, as the norms and the facts should be translated to data for the algorithm, so the data outputs should be back-translated to a proper legal justification.

**KEYWORDS** Algorithmic decision making; ADM; algorithmic explainability; transparency of artificial intelligence; legal reasoning; translation theory

**CONTACT** Zsolt Zódi  [zodi.zsolt@uni-nke.hu](mailto:zodi.zsolt@uni-nke.hu)  Institute of the Information Society, University of Public Service, Budapest, Hungary

© 2022 Informa UK Limited, trading as Taylor & Francis Group

## 1. Introduction

The explainability of algorithms<sup>1</sup> is a discourse that has existed for decades<sup>2</sup> and can be classified as a sub-field in ‘human-machine interaction’<sup>3</sup> within computer science. With the proliferation of applications based on artificial intelligence, the importance and weight of the topic has grown significantly in the last decade.

Within this discourse, a well-identifiable important subfield deals with the explainability of machine-made decisions and machine outputs that affect a person’s legal position or that are legally relevant. Such decisions include certain administrative decisions made by machine, or proposals for systems used in the course of judicial work (although no court decisions are made by machine yet). In a broader sense, the recommendations of pre-contractual expert systems in the private sector, such as the outputs of recruitment or credit assessment systems, can also be included here. This covers the area of algorithmic legal decisions, or recommendations. In this paper, I will deal only with this sub-area.<sup>4</sup>

Machine-made decisions and recommendations have many benefits: their application may even seem to fulfil Montesquieu’s dream that those who apply the law (the judges in their writing) are really just ‘mouths of the law’ (*bouches de la loi*).<sup>5</sup> Machines have no hidden intentions, biases, or

---

<sup>1</sup>The literature of explainable algorithms has grown extremely large in the past few years. Barredo Arrieta and others analysed more than 400 publications in an article (Alejandro Barredo Arrieta and others, ‘Explainable Artificial Intelligence (xai): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI’ (2020) 58 *Information Fusion* 82).

<sup>2</sup>Some other fresh publications: Pentelis Linardatos, Vasilis Papastefanopoulos, and Sotoris Kotsiantis, ‘Explainable AI: A Review of Machine Learning Interpretability Method’ (2021) 23(1) *Entropy* <<https://www.mdpi.com/1099-4300/23/1/18>> accessed 10 December 2021; Amina Adadi and Mohammed Berrada, ‘Peeking Inside the Black-box: A Survey on Explainable Artificial Intelligence, (xai)’ (2018) 6 *IEEE Access* 52138; Jorg Cassens and Rebekah Wegener, ‘Intrinsic, Dialogic, and Impact Measures of Success for Explainable AI’ in Jörg Cassens and Rebekah Wegener and Anders Kofod-Petersen (eds.), *Proceedings of the Twelfth International Workshop Modelling and Reasoning in Context (MRC 2021)*.

<sup>3</sup>The field has recently been called ‘human-computer interaction.’ It has a dedicated journal (*Human-computer interaction*, published by Taylor and Francis). Illustrative books on the topic: Alan Dix and others, *Human-computer Interaction* (4th edn., Pearson Education 2004); Silvia Pflieger, Joao Goncalves, Kadamula Varghese (eds.), *Advances in Human-Computer Interaction* (Springer 1995); Recently: Sergio Sayago, *Perspectives on Human-Computer Interaction Research with Older People* (Springer Nature 2019). For a good overview of cases of explainability in various situations of life see Mersedeh Sadegh, Verena Klös and Andreas Vogelsang, ‘Cases for Explainable Software Systems: Characteristics and Examples’ available online <<https://arxiv.org/pdf/2108.05980.pdf>> accessed 10 December 2021.

<sup>4</sup>There is an extensive literature about explainability of recommendations and diagnosis made by AI based on image-recognition. See for example, José Luis Solorio-Ramírez and others, ‘Brain Hemorrhage Classification in CT Scan Images Using Minimalist Machine Learning’ (2021) 11 *Diagnostics*, 1449; Amitojdeep Singh, Sourya Sengupta and Vasudevan Lakshminarayanan, ‘Explainable Deep Learning Models in Medical Image Analysis’ (2021) 6 *Journal of Imaging* 52 online available <<https://www.mdpi.com/2313-433X/6/6/52>> accessed 10 December 2021; Fabrizio Nunnari, Abdul Kadir and Daniel Sonntag, ‘On the Overlap Between Grad-CAM Saliency Maps and Explainable Visual Features in Skin Cancer Images’ in Andreas Holzinger and others (eds.), *Machine Learning and Knowledge Extraction. CD-MAKE 2021. Lecture Notes in Computer Science*, vol. 12844 (Springer 2021).

<sup>5</sup>‘Mais les juges de la nation ne sont, comme nous avons dit, que la bouche qui prononce les paroles de la loi’ Charles de Secondat de Montesquieu, *De l’esprit des lois* Book XI, Chapter 6. (online edition)

agendas, operate extremely rationally, and have no data-processing limitations. At the same time, it has become clear that although people may long for a completely impartial and rational decision-maker, in reality this is not always to their liking. One of the neuralgic points of such seemingly perfect decision-making, besides other fears<sup>6</sup>, is that machines cannot properly explain and justify their decisions.

The last decade and a half has brought a turning point in the subject of algorithmic explainability, as explainability has become a central theme and one of the important requirements of ‘big data ethics’<sup>7</sup> and later of ‘artificial intelligence ethics’.<sup>8</sup> Although certain information rights related to automated data processing were already included in the old data protection rules, with the entry into force of the GDPR disputes over clarity have flared up.<sup>9</sup> In these debates, explainability is defined as an ethical requirement, together with other ethical requirements such as interpretability, fairness, transparency, accountability, freedom, respect for human dignity or the right to self-determination. Behind all these requirements is the basic idea that people have a right to be aware of the rules that they are expected to follow so that they can understand the reasons for decisions that affect their lives, and how those decisions can be traced back to rules.

In this study, I would like to explain two interrelated arguments about algorithmic explainability in law. The first is that, over the past few years, this concept has become very confused, or in other words semantically overloaded. For this reason, I will try to make a clear distinction between the concepts of explainability, transparency, accountability and fairness. I suggest applying the requirement of explainability only in the case of machine decisions and hence separating it from content-substantive expectations.

My second argument is that clarifying and narrowing the meaning of the term provides a way to regain and expand the legal and linguistic dimension of the subject, since the legal dimension is primarily a tradition of theories of

---

116 <[https://archives.ecole-alsacienne.org/CDI/pdf/1400/14055\\_MONT.pdf](https://archives.ecole-alsacienne.org/CDI/pdf/1400/14055_MONT.pdf)> accessed 14 December 2021.

<sup>6</sup>Henrik Palmer Olsen, Jacob Livingston Slosser and Thomas Troels Hildebrandt, ‘What’s in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration’ in Hans-W. Micklitz and others (eds), *Constitutional Challenges in the Algorithmic Society* (CUP 2022) 220–221.

<sup>7</sup>Neil M. Richards and Jonathan H. King, ‘Big Data Ethics’ (2014) 49 *Wake Forest Law Review* 393; Solon Barocas and Andrew D. Selbst, ‘Big Data’s Disparate Impact’ (2016) 104 *California Law Review* 671.

<sup>8</sup>Vincent Müller, ‘Ethics of Artificial Intelligence and Robotics’ in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (2021 summer edition), <<https://plato.stanford.edu/archives/sum2021/entry/ethics-ai/>> Mark Coeckelbergh, *AI Ethics* (The MIT Press 2020) and European Commission, ‘Proposal for a Regulation of the Parliament and the European Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), and amending certain Union legislative acts’, SEC(2021) 167 final) – {SWD(2021) 84 final} – {SWD(2021) 85 final} ‘to address the opacity that may make certain AI systems incomprehensible to or too complex for natural persons, a certain degree of transparency should be required for high-risk AI systems. Users should be able to interpret the system output and use it appropriately’. (recital 47).

<sup>9</sup>See section 2.1.

legal reasoning, which can be the source of many inspiring ideas. By linguistic dimension I mean the application of the language translation theory developed by Roman Jakobson to algorithmic explainability.<sup>10</sup>

The study will proceed as follows. The first part reviews the process of the semantic overloading of explainability. To do this, I will first summarise the history of the legal treatment of machine (automated) decisions and then reconstruct the process whereby explainability became intertwined with other ethical requirements, such as accountability, legality, and transparency. Finally, I will propose a clearer use of the term.

The second part analyses the legal theoretical and linguistic aspects of explainability. My starting point here is that translating the outputs of algorithms into human language is an intersemiotic (inter-systemic) translation in which a description of the initial rules and facts, as well as the internal processes within the ‘black box’, must be present at the same time. Accordingly, the reasoning of machine decisions must differ from the traditional legal reasoning in several respects.

## 2. Explainable algorithms: a semantically overloaded concept

### 2.1. *The explainability of machine-made legal decisions – a short history*

The automated individual decision appeared in Europe as a legal concept in Article 15 of the first Data Protection Directive.<sup>11</sup> As a general rule, the data subject had a primary right to ‘opt-out’, of such a decision.<sup>12</sup> The wording of ‘right to an explanation’ remained rather vague and uncertain in Article 12, which states that the data subject has the right ‘to obtain information [...] of the logic involved in any automatic processing of data concerning him’.<sup>13</sup> There is no established case law on the provision, so it is difficult to judge what the information on the logic involved would have looked like under the old rules.

This provision has been transferred to the GDPR<sup>14</sup> with some minor changes. The GDPR has, on the one hand, supplemented the provisions on automatic data processing with some new stakeholder licenses and on

---

<sup>10</sup>Roman Jakobson, ‘On Linguistic Aspects of translation’ in Reuben Arthur Brower (ed.), *On Translation* (Harvard University Press 1959) 232–239. Online available: <<https://web.stanford.edu/~eckert/PDF/jakobson.pdf>> accessed 10 December 2021.

<sup>11</sup>Directive 95/46/EC of the European Parliament and the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

<sup>12</sup>‘Member States shall grant the right to every person not to be subject to a decision [...] which is based solely on automated processing of data’ *ibid* 15(1).

<sup>13</sup>*ibid* 12(a).

<sup>14</sup>Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (GDPR).

the other hand, introduced some new rules on ‘profiling’ within the category of automated decision-making.<sup>15</sup> The right to an explanation is only mentioned in the GDPR among the non-binding recitals in the context of automated data processing:

such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, *to obtain an explanation of the decision reached* after such assessment and to challenge the decision.<sup>16</sup>

The Article 29 Working Party’s<sup>17</sup> expert material on automated decisions also interpreted the older provisions in the GDPR as meaning that the explanation does not apply to individual decisions, but that the controller ‘should provide the data subject with general information (notably, on factors taken into account for the decision-making process, and on their respective ‘weight’ on an aggregate level) which is also useful for him or her to challenge the decision’.<sup>18</sup>

It remains unclear whether the provision has raised problems in (judicial) practice. Machine processes, where they exist at all (e.g. in automatic traffic control and fine systems<sup>19</sup>) are usually designed so that there is always the opportunity to appeal to a human decision maker. In addition, these decisions are so simple (they use some easily accessible, verifiable parameters) that although they are very easy to understand, yet at the same time quite difficult to dispute. In the case of automatic traffic control systems, the decision itself usually takes the form of a standard official decision, and the explanation therefore follows this pattern.<sup>20</sup>

There are two noteworthy points on the rules of automated decision-making, and both exhibit confusion about the ‘right to be explained.’

One of the more general problems is that the law has started to address the whole issue in the context of data protection, although at first glance it is

---

<sup>15</sup>Bryce Goodman and Seth Flaxman, ‘European Union regulations on algorithmic decision-making and a “right to explanation”’ (2016) arXiv:1606.08813v3 1; Bryan Casey, Ashkon Farhangi and Roland Vogl, ‘Rethinking Explainable Machines: The GDPR’s ‘Right to Explanation’ Debate and the Rise of Algorithmic Audits in Enterprise’ (2019) 34 Berkeley Technology Law Journal 143.

<sup>16</sup>ibid recital 71, italics added.

<sup>17</sup>Article 29 Working Party, or WP 29 was a working group set up under Article 29 of the Directive to deal with privacy and personal data protection issues until 25 May 2018 (entry into force of the General Data Protection Regulation). It issued hundreds of opinions, including on automated individual decisions.

<sup>18</sup>Article 29 Working Party ‘Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation’ 2016/679, WP251rev.01. Available at: <<https://ec.europa.eu/newsroom/article29/items/612053/en27>> accessed 10 December 2021.

<sup>19</sup>These are systems that usually monitor roads with cameras and other image recognizers and are then able to detect some simple traffic violations (e.g., not signalling or speeding) and the offender’s license plate and then manage the entire first-degree fines process.

<sup>20</sup>On the automatic traffic enforcement systems, see Jennifer M. Lancaster, ‘You have Got Mail: Analysis of the Constitutionality of Speeding Cameras in City of Moline Acres v. Brennan, 470 SW3D 367 (MO.2015)’ (2017) 41 Southern Illinois University Law Journal 485. 493.

quite clear that this is not a classic data protection issue. If the ultimate goal of data protection is to have clear rules on the processing of personal data (acquisition, storage, use, transfer, etc.), then the right to an explanation goes far beyond that. It would establish a special requirement for a substantive decision concerning the rights of the individual, although this decision, like any other decision, is based in part on personal data.

The other noteworthy point is that even in the context of data protection it is disputed whether the right to an explanation exists at all as a legal requirement: there is an extensive literature around it<sup>21</sup> and some authors argue very convincingly that the current data protection regime does not include a mandatory ‘explanatory statement’ rule,<sup>22</sup> while others dispute this.<sup>23</sup>

Legal and ethical concerns about big data and artificial intelligence multiplied in the mid-2010s. Countless articles,<sup>24</sup> policy documents,<sup>25</sup> codes of ethics,<sup>26</sup> legislative proposals,<sup>27</sup> and other texts began to discuss the potential risks of big data and artificial intelligence, and most of these documents did so in the context of ethics, ethical expectations, and principles. Almost all of them enshrined the requirement of explainability,<sup>28</sup> albeit in very varied

<sup>21</sup>Goodman and Flaxman (n 15) 1; Casey, Farhangi and Vogl (n 15) 145; see also Celine Castets-Renard, ‘Accountability of Algorithms in the GDPR and beyond: A European Legal Framework on Automated Decision-Making’ (2019) 30 *Fordham Intellectual Property, Media & Entertainment Law Journal* 91; Thomas D. Grant and Damon J. Wischik, ‘Show Us the Data: Privacy, Explainability, and Why the Law Can’t Have Both’ (2020) 88 *George Washington Law Review* 1350.

<sup>22</sup>Sandra Wachter, Brent Mittelstadt and Luciano Floridi, ‘Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation’ (2017) 7 *International Data Privacy Law* 76.

<sup>23</sup>Gianclaudio Malgieri and Giovanni Comand, ‘Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation’ (2017) 7 *International Data Privacy Law* 243.

<sup>24</sup>See, e.g. Joshua A. Kroll and others, ‘Accountable Algorithms’ (2017) 165 *University of Pennsylvania Law Review* 633; Cary Coglianese and David Lehr, ‘Regulating by Robot: Administrative Decision Making in the Machine-Learning Era’ (2017) 105 *Georgetown Law Journal* 1147; Andrew D. Selbst and Solon Barocas, ‘The Intuitive Appeal of Explainable Machines’ (2018) 87 *Fordham Law Review* 1085; Ashley Deeks, ‘The Judicial Demand for Explainable Artificial Intelligence’ (2019) 119 *Columbia Law Review* 1829; Katherine J. Strandburg, ‘Rulemaking and Inscrutable Automated Decision Tools’ (2019) 119 *Columbia Law Review* 1851.

<sup>25</sup>Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights Executive Office of the President May 2016 Online available: <[https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016\\_0504\\_data\\_discrimination.pdf](https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf)> Accessed 13 December 2021; Ronan Hamon, Henrik Junklewitz and José Ignacio Sanchez Martin, *Robustness and Explainability of Artificial Intelligence*, EUR 30040 EN (Publications Office of the European Union 2020), doi:10.2760/57493, Online available at: <<https://publications.jrc.ec.europa.eu/repository/handle/JRC119336>> accessed 13 December 2021.

<sup>26</sup>See the nearly 200 ethical guidelines on AI Ethics Guidelines Global Inventory <<https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/>> accessed 13 December 2021.

<sup>27</sup>see for example the recent European Artificial Intelligence Act proposal (n 8).

<sup>28</sup>A few examples include Accenture, *Responsible AI and Ethical Framework*: ‘Transparency: When complex machine learning systems have been used to make significant decisions, it may be difficult to unpick the causes behind a specific course of action. The clear explanation of machine reasoning is necessary to determine accountablity.’ Online available: <<https://www.accenture.com/gb-en/company-responsible-ai-robotics>> accessed 13 December 2021; Advisory Board on Artificial Intelligence and Human Society, Japan, *Report on Artificial Intelligence and Human Society*. ‘R&D should be conducted to develop technologies that enable people [...] to explain the processes and logics of calculations inside AI technologies’ Online available: <<https://www8.cao.go.jp/cstp/tyousakai/ai/>>



forms and contexts. As the documents contained dozens of different ethical principles, in this ‘principle proliferation’<sup>29</sup> explainability became intricately intertwined with other, related concepts and principles. By now, pretty much all we know for sure about the criterion of explainability is that it is an important ethical principle, but a number of unclear questions remain, starting with the question of whether or not it is currently part of the law. It is safe to say that the term has become semantically ‘overloaded’.<sup>30</sup>

## 2.2. The roots of semantic overload – algorithmic aversion

Confusion has arisen in several areas. It is not clear where explainability should be applied: in which area of life there should be an expectation, which part of a given system/process/decision should be explained, what is the opposite or negative state that it aims to avoid, and finally what the ultimate purpose of an explanation: that is, why it needs to be explained.

In this section, after identifying the areas where confusion has arisen and its causes, my main goal is to demonstrate that in order to better exploit this otherwise useful concept, the requirement for explainability should be kept as simple as possible. Therefore, I suggest that it should be interpreted first as a formal-procedural criterion (and not confuse substantive requirements such as fairness, justice or non-discrimination in the decision) and, secondly, distinguish it from accountability and transparency, which are closely related to it but which ultimately have different functions, and finally, thirdly, to reserve the term exclusively for machine decisions or outputs. Before I

---

[summary/aisociety\\_en.pdf23](#)> accessed 13 December 2021; *Artificial Intelligence – Australia’s Ethics Framework. A Discussion Paper: ‘Transparency & Explainability. People must be informed when an algorithm is being used that impacts them and they should be provided with information about what information the algorithm uses to make decisions.’* Online available: <[https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting\\_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf6](https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf6)> accessed 13 December 2021; High Level Expert Group on Artificial Intelligence, *Ethics guidelines for trustworthy AI*. ‘Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system’s explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an AI system has a significant impact on people’s lives, it should be possible to demand a suitable explanation of the AI system’s decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency).’ online available: <<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>> accessed 13 Dec 2021.

<sup>29</sup>Luciano Floridi and Josh Cows, ‘A Unified Framework of Five Principles for AI in Society’ (2019) 1 *Harvard Data Science Review*, 2; Luciano Floridi and others, ‘AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations’ (2018) 28 *Minds and Machines* 687. online available <<https://link.springer.com/article/10.1007/s11023-018-9482-5>> accessed 15 December 2021.

<sup>30</sup>Dwight Bolinger, ‘Semantic Overloading: A Restudy of the Verb Remind’ (1971) 47 *Language* 522.



outline the areas and manifestations of the confusion, it is worth saying a few words about why this requirement has become so important.

When addressing decisions taken by humans, although the ultimate ideal and goal is for a decision to be substantively correct and fair, we do not demand perfection in every moment and situation. If a judge follows the rules of procedure, examines the case thoroughly and carefully, and in the end makes a decision that is not entirely fair, but whose reasoning is coherent<sup>31</sup>, the society will accept even such imperfect decision.<sup>32</sup> If a doctor acts to the best of his knowledge and makes a diagnosis according to established protocols, ('customs', later 'duty of care')<sup>33</sup> it is still possible that the patient will not recover in the end, but the doctor is not usually held responsible for such unfortunate outcomes.

This is not the case with machines, where we do not seem to tolerate errors and are not satisfied with mere compliance with rules and criteria if the end result is fundamentally incorrect. We seem to expect more from machines than from humans.<sup>34</sup> Pope's famous quote<sup>35</sup> could be amended to 'to err is human, but not allowed for the machines' The phenomenon has been called 'algorithmic aversion' in the literature.<sup>36</sup>

Several factors have led to this expectation. First, algorithms are very efficient behavioural control tools, as Lessig recognised in his theory as early as 1998.<sup>37</sup> According to him, codes (algorithms) are tools for controlling behaviour in cyberspace like 'objects' ('architecture') in the physical world: in many respects they are far more effective than the law, or other social norms. Mirelle Hildebrand has recently developed a very similar theory in which she claims, that 'standard-setting, monitoring and behaviour modification'<sup>38</sup> can be performed by algorithms.

<sup>31</sup>The recently dominant theory of legal argumentation, coherence theory, rests on this assumption. See e.g. Ronald Dworkin, *Taking Rights Seriously* (HUP 1977), Ronald Dworkin, *Law's Empire* (HUP 1986), Neil MacCormick, 'Coherence in Legal Justification', in Alexander Peczenik, Lars Lindahl and Bert Van Roermund (eds.), *Theory of Legal Science* (D. Reidel Publishing 1984); Joseph Raz, 'The Relevance of Coherence' (1992) 72 Boston University Law Review 273–321.

<sup>32</sup>According to Dworkin, sometimes good procedures can lead to unfair decisions. In this case, the principle of integrity applies: in that case, we must at least be consistent. Dworkin, *Law's Empire* (n 31) 176.

<sup>33</sup>Peter Moffett and Gregory Moore, 'The Standard of Care: Legal History and Definitions: the Bad and Good News?' (2011) *The Western Journal of Emergency Medicine* 12 109.

<sup>34</sup>see e.g. John Naughton, 'To Err is Human – is that Why We Fear Machines that Can Be Made to Err Less?' *The Guardian*, 14 December 2019 online available: <<https://www.theguardian.com/commentisfree/2019/dec/14/err-is-human-why-fear-machines-made-to-err-less-algorithmic-bias>> accessed 13 December 2021.

<sup>35</sup>Alexander Pope, 'An Essay on Criticism' online available <<https://www.poetryfoundation.org/articles/69379/an-essay-on-criticism>> accessed 13 December 2021.

<sup>36</sup>Laetitia A. Renier, Marianne Schmid Mast and Anely Bekbergenova, 'To Err is Human, Not Algorithmic – Robust Reactions to Erring Algorithms Computers in Human Behaviour' (article in press) <https://doi.org/10.1016/j.chb.2021.106879>.

<sup>37</sup>Lawrence Lessig, *Code 2.0*. (Basic Books 2006) 124–125.

<sup>38</sup>Mirelle Hildebrandt, 'Algorithmic Regulation and the Rule of Law' (2018) 376 *Philosophical Transactions of the Royal Society A* 20170355.

Second, machines work with data that involve quantitative and numerical categories. The arbitrariness of such categories exacerbates the problem. This phenomenon was recognised by the Ancient Greeks and later by the greats of legal theory: it is the difference between *physis* (φύσις) and *nomos* (νόμος), the natural and the arbitrary, the ‘reason and the fiat’.<sup>39</sup> If machine learning is based on samples of data, machines themselves draw seemingly arbitrary boundaries with certain quantities ‘in mind’ and form arbitrary categories – ones that would otherwise have been drawn by man. Such arbitrary delimitations are easier to accept from a person than from a machine. Olsen, Slosser, and Hildebrandt also assume in their study that there are ‘abstract concerns’ about machine decision-making that stem from arguments of control, human dignity, and the manipulability of decision: ‘(t) he fears surrounding the adoption of ADM systems, while varied, can be broadly grouped into three categories: the argument of control, the argument of dignity, and the argument of contamination’.<sup>40</sup>

Third, there is a fear of detachment of the decision from its original values. Algorithmic decision making has no separate set of values. The rationale or control values for algorithmic decisions come from the outside world, often from the law itself, but through multiple referrals. For example, in the operation of the COMPAS software, the original values that govern the sentencing of a convict derive from the law.<sup>41</sup> However, the system made predictions of the offenders’ likelihood of reoffending based on thousands of cases, as a result of a machine learning process, which means that the decision is only controlled by the cases (or rather the data in these cases) after a while. Thus, there is a multiple (intersemiotic) translation between the rules and algorithms on the one hand, and the rules and data on the other, which is described in detail in Section 3.3.2.

Fourth, one of the main causes of algorithmic anxiety is that machine decisions can sometimes simply not be explained by common sense narratives, as there is a huge difference between ‘statistical’ and ‘common sense’ explanations. Pasquale formulates this in his seminal book in the following way:

[l]aw has begun to address this issue in the credit context, where applicants tend to get basic, very brief rationales for adverse actions. “Explanations” like “too many revolving accounts” or “time since last account opened too short” are reason codes; rather than explain what happened in a straightforward way, they simply name a factor in the decision. We know it was more

<sup>39</sup>Lon L. Fuller, ‘Reason and Fiat in Case Law’ (1946) 59 Harvard Law Review 376.

<sup>40</sup>Olsen, Slosser and Hildebrandt (n 6) 220.

<sup>41</sup>The COMPAS software played an important role in *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016). For a description and detailed critique of the software, see ProPublica (Jeff Larson and others, ‘How We Analyzed the COMPAS Recidivism Algorithm’ (2016) Pro Publica available online <<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>>, May 23 2016.

important than other, unnamed factors, but we have little sense of how the weighing went.<sup>42</sup>

### 2.3. The domains of semantic overload of explainability

Partly as a result of algorithmic aversion, the requirement for explainability and the problem of unexplained algorithms/systems/decisions has arisen in various fields and contexts in the last decade.

The earliest context for this is the big data narrative.<sup>43</sup> Within this framework, the areas of credit assessment decisions, higher education admissions, or employee evaluations have almost always come to the fore.<sup>44</sup> The infamous Target case (in which an algorithm found out that a teenager was pregnant)<sup>45</sup> was also oftentimes cited. It has frequently been stated that algorithms based on big data methodology can unfairly discriminate against individuals or manipulate them unnoticed.<sup>46</sup> Therefore, in ‘big data ethics,’ transparency has become the foremost value, i.e. the requirement that the person concerned must always be aware of the data on which the machine grouping or prediction was based. Another important element of ‘big data ethics’ is the need to give the subject of the decision the opportunity to question it – a prerequisite for which is that the decision must be justified in a comprehensible language. This narrative framework is a strong reminder of the way in which European data protection law has long been dealing with the problem, and which is set out in section 2.1.<sup>47</sup>

This initial narrative was joined by two new aspects in the mid-2010s. The big data narrative has been mostly replaced by an explanatory framework for artificial intelligence that focuses on its unpredictable, ‘black box-like’ operation and non-deterministic output.<sup>48</sup> This narrative focused on the phenomenon of systems based on machine learning. It emphasised the need to make the outcomes of AI understandable to ordinary people as well.

---

<sup>42</sup>Frank Pasquale, *Black box society* (HUP 2015) 149. and Frank Pasquale, ‘Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society’ (2017) 78 *Ohio State Law Journal* 1243.

<sup>43</sup>Richards and King (n 7).

<sup>44</sup>These are mentioned in the GDPR (Recitals 71). See also Executive Office of the President (of the USA) *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*. Online available <[https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016\\_0504\\_data\\_discrimination.pdf](https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf)> accessed 13 December 2021.

<sup>45</sup>E.g. Charles Duhigg, ‘How Companies Learn Your Secrets’ (2012) *The New York Times*, 16 Feb 2012 online available <<https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>> accessed 13 December 2021; Michael Mattioli, ‘Disclosing Big Data’ (2014) 99 *Minnesota Law Review* 535.

<sup>46</sup>Federal Trade Commission Report Provides Recommendations to Business on Growing Use of Big Data, available online <<https://www.ftc.gov/news-events/press-releases/2016/01/ftc-report-provides-recommendations-business-growing-use-big-data>> accessed 13 December 2021; Cathy O’Neil, ‘Big-Data Algorithms Are Manipulating Us All’ (2016) *Wired* 2016. 10. 10. online available <<https://www.wired.com/2016/10/big-data-algorithms-manipulating-us/>> accessed 13 December 2021.

<sup>47</sup>Directive 95/46/EK (n 11) 15(1).

<sup>48</sup>AI HLEG Ethics Guidelines (n 28) 21.

Finally, in the late 2010s, this narrative was complemented by another element, that of vulnerability to Internet platforms. The need to explain data, processes, algorithms and outputs within platforms (e.g. ranking) is one of the key recurring elements of the European Commission's latest platform regulatory package.<sup>49</sup>

All this has led to the requirement for explainability becoming a semantically overloaded concept. There are at least three aspects in which this overload manifests itself.

First, it is not clear *what* should be explained. On some occasions, the requirement applies to the entire system (processes and/or system architecture). This is the case, for example, on Internet platforms. We do not know exactly what happens to our post when we click on the post button or report a post. What systems and processes start to work at this juncture? When does a person intervene?

At other times, the requirement of explainability is applied not to the system as such but to the decisions or outputs of the system. In this context, 'explainability' means that a given output must be explainable, just as a decision maker must be able to explain its decision – this means that a proper justification is required for a particular decision.<sup>50</sup> In such cases, in line with the structure of the legal argument, the focus is on the rules which guided the decision, and what previous events and happenings were considered.

Another source of confusion is that the 'explainability' or rather the 'incomprehensibility' of an algorithmic process or decision can refer to different things in terms of what we wish to avoid. As Selbst and Barocas point out, 'inscrutability' can mean at least three different things in this regard.<sup>51</sup> First, algorithms can be inaccessible simply because they are owned by someone, and that person does not want to reveal how they work. Second, even if the operation of the algorithm is not secret, it is still possible that it is *too complicated* for some reason, and this is a 'situation in which the rules that govern decision-making are so complex, numerous, and interdependent that they defy practical inspection and resist comprehension'.<sup>52</sup> Finally, an algorithm can be "nonintuitive" when there is 'an inability

---

<sup>49</sup>See e.g., Recital 99 of the new draft DSA (Regulation of the European Parliament and of the Council on the single market for digital services (amending the Digital Services Act and Directive 2000/31 / EC) 'The Commission should be empowered to request access to and explanations relating to, databases and algorithms ...' and recital 52: 'recipients of the service should have information on the main parameters used for determining that specific advertising is to be displayed to them, providing meaningful explanations of the logic used to that end.'

<sup>50</sup>For a different answer to the 'what should be explained?' question see Hamon and others (n 25) and Bernhard Waltl and Roland Vogl, 'Explainable Artificial Intelligence – the New Frontier in Legal Informatics' (2018) *Jusletter IT* 22 Feb 2018.

<sup>51</sup>Selbst and Barocas (n 24).

<sup>52</sup>*ibid* 1094.

to weave a sensible story to account for the statistical relationships in the model<sup>53</sup>

To further complicate the situation, expecting an explanation also affects what kind of explanation will be considered acceptable and what is considered to be the purpose of the explanation. The explanation may be regarded as a guide to future behaviour or a basis for challenging the decision, but it can also serve (independently of these, but mostly together with the previous ones) to maintain the acceptance and legitimacy of the entire decision-making system, as Grant and Wischick point out in their article.<sup>54</sup> The characteristic of the law is that these objectives are present in a legal statement at the same time. It is no coincidence that the requirement of clarity and transparency is interpreted very differently by stakeholders.<sup>55</sup>

#### ***2.4. The neighbouring concepts of explainability and a proposal for simplification***

In order to resolve this semantic overload, it is advisable to separate explainability from related concepts. It would be useful to divide these related concepts into two groups: those that are more procedural and formal in nature, and those that are more content-substantive. I will outline the benefits of this dismantling below.

It is worth starting the clarification by placing explainability in the procedural-formal group in this grouping. Of the content-substantive concepts, lawfulness, fairness, non-discrimination, and autonomy often emerge as expectations closely related to explainability.<sup>56</sup> Although it is immediately apparent that these are of a different nature to explainability, the strong connection exists because, as we shall see later, explainability means not only providing *any* explanation to complement machine decisions, but also that this explanation must meet a minimum level of some substantial values. Essentially, Floridi and Cowls take the same view when they write, while clarifying the relationship between the principles, of explainability (they use the word explicability) ‘enabling the other principles’ (beneficence, non-maleficence, autonomy and justice) to prevail.<sup>57</sup>

However, this does not mean that clarity (along with transparency and accountability) does not differ radically from substantive expectations such

<sup>53</sup>ibid 1097, and similarly, distinguishing between causal and legal explanation: Olsen, Slosser and Hildebrandt (n 6) 224.

<sup>54</sup>Grant and Wischick (n 21) 1419. The author’s main argument here is to point out the tension within GDPR between privacy and explainability.

<sup>55</sup>Heike Felzmann and others, ‘Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns’ (2019) *Big Data & Society* 5.

<sup>56</sup>A good example of the confusion between substantive principles and explainability is given by Casey and others who argue that the right to explanation is existing in GDPR that it is ‘a promising new mechanism for promoting fairness’ Casey and others (n 15) 148.

<sup>57</sup>Floridi and Cowls (n 29) 8.

as autonomy and justice. The same relationship can be discovered between the kind of ‘procedural’ and substantive principles that Lon Fuller based his ‘procedural legal theory’ on. In his famous study, Fuller distinguished between the ‘inner or internal morality of law’<sup>58</sup> and the substantive goals or values that law must achieve. Even though the latter are open to debate, it would be difficult to imagine a legal system without the former (e.g. without promulgating the law). In this sense, it can be said that some kind of explanation is – quite rightly – expected for legal decisions. At this point it is worth recalling the algorithmic aversion discussed earlier: although we sometimes accept it when human decisions are not explained (e.g. the jury does not justify its decision), this would be difficult to imagine for machines.

Having drawn a distinction between explainability and the substantive criteria, a number of principles remain from which it also needs to be differentiated in some way. The four closest such concepts are transparency, accountability, interpretability, and comprehensibility (or intelligibility). These concepts can be arranged into a wide variety of different relationships. They are sometimes treated as synonyms or as complementary concepts lying on the same level.<sup>59</sup> Sometimes transparency is held to be the main principle, and accountability and explainability are its auxiliary principles.<sup>60</sup> Other interpretations give explainability priority, and accountability and transparency are placed within it.<sup>61</sup>

To clarify this, it is necessary to return to the issues raised in point 2.3. As was demonstrated, it is not clear whether comprehensibility should be interpreted as the intelligibility of an entire process or system, or only of a system output (decision or recommendation) arising from an individual case. At this point, it is clear that the structure and operation of an entire process or system would better be expressed by the notion of *transparency*. Correspondingly, when the question arises as to who is the owner of a process and who is responsible for its final operation, the principle of accountability is more relevant. Explainability involves *finding a reasonable explanation* for a machine output.<sup>62</sup> Consequently, it includes the expectation that the

---

<sup>58</sup>Lon L. Fuller, *The Morality of Law* (Yale University Press 1964).

<sup>59</sup>E.g., Amanda Thomas in her blogpost ‘(Model) Transparency and Explainability’, online available <<https://ople.ai/blog/model-transparency-and-explainability/>> accessed 20 June 2021) mentions explainability as a synonym with interpretability; In the same vein a course material published by University of Helsinki (Chapter 4: Should we know how AI works, online available <<https://ethics-of-ai.mooc.fi/chapter-4/2-what-is-transparency>> accessed 10 December 2021). Also: Mike Ananny and Kate Crawford, ‘Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability’ (2018) 20 *New Media & Society* 973.

<sup>60</sup>E.g., Joana Hois, Dimitra Theofanou-Fuelbier and Alicia-Janin Junk, ‘How to Achieve Explainability and Transparency in Human AI Interaction’ in Constantine Stephanidis (ed.), *HCI International 2019 – Posters. HCI 2019. Communications in Computer and Information Science*, vol 1033. (Springer 2019).

<sup>61</sup>Floridi and Cowls (n. 29).

<sup>62</sup>Barredo-Arrieta and others (n 1) (in general AI context) talk about ‘transparent models’ and ‘post-hoc explainability techniques for machine learning models’. I use basically the same distinction here.

decision or recommendation should give an account of the rules governing the system and the events underlying the output. It should also be clear that of the three negative states mentioned, explainability primarily seeks to address the problem of overly complex and non-intuitive outputs. The fight against secret algorithms, however, can be pursued more effectively with the requirement of transparency and/or accountability.

At the same time, some concepts are truly interchangeable with explainability. These are ‘interpretability’ and ‘explicability’. These concepts are based on the same considerations, with only slightly different emphases: while interpretability emphasises bridging the gap between technology and ordinary language, clarity emphasises the ability to incorporate it into ordinary narratives.<sup>63</sup>

It is worth noting that the Floridi and Cowls study,<sup>64</sup> cited several times above, while correctly recognising the difference between explainability and substantive principles, considers explicability to be a general principle that includes intelligibility and accountability. I would assert, however, that the three procedural principles are certainly not related to each other in this way, although there is a relationship between them. Comprehensibility (intelligibility) focuses primarily on the linguistic dimension, while accountability seeks out the ultimate human responsibility. To take an example, in the case of COMPAS recommendations, transparency means knowing in what cases and in what procedure it is used.<sup>65</sup> Accountability means that it is clear who can use it and for what purpose, and that we know who is responsible if the system makes a wrong recommendation. The criterion of explainability reflects that a clear and understandable connection can be drawn between the rules on the input side and the input data and the decision on the output decision. Finally, intelligibility means presenting the whole thing in a linguistically understandable form.

As should be obvious, the decision may meet one requirement but not necessarily all. The process can be transparent but at the same time difficult to explain. The ultimate responsibility may be lost in the process, but the whole thing can be well explained. An explanation may substantially good, while its language is difficult to understand.

---

<sup>63</sup>The concepts of interpretability and explicability are treated here as synonyms for their explainability, with the difference that as if interpretability were used more often by technicians and programmers, but in essentially the same sense as explainability. (Chris Olah and others, ‘The Building Blocks of Interpretability’, (2018) Distill available online: <<https://distill.pub/2018/building-blocks/>> accessed 13 December 2021. In the material of the High Level Expert Group (AI HLEG) (n 28), this is quite spectacular, as ethical requirements are explained throughout, but at the end of the material, where the details of the requirements of technical robustness and safety are discussed, the term interpretability is used. ‘Explainability’ also seems to emphasize the role of ordinary narratives, which will play an important role later.

<sup>64</sup>Floridi and Cowls (n 29).

<sup>65</sup>see n 39.



Adopting this narrow interpretation of explainability has two advantages. First, by applying the requirement of explainability exclusively to machine decisions and recommendations, this discourse can easily be related to the discourse of legal reasoning, which seeks answers to the following questions: Why reasoning is needed, and why it is important in the legal domain? When is an argument correct and sound? What are the specifics of reasoning in law? This discourse can be complemented in the field of machine decisions with questions such as what are the differences and commonalities between judicial and machine decisions? What can we learn from judicial reasoning and what can we use from it in the context of algorithmic explainability, and what is not applicable? The second advantage of this limitation is that it allows us to regain the linguistic dimension and make use of some linguistic insights, above all Jakobson's translation theory<sup>66</sup> and its central category, the concept of equivalence.

### 3. Legal theoretical and linguistic aspects of algorithmic explainability

#### 3.1. Characteristics of judicial reasoning

Examining the problem of explainable algorithms from a legal theory perspective ultimately leads to the problem of practical and judicial reasoning. When people make decisions, they are expected to give reasons for them, and this is especially true of law, which is a very important area of practical reasoning.<sup>67</sup> Law itself is sometimes perceived as a discursive practice based on coherent reasoning.<sup>68</sup> When machines make decisions, we expect them to be reasoned in the same coherent way: in fact, as has been noted, we expect *more* of machines than of humans, because in some cases, such as jury decisions, it is accepted that no justification will be provided.<sup>69</sup>

A legal argument rests on three pillars: the presentation of the norms, the presentation of the facts, and the demonstration that the norms have been correctly applied to the facts, and hence the legal consequence, the decision is correct.<sup>70</sup> Both are written in human language, and there is a growing demand for both to be understandable: for legal norms, this means that they must use a language that is close to ordinary language. Facts are also

<sup>66</sup>Jakobson: On translation (n 10).

<sup>67</sup>Frederick Schauer, 'Giving Reasons' (1995) 47 *Stanford Law Review* 633; Julie Dickson, 'Interpretation and Coherence in Legal Reasoning' (2016) Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (2021 summer edition), available online <<https://plato.stanford.edu/entries/legal-reas-interpret/>> accessed 13 December 2021, and Müller (n 8).

<sup>68</sup>Dworkin: *Law's Empire* (n 31).

<sup>69</sup>Schauer (n 67) 634.

<sup>70</sup>The facts, the laws, and application of laws to facts (facts, laws, and inference) are constant elements of decisions in almost every legal system. Judgments of the European Court of Human Rights, for example, use the following structure: The Facts, Relevant Legal Framework and Practice, The Law, and within the latter: The Parties' submissions, The Court's assessment. In the same vein Olsen, Slosser and Hildebrandt (n 6) 227–230.

expected to be ‘packaged’ into an easy-to-understand ‘story’.<sup>71</sup> The argument itself is good if it convinces the reader that the norm (which is either enshrined in written law or encapsulated in previous decisions) is applicable and that the norm has been carefully interpreted, the evidence has been properly considered, the facts have been thoroughly disclosed, and therefore, in individual cases, that the decision complies with the law. A very rich theoretical tradition deals with the microstructure and details of this process,<sup>72</sup> but it is based on a few simple assumptions.

- (1) The person making the decision has a certain degree of freedom, or at least room for manoeuvre, and the reasoning is partly an explanation of the solution chosen. The explanation in many legal cultures is tied to a particular person that reflects his or her personality.<sup>73</sup>
- (2) The rules by which a decision is made are written in human language and are therefore subject to human interpretation.<sup>74</sup> Interpretations are based on or embedded in the language usage of the legal community, and this community may, for example, limit or expand the meaning of a particular word in the context of the case.<sup>75</sup>
- (3) The facts are transformed into a coherent narrative. Facts are not ‘raw facts’, but more or less coherent and believable narratives that fit into ‘ordinary stories’.<sup>76</sup>
- (4) The judicial decision-making process is circular. The interpretation of the norm and the selection and presentation of the facts are made with reference to each other: the facts are interpreted using legal concepts, while the text of the law is selected and constructed in the light of the interpreted facts.<sup>77</sup>
- (5) Judicial reasoning is addressed to several target groups at once. It serves not only as a means of persuading those affected, but also to guide the future behaviour of other people. Thus, the decisions and reasoning are designed to meet the needs of multiple target groups.<sup>78</sup>

<sup>71</sup>J. Christopher Rideout, ‘Storytelling, Narrative Rationality, and Legal Persuasion’ (2008) *Legal Writing: The Journal of the Legal Writing Institute* 53.

<sup>72</sup>Robert Alexy, *A Theory of Legal Argumentation: The Theory of Rational Discourse as Theory of Legal Justification* (OUP 1989).

<sup>73</sup>Although in some legal cultures, like the French, legal decisions are presented as impersonal decisions.

<sup>74</sup>Ronald Dworkin, ‘Law as Interpretation’ (1982) 60 *Texas Law Review* 527.

<sup>75</sup>see e.g. Neil MacCormick and Robert S. Summers (eds.), *Interpreting Statutes. A Comparative Study* (Routledge 1991).

<sup>76</sup>‘They don’t just have to prove that something happened, but that what happened is believable. Proof cannot be focused on certainty, only on (maximum) probability – and the great errors of large lawsuits warn against this.’ Miklós Szabó: *Ars Iuris, A jogdogmatika alapjai* (Bibor 2005) 257.

<sup>77</sup>Wolfgang Fikentscher, *Methoden des Rechts IV*. (J.C.B. Mohr (Paul Siebeck) 1975–1977) 198 Fikentscher adapts the theory of Gadamer on the hermeneutic circle to the application of the laws to a particular case.

<sup>78</sup>Lawrence Baum, *Judges and their Audiences. A Perspective on Judicial Behaviour* (Princeton University Press, 2008) 21, 50.

- (6) Explanation is a form of ‘generalization’. Both storytelling and interpreting the laws, as well as demonstrating that the laws are being applied properly, are based on common assumptions, a commitment to higher narratives, the constant use of metaphors, and is done in a multi-value space.<sup>79</sup> In addition to the wording of the laws, there are common reasons as to which arguments and explanations are acceptable.<sup>80</sup>

### 3.2. Characteristics of machine-made decisions

If machine decisions are compared with the characteristics of the legal decision summarised in the six points above, several important differences arise.

- (1) First, the argument presupposes that another decision could have been made and the decision-maker explains why he chose that particular decision. Machines, when based on rule-based algorithms, simply do not have the discretion and ability to do so. If the decision is an open-ended process (since it is based on certain statistical contexts or machine learning processes), the explanation may, of course, give an account of this context (which may or may not be understandable), but it will not be a process of reasoning supported by a good rhetorical performance, but an account of the mathematical process that led to that result.
- (2) Second, closely related to the first point above, is that the human decision (among others) is open-ended because the rules of the law can be interpreted in different ways, since they are written in human language. Natural languages, like law, are always subject to various interpretations, and an interpretation is always debatable because it is embedded in the practice of a community – in the case of law, in the practice of the legal community. Machines have no community and do not speak the language, even if they use natural language processing tools. The interpretation of a particular legal rule which a machine applies is predetermined in the machine process and is decided when the translation of that particular rule is coded, or when the rules are created from data patterns, during the machine learning process. Therefore, the interpretation made during programming, or the code that has been created on the basis of the data patterns can be debatable, but the result of the decision is not debateable in this sense. There may be rules

---

<sup>79</sup>Kiel Brennan-Marquez, ‘Plausible Cause: Explanatory Standards in the Age of Powerful Machines’ (2017) 70 *Vanderbilt Law Review* 1249.

<sup>80</sup>Robert M. Cover, ‘The Supreme Court, 1982 Term – Foreword: Nomos and Narrative’ (1983) 97 *Harvard Law Review* 4.

written in human language by which the algorithm was programmed, but the algorithm itself is not written in human language, so it cannot be interpreted differently. In fact, there is no interpretation at all when making a decision: algorithms are ‘used’ or ‘run’.

- (3) Third, constructing facts for machines is not a storytelling process. Storytelling and fact-setting are often based on data, but the underlying ordinary story patterns are much more important than that. The data can outline a story or make it more believable, but the data alone says nothing.<sup>81</sup> Although fact-finding is not entirely absent from machine decisions, the nature and rationale for fact-finding is dramatically different. There are no facts and stories in the traditional sense, just data. The individual data is then compared by the machines to previous data or databases, and in the case of machine-learning systems, this data also partially replaces and (re)shapes the rules. It is interesting to note that there is a circularity in this process, as in a legal decision: if the system is based on machine learning, the data forms the rules (the algorithm) and then continuously shape them during operation. For example, very often value thresholds are set during the learning process based on expected outcomes. Because, as we have seen, we accept arbitrariness from humans, but not from machines, it is difficult to tolerate these determinations of quantity. It should be clear, however, that this circularity lacks the human level of abstraction that legal decisions contain – which leads on to the point that legal generalisations are embedded in our ordinary narratives.
- (4) Fourth, it follows that storytelling and the interpretation of laws take place in relation to each other, (as was described in point iv previously), as the theory of the hermeneutical circle claims.<sup>82</sup> If there is no way to change the interpretation of the norms and no room to find alternative narratives of the story, however, it lacks the dovetailing process well known from the legal argument.
- (5) Fifth, the decision of a machine cannot be forward-looking to influence future behaviour. The forward-looking element of a particular decision means that the decision-maker wants to reinforce or change a certain behaviour, or wants to avoid it in the future through the decision, in reference to certain social values. A machine decision is always based on previous data, even if it is not deterministic (or seems non-deterministic.) However, this non-deterministic feature is more the result of individualisation and is more like randomness than a normative gesture which bears in mind social values. While machines are better

---

<sup>81</sup>Caryn Devins and others, ‘The Law and Big Data’ (2017) 27 *Cornell Journal of Law and Public Policy* 357.

<sup>82</sup>Hans-Georg Gadamer, *Truth and Method* (Continuum 2006) 267 and Fikentscher (n 77).

able to ‘predict’ than humans, that does not mean that they can make wiser recommendations for people for the future.<sup>83</sup>

- (6) Finally, machines cannot use metaphors and narratives, and cannot perceive value considerations. This is a huge obstacle to them being able to explain their decisions well.<sup>84</sup> The importance of narratives has already been mentioned in relation to fact-making, but I would like to draw attention to another problem here. The text of the laws and the meta-texts behind the laws, such as political statements, texts of doctrinal science, texts of public discourse like journal articles, documents of official organisations, and countless other texts, continuously enrich the narrative universe around the law, and affect those who take legal decisions. There are no such metatexts in machine made decisions.

### 3.3. *The man-machine interaction as an intersemiotic translation*

#### 3.3.1. *What does translation theory have to do with this?*

Machines work with algorithms and data, while human decision makers work with rules and narratives expressed in natural language. Both the generation of the data and algorithms required for machine decision and the explanation of machine output therefore require translation to another signal system. First, the rules and facts of law must be transformed into data and algorithms, and then machine output must be translated back into the medium of natural language. The transposition or translation between individual signal systems has generated a rich literature that can help to understand the problem of explainable algorithms.

Jakobson’s seminal work distinguished between three kinds of translations: intralingual, interlingual, and intersemiotic. Interlingual translation is a standard translation activity that takes place between two languages. Intralingual translation means that the text is reworded or interpreted in the same language. Examples of this are paraphrase or narration. Finally, intersemiotic translation means translation between different signal systems. This takes place, for example, when text written in natural language is transformed into flag signs or pictograms (figures).<sup>85</sup>

Eco modifies this division, in that he does not address translation, especially not in the case of transformation between different sign systems, but interpretation. He distinguishes between two types of interpretation:

---

<sup>83</sup>Marquez (n 79) 1250.

<sup>84</sup>Recently the papers collected in *Narrative and Metaphor in Law* have demonstrated how important narratives and metaphors are in the law. Michael Hanne and Robert Weisberg (eds), *Narrative and Metaphor in the Law* (CUP 2018) and in a similar vein: Peter Brooks, ‘Narrative in and of the Law’ in James Phelan and Peter J. Rabinowitz (eds.), *A Companion to Narrative Theory* (Blackwell Publishing 2005).

<sup>85</sup>Jakobson (n 10) 233.

intersystemic (between different systems) and intrasystemic (within a system). An intrasystemic interpretation can be, for example, a paraphrase within a language, or a conversion within a music sign system when a piece of music is transposed into another tone. Intersystemic interpretation can again take several forms: when the interpretation essentially aims at reconstructing the original (such as a photo-reproduction of a painting, or a translation between traditional languages), or when the intention is to create a new quality: for example, during a stage or film adaptation of a novel., transitions between the two are also conceivable: for example, the recitation of a poem, where the reconstruction of the original poem and the creation of a new quality are present at the same time, or where the written word and spoken language (as two sign systems) are both used.<sup>86</sup>

The operation of law can similarly be understood as a linguistic activity, either as an intralinguistic translation (in the Jakobsonian sense) or an intrasystemic interpretation (in the Ecomian sense). A legal decision (but also the production of another legal document) is preceded by a series of translations (or interpretations):

Legal activity is linguistic activity. It is a series of transformations of linguistic expressions and texts ending up in the final event of court judgement. If we understand legal activity as a series of translations, we can separate two — otherwise intertwined — chains of translations. The first runs along the question of law, with the texts of norms in the centre [...] The other chain runs from an empirical act (event) to the decision of the court (as an event and as a text). To connect the loops of the second chain — which is not a *par excellence* legal job — is to reconstruct past events, to turn them into facts and to transform the latter into a legal state of affairs.<sup>87</sup>

In this explanatory framework, while it is important to ponder what can be regarded as the same ‘signal system’ or ‘system,’ this question is less relevant to the explainability of machine decisions because the explanation of a machine output is definitely a translation between two systems. I argue that this perspective helps to understand what a good explanation of machine decisions should look like.

### 3.3.2. Machine-made decisions. Two streams and two translations

When the explanation of machine decisions is treated as a translation between sign systems, the most striking thing is that not one, but two translations are involved (‘chain translation’<sup>88</sup> according to Jakobson) and two threads within this chain: the translation of the norms, and the facts.

<sup>86</sup>Umberto Eco ‘Traduzione e interpretazione’ [Translation and Interpretation]. Versus 85–87. 55–100 cited by Nicola Dusi, ‘Intersemiotic Translation: Theories, Problems, Analysis’ (2015) 206 *Semiotica* 181–205.

<sup>87</sup>Miklós Szabó, ‘Law as Translation’ (2004) 91 *Archiv für Rechts- und Sozialphilosophie* 65–66.

<sup>88</sup>Jakobson (n 10) 236.

Machines work with signals and quantities, data, and data patterns, so machines can only handle legal cases if those cases are translated into data by an intersemiotic translation at the outset. A machine taking decisions on cases requires a similar transposition of the legislation into algorithms beforehand to operate. In addition, the two inversions raise different issues.

The problem of translating legal rules into machine rules has long been known in the literature. Shay and her colleagues for example conducted an experiment in which they modelled a simple situation where a fine should be imposed (writing the algorithm for a speeding monitoring system mentioned several times in this study) and found that programmers interpreted even these simple rules very differently.<sup>89</sup>

It is no coincidence that this study, and another similar one,<sup>90</sup> suggest that governments should establish the applicable rules in a ‘machine-consumable’<sup>91</sup> (vs purely ‘machine readable’) form from the outset. In this case, the responsibility for translating the rules into code will be vested in governments (legislators), rather than the organisations that develop or operate legal decision-making systems. But the translation itself should be done in any case.

Turning facts into data is no less problematic a process. As Beniger pointed out, in order for modern bureaucracy – and modern law – to function effectively, it destroys or ignores information ‘in order to facilitate its processing’.<sup>92</sup> When machines are involved, the destruction is even more dramatic than in a traditional bureaucratic process, because on the one hand, machines can only work with inputs that are completely stripped of data, while on the other, they can handle much more data than a bureaucratic organisation where the ‘processors’ or ‘inference engines’ are human.

Based on the rules converted into machine code and the facts converted into processable data, the machine then makes its decision or recommendation. This output must then be translated back into human language. The process is illustrated in [Figure 1](#).

If all this is applied to a specific system, the operation of the COMPAS system already mentioned,<sup>93</sup> when the risk of recidivism is predicted from a series of 137 questions completed by the accused and from other sources (data on the perpetrator’s crime and hundreds of past cases), then the legal proposition that the ‘risk of recidivism is high’ is actually translated

---

<sup>89</sup>Liza A. Shay and others, ‘Do Robots Dream of Electric Laws? An Experiment in the Law as Algorithm’ in Ryan Calo, Michael Froomkin and Ian Kerr (eds.), *Robot Law* (Edward Elgar 2016) 274–305.

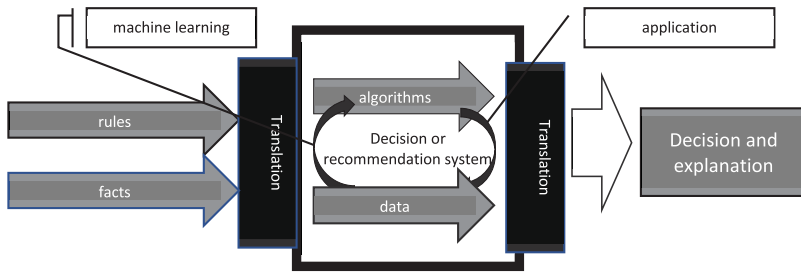
<sup>90</sup>James Mohun and Alex Roberts, *Cracking the Code. Rulemaking for humans and machines* OECD Working Papers on Public Governance No. 42 (OECD 2020) online available <[https://www.oecd-ilibrary.org/governance/cracking-the-code\\_3afe6ba5-en](https://www.oecd-ilibrary.org/governance/cracking-the-code_3afe6ba5-en)> accessed 14 December 2021.

<sup>91</sup>*ibid* 18.

<sup>92</sup>James Beniger, *The Control Revolution, Technological and Economic Origins of the Information Society* (HUP 1986) 15.

<sup>93</sup>For the description of COMPAS see (n 41).





**Figure 1.** The process of automated decision making in law.

into algorithms, and the factual elements of the perpetrator, the crime, and other circumstances are also translated into data.<sup>94</sup> The problem with COMPAS was that the translation process was not transparent, and later proved to be biased.

### 3.3.3. *The problem of equivalence*

The central issue of translation theory is the issue of equivalence. Explainability can also be understood as a question of equivalence.

Equivalence, in inter- and intralingual translations, means that the translation contains two equivalent *messages* in two different codes, according to Jakobson,<sup>95</sup> or in other words, that the source and target language items are ‘interchangeable in a given situation’.<sup>96</sup> As such, it is a pragmatic problem. Others state more directly that the quality of a translation is affected by whom it is made for. There is no objectively good translation, only a translation appropriate for one person or another (for the target group).<sup>97</sup>

As we have seen above in Eco’s division, in both intrasystemic (within a signal system) and intersystemic (between signal systems) interpretations, an important consideration is whether the translator’s intention is to create an entirely new object, or to faithfully reproduce the source material. With machine systems, there can be no question of creativity or the creation of a new quality different from the original. In fact, machine recommendation or decision systems should work as if they were not there, as if the decision had been made by a human.<sup>98</sup>

How, then, should equivalence be defined for intersemiotic translation? Jakobson, as Dusi points out, used a separate word for this activity – he called it a transmutation.<sup>99</sup> When translating a message or information

<sup>94</sup>The questionnaire is online available <<https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html>> accessed 14 December 2021.

<sup>95</sup>Jakobson (n 10) 233.

<sup>96</sup>J.C. Catford, *A Linguistic Theory of Translation* (OUP 1965) 49.

<sup>97</sup>Eugene A. Nida and Charles Taber, *The Theory and Practice of Translation* (Brill 2003) 1.

<sup>98</sup>Olsen, Slosser and Hildebrandt (n 6) 234.

<sup>99</sup>Dusi (n 86) 182.

into another sign system there are also theoretical extremes. While turning text into Morse code clearly results in almost complete equivalence, adapting a novel into a film is more of a re-creation. Describing the same novel in a high school textbook lies perhaps somewhere in between the two, insofar as it allows less creative freedom but at the same time involves a huge loss of information.

In this context, another question is what the output should be equivalent to: the initial rules and principles, or the data and processes of the algorithmic phase. I would again note that this results in a double, or chain translation, that can completely distort the original message. The explainability problem of machine decisions seems to be related only to the second translation at first sight: how machine outputs be can interpreted. In reality, however, this is not the case: and in terms of the understanding, and legitimacy of machine-made decisions, this is a cardinal question. The explanation must not only describe the relationship of output to machine operation ('what happened inside the machine'), but more importantly, *its connection to the original rules and facts*. It should also demonstrate that the initial facts have been equivalently translated to data, and the initial rules have been equivalently translated to algorithms. A good explanation therefore refers to the initial rules and facts, their equivalents in data and code, and convincingly demonstrates that the transmutations has been made in a way, that the content is equivalent.

### 3.4. What should the explanation for machine decisions be?

While some have suggested that machines should not enter certain decision areas at all,<sup>100</sup> others have argued for a new kind of algorithmic accountability, highlighting the different characteristics of black boxes (algorithmic systems).<sup>101</sup>

Two options seem to suggest themselves. The first is to simply ban machines from making decisions about people's rights. While this is tempting opportunity, it cannot be taken seriously here, and it may be too late for this in any case. The other possibility is that, although machines cannot, at first sight, justify their decisions in a traditional, 'legal way', it is possible to use methods to translate their decisions into a common, understandable language. It remains to be determined what exactly this translation would look like. What is a good rationale for a machine decision? What form should explanations for machine decisions take? Of course, I cannot give an exhaustive answer to these questions in this paper, but I will try to formulate some simple principles as a starting point.

---

<sup>100</sup>Marquez (n 79) 1280.

<sup>101</sup>Ananny and Crawford (n 59) 12.

First, it has become clear from examining the translation theory approach above that what matters is not only what happened in the machine, but also what data was entered into the machine and the broader institutional background in which the machine decision took place. This being so, until the use of machine decisions becomes widespread, the institutional and value context of each decision must be given: the organisation, or anyone who uses algorithms to decide matters, must explain the broader context of the decision. This requirement therefore entails, in fact, that transparency is very closely linked to explainability. What was the reason for introducing the algorithmic decision? What was the predecessor of the algorithmic procedure? What are the values and interests that guide the decisions?

Second, an account must be given of the rules that guided the decision, as well as of the translation of those rules into an algorithm. It is not enough to refer to the first form of the rule laid down in law, but it is also necessary to describe how it became an algorithm. How did the intersemiotic translation<sup>102</sup> work? If the above-mentioned ‘dual legislation’ is implemented, i.e. governments create a ‘machine-consumable’ version of the legislation,<sup>103</sup> it should not be the responsibility of individual system operators but of governments to explain this: for example in an explanatory memorandum attached to the legislation together with the relevant codes.

Third, the statement of reasons must contain a description of the facts and, again, a translation of those facts into the data on the basis of which the decision was taken. This would include both data collected about the subject and the external data on which the decision is based. The data (examples, cases, previous decisions) should also be provided and characterised. It should be noted that it is not necessary and probably not possible to list all the data (although this data must be provided on request).

Finally, the decision-making process needs to be explained in a step-by-step manner. One of the most important parts of the rationale is what the ‘explainable AI’ literature deals with extensively: demonstrating the learning processes of self-learning systems. This is the part of the justification that will be the least similar to traditional legal justification. This is because the rules and the data patterns that form the basis of the (often seemingly arbitrary) categorizations need to be described there, which also means that these data must be extracted in some way from systems based on machine learning.

This solution is in many points similar to the one proposed by Olsen, Slosser and Hildebrand.<sup>104</sup> However they go even further, proposing that

---

<sup>102</sup>For example, it is well known that some tolerance is built into speed control systems and speeding is generally allowed by 10%. In this case, the driver should know that he can actually drive 66 km/h at sign 60.

<sup>103</sup>see Mohun and Roberts (n 90).

<sup>104</sup>Olsen, Slosser and Hildebrandt (n 6).

the administrative decisions should be split up to two loads, where one is fed to an algorithm, and the other is fed to a human team, and ‘final decisions are pooled and used to regularly update the algorithm used.’<sup>105</sup>

#### 4. Conclusions

Machine-made decisions and recommendations have many benefits. Machines have no hidden intentions, biases, or agendas, operate extremely rationally, and have no data-processing limitations. At the same time, a neuralgic point of this decision-making is that machines cannot properly explain and justify their decisions.

In the last decade, the literature on algorithmic explainability, and on its subdomain explainability of machine-made legal decisions has grown exponentially. In part, this is why the concept of explainability is almost inextricably entangled with other ethical expectations such as transparency, fairness, non-bias, or accountability. In this article, I call this phenomenon ‘semantic overload’ and attempt to resolve it. I argue that explainability is more of a formal-procedural principle, which must first be distinguished from substantive requirements such as fairness or non-discrimination. Second, it is worth separating it from two other formal-procedural principles, namely transparency and accountability. I recommend to use the former to larger processes and systems, and the latter to the ultimate human responsibility for machine-made decisions. Explainability should be used as a measure of the rationale for machine outputs (decisions and recommendations).

If we do so, two theoretical perspectives will open up for us. The second part of the study addresses these two theoretical perspectives. On the one hand, it places the explanation of machine made decisions in the context of legal reasoning. It identifies six areas where machine-made decisions are different from human ones. Machines have no discretion in setting up the facts and interpreting the rules, their decision is not based on narratives, they are not interpreting the rules and the facts in relation to each other, they do not make decision in a forward looking way, consider higher values, and speak to multiple audiences, as human legal decision makers do. The study argues, however that despite of all these differences, the structure of the justification for legal decisions, should be maintained in case of machine-made decisions. The justification of a legal decision is based on three pillars: it should present the facts that the decision rests, the law applied and to give an account that the law has been correctly applied to the facts.

The last part of the study argues that the problem of machine made decisions and the explanation of these decisions can be perceived as a

---

<sup>105</sup>ibid 232.

translation-theory issue, because in an ADM process multiple intersemiotic (or intersystemic) translations are happening. The first intersemiotic translation occurs when we create data and algorithms from facts and rules for the machine, and another one, when we translate machine outputs back into human language. The explainability problem of machine decisions seems to be related only to the second translation, but this is not the case: the explanation must not only describe the relationship of output to machine operation ('what happened inside the machine'), but more importantly, its connection to the original rules and facts. It should also demonstrate that the initial facts have been equivalently translated to data, and the initial rules have been equivalently translated to codes. A good explanation therefore refers to the initial rules and facts, their equivalents in data and code, and convincingly demonstrates that the transmutations has been made in a way, that the contents are equivalent.

### Disclosure statement

No potential conflict of interest was reported by the author(s).

### ORCID

Zsolt Zódi  <http://orcid.org/0000-0003-3978-5493>