AUTOMATED ADMINISTRATIVE DECISION-MAKING: WHAT IS THE BLACK BOX HIDING?[1]

JAN NEŠPOR

Abstract: The exploration of the "black box" phenomenon underscores opacity challenges in automated administrative decision-making systems, prompting a discussion on the paradox of transparency. Advocating for the concept of "qualified transparency", the article aims to navigate the delicate balance between understanding and safeguarding sensitive information. Ethical imperatives, including respect for human autonomy, harm prevention, fairness, and explicability, are considered, culminating in recommendations for human participation, ethicality or accountability by design considerations, and the implementation of regulatory sandboxes to test such models prior to broad integration. Ultimately, the article advocates for a comprehensive discourse on transitioning from a human-centric to an automated public administration model, acknowledging the complexity and potential risks involved.

Keywords: automated administrative decision-making; artificial intelligence; transparency

DOI: 10.14712/23366478.2024.23

1. INTRODUCTION[2]

The introduction of artificial intelligence (AI) into decisions made through automated means with little or no human involvement also known as automated decision-making raised several issues. On one hand, the potential of AI is to evaluate the inputs and all variables to make decisions in complex situations and thus enabling the decision-makers (public administration bodies for the purpose of this article) to make faster and more consistent decisions. On the other hand, the delegation of decision-making power to an algorithm (AI) challenges the legality and the very essence of public administration's decision-making process.

This article delves into prevalent issues associated with the utilization of AI, particularly in the context of automated administrative decision-making (AADM), highlighting concerns such as the lack of transparency or explicability also known as the black box phenomenon. The aim of this article is to provide a comprehensive understanding of

[1] This article was written under the umbrella of the project "Regulatory Sandboxes: Mirage and Reality in Public Law", supported by Charles University's 4EU+ Mini-grants Programme.

[2] The AI generative language model Chat GPT was used in the processing of the paper for the purpose of text proofreading. E-mail: nesporjan@prf.cuni.cz.

the primary challenges arising from the black box while simultaneously suggesting solutions which may minimize its negative aspects.

The first part emphasizes the diverse spectrum of AADM and analyses its complexities which may arise when integrated. Further it provides a showcase of legal perspectives from different countries, ethical considerations, and the challenges posed by the black box phenomenon.

The second part discusses the challenges related to transparency, explicability, and justification in AADM, emphasizing the need for careful examination and regulatory considerations. This part also delves into the balance between the need for transparency and the protection of legally safeguarded interests.

And finally, the third part explores ethical and legal imperatives imposed on trustworthy AADM introducing concept of transparency ensuring fulfilment of said imperatives and security of safeguarded interest. Further it proposes solutions potentially ensuring fairness and mitigation of risks concerned with the evolving nature of machine learning AADM models.

2. UNDERSTANDING THE COMPLEXITIES OF AUTOMATED ADMINISTRATOVE DECISION-MAKING IN THE 21ST CENTURY

The omnipresence of AI throughout the 21st century naturally permeates the public sector and more specifically the public administration. This goes hand in hand with the ever more present digital tools and solutions enhancing the effectivity of public administration often driven by the public demand or the demand of the public administration from within. Such process is often described as a digitalisation of public administration whilst automatization may be just a small part of such.

The use of AI changed the process of automatization dramatically. One of the reasons might be the change of society's understanding of what AI actually is from Alan Turing's code breaking machine invented during the Second World War[3] to today's large language models[4] or deep learning technologies[5]. The more enhanced the AI is the

[3] HAENLEIN, M. – KAPLAN, A. A Brief History of Artificial Intelligence: on the Past, Present, and Future of Artificial Intelligence. *California Management Review* [online]. 2019, Vol. 61, No. 4, pp. 5–14 [cit. 2024-02-26]. Available at: https://doi.org/10.1177/0008125619864925.

[4] *"[…] models are trained on massive amounts of text data and are able to generate human-like text, answer questions, and complete other language-related tasks with high accuracy"* (see KASNECI, E. et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* [online]. 2023, Vol. 103, p. 102274 [cit. 2024-02-26]. Available at: https://doi.org/10.1016/j.lindif.2023.102274.

[5] *"Often known as deep neural network, deep learning consists of many layers with a number of neurons in each layer. Such layers may range from a few to thousands, and each layer may contain thousands of neurons (processing unit) in addition. Multiplying the input values with the allocated weight to each input and summing up the result are the simplest process in a neuron. This result will be further scrutinized by the activation function."* (see PATIL, T. et al. A Review on Basic Deep Learning Technologies and Applications. In: KOTECHA, K. et al. (eds.). *Data Science and Intelligent Applications* [online]. Lecture Notes on Data Engineering and Communications Technologies, Vol. 52. Singapore: Springer, 2021 [cit. 2024-02-26]. Available at: https://doi.org/10.1007/978-981-15-4474-3_61.

more the number of AI public administration and/or automated system use cases offers. This does not come as a profound provocative theory but rather a simple course of things due to the impact of technological developments.

However, the simplicity of such outcomes should not divert attention from the profound influence that contemporary technologies, particularly advanced AI, exert on the public administration's decision-making and the role of human-centric systems (such as public administration) within today's society.

As Olsen et al. stresses, the automated (administrative) decision-making (AADM) comes in a wide range of formats.[6] Rather than one specific model or a solution, the AADM is more of a spectrum of solutions for the automatization of administrative decision-making processes.

Understanding the AADM spectrum involves considering a set of variables that simultaneously alter its structure and limits. When examining the spectrum in terms of AADM complexity, one end can be defined by the straightforward decision tree mechanisms, while the other end could be characterized by deep neural networks. Whilst decision trees are models where the automated decision is made by simple "if this then that" algorithm, operating in a sort of binary or linear decision mechanism,[7] the deep neural networks is "brain inspired" model of machine learning algorithm operating in a non-linear way. In simple words, machine learning is model of AI which allows the algorithm to gradually improve its accuracy by learning (the similar way humans do) and deep neural network is a type of machine learning that simulates the functioning of a human brain by transporting input data between multiple layers and units (neurons) each time weighting the transferred information.[8]

Another important differentiation is by AADM model's application in the actual administrative decision-making process. The question lies whether the AADM model is used only for one part of the process (i.e., delivering the final decision) or if it's used for the whole administrative procedure from the initiation, gathering of the factual information up until the final decision.[9]

Both the complexity of AADM and the range of parts of decision-making procedures, where the AADM is deployed are subject to questions of technical character of

[6] OLSEN, H. P. et al. *What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration*. iCourts Working Paper Series, No. 162. University of Copenhagen, Faculty of Law, 2019, p. 9.

[7] HILDEBRANDT, M. Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society A* [online]. 2018, Vol. 376, No. 2128 [cit. 2024-02-26]. Available at: https://doi.org/10.1098/rsta.2017.0355.

[8] See DENG, L. – YU, D. Deep Learning: Methods and Applications. *Foundations and Trends in Signal Processing* [online]. 2014, Vol. 7, No. 3–4, pp. 197–387 [cit. 2024-02-26]. Available at: http://dx.doi.org/10.1561/2000000039; or AOUICHAOUI, A. R. N. et al. Comparison of Group-Contribution and Machine Learning-based Property Prediction Models with Uncertainty Quantification. *Computer Aided Chemical Engineering* [online]. 2021, Vol. 50, pp. 755–760 [cit. 2024-02-26]. Available at: https://doi.org/10.1016/B978-0-323-88506-5.50118-2.

[9] As Hofmann states, at the beginning of last year, there was not an AADM model deployed, that would cover the entire administrative procedure's cycle, from initiation to implementation of measures (see HOFMANN, H. C. H. Comparative Law of Public Automated Decision-Making. An Outline. *Rivista Interdisciplinare sul Diritto delle Amministrazioni Pubbliche* [online]. 2023, No. 1, pp. 1–12 [cit. 2024-02-26]. Available at: https://doi.org/10.13130/2723-9195/2023-1-3.

the AI. From a legal perspective and based on the regulation from some of the European countries, some of the often legally addressed issues are associated with other variables.

Kischel summarizes that Germany's jurisprudence and literature widely advocates for the more detailed explanation the wider the discretion of an administrator is.[10] Germany adopted a provision in Administrative Procedural Act (*Verwaltungsverfahrensgesetz*) proclaiming, that an administrative decision may be adopted entirely automatically, but only in cases without any room for discretion or assessment,[11] thus highlighting the risk of inability of AADM to sufficiently assess all relevant circumstances and in accordance to exercise discretional powers.

The French Code of Relations Between the Public and the Administration (*Code des relations entre le public et l'administration*) allows administrative individual decision to be adopted automatically, however under the condition, that the subject of such decision is informed about the automatic nature and the purpose of such automatization.[12] Furthermore this subject also has the right to be provided with the information about the (i) degree and method of this automation, (ii) the data processed and their sources, or (iii) processing parameters.[13] Some nations like Sweden adopted only the possibility to deliver administrative decisions by automatic means, however without any specific provisions as France or Germany did.[14] Choosing a more technological neutral approach while leaving the regulation of AADM to the general norms of administrative law. Sweden's approach corresponds with Olsen et al. who argue that the AADM should not be a subject to different legal standards than solely human decisions, as this is further addressed in the third part of this article.[15]

The provided overview of some nation's regulation of AADM is just a showcase of a multiple approaches. Whilst some nations tend to regulate the most essential risks such as the lack of discretion, possible infringement of data privacy and/or lack of transparency, others rely on general norms that ensure the legality of decisions both with and without the use of AADM models.

Some authors argue that the use of AI in decision-making might be from a deontological point of view unethical because of the inefficiency to identify uniqueness and of potentially causing harm to its subjects.[16] The deontological logic follows an approach that some things are either good or bad disregard of what the actual outcomes are. I argue that this might create problems mainly towards the general trust of public

[10] KISCHEL, U. *Die Begründung: Zur Erläuterung Staatlicher Entscheidungen Gegenüber Dem Bürger*. Tübingen: Mohr Siebeck, 2003, pp. 223–224.

[11] Article 35a of Administrative Procedure Act (Germany) in the version published on 23 January 2003, as amended.

[12] Articles L311-3-1 and R311-3-1 of Code of Relations Between the Public and the Administration (France) in the version published on 1 January 2016, as amended.

[13] Ibid., Article R311-3-2.

[14] Section 28 of the Administrative Procedure Act (Sweden): *"A decision can be made by an officer on their own or by several jointly or be made automatically…"*

[15] OLSEN et al., *c. d.*, p. 6.

[16] YAN, C. et al. When the Automated fire Backfires: the Adoption of Algorithm-based HR Decision-making Could Induce Consumer's Unfavourable Ethicality Inferences of the Company. *Journal of Business Ethics* [online]. 2023 [cit. 2024-02-24]. Available at: https://link.springer.com/article/10.1007/s10551-023-05351-x.

administrations addressee towards the validity of automatically decided matter. However, the analysis of the relationship between the automatization of public administration decision-making and the trust of its addressees is not aim of this article but rather a one of the emphasized challenges of AADM.

The German example of AADM's regulation shows a concern that the absence of discretion when using an AADM model might severely affect the legality and the factuality of a given decision. As Henman emphasizes, a great number of administrative areas require human decision-makers to exercise discretion and to personalize the decision in complex situations. Therefore, the lack of discretion raises concerns about the appropriate consideration of all variables in an administrative decision-making.[17]

While the ethical and discretionary concerns when using AADM, as well as the emphasis on the security and protection of personal data are substantiated, this article predominantly centres on the phenomenon of the black box in AADM. This term was used by Pasquale as a metaphor to describe a system *"whose workings are mysterious; we can observe its inputs and outputs, but we cannot tell how one becomes the other"*.[18]

Given the sometimes-complex nature of AADM based on neural networks, deep neural networks, or any other advanced model of AI, the inherent opacity challenges fundamental principles of administrative decision such as the transparency, explicability, and/or accountability.[19]

3. THE PARADOX OF TRANSPARENCY: BALANCE BETWEEN JUSTIFICATION AND PROTECTION

Naturally the range of opacity surrounding the AADM differs based on the complexity of the AI model used. This is sometimes addressed by Dyson as the "third law of artificial intelligence" claiming that: *"Any system simple enough to be understandable will not be complicated enough to behave intelligently, while any system complicated enough to behave intelligently will be too complicated to understand."*[20] Dyson also advocates that the relationships between AI and humans are rather a matter of faith than of proof, meaning that humans are perfectly capable with using or being a subject to things, which we cannot understand. Innerarity argues that opacity and invisibility are not an epistemic anomaly, but they are part of daily life and that using only fully comprehensible mechanisms limits benefits of any technology.[21] Take a human brain for example. After hundreds of years, the neuroscience still has difficulties

[17] HENMAN, P. Improving public services using artificial intelligence: possibilities, pitfalls, governance. *Asia Pacific Journal of Public Administration* [online]. 2020, Vol. 42, No. 4, pp. 209–221 [cit. 2024-02-24]. Available at: https://doi.org/10.1080/23276665.2020.1816188.

[18] PASQUALE, F. *The Black Box Society*. Cambridge: Harvard University Press, 2015, p. 3.

[19] For the sake of simplicity, the advanced AI-based AADM model will henceforth be denoted simply as the AADM or AADM model.

[20] DYSON, G. The Third Law. In: BROCKMAN, J. (ed.). *Possible minds: 25 Ways of looking at AI*. New York: Penguin Press, 2019.

[21] INNERARITY, D. Making the black box society transparent. *AI & Society* [online]. 2021, Vol. 36, pp. 975–981 [cit. 2024-02-26]. Available at: https://doi.org/10.1007/s00146-020-01130-8.

to answer simple question “how does a human brain work?”. And yet, human brain has demonstrated its capacity to discover penicillin, achieve manned moon landings, or establish a set of universally accepted rules, commonly referred to as the law.

A full comprehension of AADM inherits the need to open the black box and provide full transparency. Yet, supposing one were to argue for the contrary stance, insisting on complete transparency for AADM, asserting that every individual should grasp “how the inputs transform into outputs” or “how the data inputted into the AADM model translates into the administrative decision”, the issue of transparency necessitates a more comprehensive evaluation.

Hamon et al. refers to transparency in terms of AI as to *“possibility to have a complete view on a system, i.e., all aspects are visible and can be scrutinised for analysis”*.[22] In this sense the transparency is further distinguished into three levels: (i) the *transparency of implementation*, meaning that the technical parameters and principles of the model are known, and the outcomes are therefore predictable;[23] (ii) the *transparency of specifications* referring to the knowledge of task, objectives or context of the model as to the training dataset and training procedure; and (iii) the *transparency of interpretability* provides a general understanding of the logic behind an AADM model and provides sufficient reasoning.[24]

To offer a more pragmatic perspective, does full transparency allow the subjects of AADM to fully comprehend how the model reached the final decision? Can complete transparency potentially compromise other legally protected interests? To answer such provocative questions, the question and where or why does it exist needs to be subject to an examination.

Burrell, Lepri et al., or Innerarity concur that opacity does not inherently imply wrongdoing and, in certain instances, may have justified reasons explaining why it exists and/or that there are safeguards against potential threats. As to provide safeguards or to protect other legally protected interests or certain elements of AADM model, such opacity is referred to as *intentional* or *deliberate*. Created or existing due to legitimate concerns such as the will to protect one’s intellectual property, state secrecy, personal information, or sometimes details and information of which disclosure is limited.[25] However this does not come always as a persuasive argument. A more compelling argument is provided by Lepri et al. who suggest that the open source-ness of AADM might lead to risk of “gaming the system” meaning, that subject of AADM would have the advantage to provide information in a certain way allowing them to get the desired

[22] HAMON, R. et al. *Robustness and Explainability of Artificial Intelligence* [online]. Luxembourg: Publications Office of the European Union, 2020, p. 11 [cit. 2024-02-26]. Available at: https://doi.org/10.2760/57493.

[23] Such model is also known as “white-box model” as opposite to “black-box model”, which is subject of this article.

[24] HAMON et al., *c. d.*, pp. 11–12.

[25] See BURRELL, J. How the machine ‘thinks’: understanding opacity in machine learning algorithms. *Big Data & Society* [online]. 2016, Vol. 3, No. 1 [cit. 2024-02-26]. Available at: https://doi.org/10.1177/2053951715622512; and LEPRI, B. et al. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philos. Technol* [online]. 2018, Vol. 31, pp. 611–627 [cit. 2024-02-26]. Available at: https://doi.org/10.1007/s13347-017-0279-x.

form of a decision, therefore leaving no discretion for the public administration whatsoever.[26] Compared to human decision-making, the "gaming of a system" corresponds to manipulating or even bribing of the human administrator. That is of course forbidden, however a prohibition of "gaming the AADM" might not be as effective, since is less visible and harder to prove than human manipulation and/or bribery.

Due to the natural lack of knowledge about algorithms, AI or data science, transparency does not enable the general public to comprehend the decision-making process of an AADM model. Due to this fact, the second type of opacity is addressed as *illiterate*[27] or more precisely as an *objective*[28] transparency. Thus, even a fully transparent AADM does not ensure its comprehension by public. The understanding and interpretation of AADM models is up to educated few, but in similar sense, so is the law.

Finally, the third type of transparency – *intrinsic* or *emerging* is caused by the scalability of machine learning AI and emerging unpredictability and unintentionality. As Lepri et al. concludes this may be tackled using alternative easy to interpret machine learning models but with the inherent disadvantage of lower precision.[29] Hence, the overarching inquiry revolves around the deliberation on whether to opt for a model that prioritizes transparency at the potential cost of accuracy, or conversely, a less transparent (black box) model that may exhibit enhanced accuracy. The problem with AADM that is complicated enough to cause an *emerging opacity* is usually referred to as the *interpretability problem*. From a legal perspective, there is a need to distinguish between and interpretable AI and explainable AI. Whereas the interpretability is defined as a *"level of understanding how the underlying (AI) technology works"*[30] the explicability is the *"level of understanding how the AI-based system* [...] *came up with a given result"*.[31] Having the differentiation in mind, does the lack of interpretability raise a relevant concern at all?

European regulation found its answer in putting the emphasis on explicability rather on full transparency or interpretability. This can be demonstrated by the provisions adapted in Article 15(1)(h) together with Article 22(1) & (4) of the General Data Protection Act[32] (GDPR). Under this provision, the data subject has the right to receive information regarding any automated decision-making process of its data, including the underlying logic behind the decision, along with details about its significance and envisaged consequences, thus granting the data subject "a right to an explanation".[33]

[26] LEPRI et al., *c. d.*

[27] BURRELL, *c. d.*; and LEPRI et al., *c. d.*

[28] INNERARITY, *c. d.*

[29] LEPRI et al., *c. d.*

[30] International Organization for Standardization & International Electrotechnical Commission. Software and Systems engineering – Software testing – Part 11: Guidelines on the testing of AI-based systems. 2020, Art. 4(1)(42).

[31] Ibid., Art. 3(1)(31).

[32] 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).

[33] See GOODMAN, B. – FLAXMAN, S. EU Regulations on Algorithmic Decision Making and a "Right to Explanation". *AI Magazine* [online]. 2016, Vol. 38, No. 3, pp. 3–112 [cit. 2024-02-26]. Available at: https://doi.org/10.1609/aimag.v38i3.2741.

Innerarity does not see the right to an explanation as "autopsy of the system" but instead as a principle of self-control that lessens the knowledge gap between the subject of an AADM and its developer. Similar argument is argued by Doshi-Velez et al. framing the AI's explicability as the interpretable depiction of a process, where a decision-maker reached to a particular conclusion based on a particular set of inputs.[34] This underscores the distinction between interpretability and explicability, as mentioned above.

From a different standpoint Hildebrandt clarifies that explanation is not necessarily justification and claiming that *"knowing how the algorithm came to its conclusion does not imply that the conclusion is 'in accordance with the law'"*.[35] A case where a black box provides a superficial explanation with no justification whatsoever Pasquale calls a *"mere façade of an explanation"*.[36] It is my believe that definition provided by Doshi-Valez et al. disregards the opacity between inputs and outputs and thus leaving space for doubts regarding the lack of bias or fairness of the AADM.

The GDPR regulates only a right to an explanation in connection to any data processing which applies in several cases including, but not limited to, administrative decision-making. But in terms of AADM the law mandates the consideration of additional legal requirements for administrative decisions, irrespective of whether they are automated. On the European level such prerequisites are set forth by the Charter of Fundamental Rights of the European Union (CFR) or by Treaty on the Functioning of the European Union (TFEU).

The CFR stipulates in Article 41 the right to good administration containing the obligation of the administration to give reasons for its decision and the Article 296 of TFEU enshrines a duty to give reasons upon which a legal act was based. Court of Justice of the European Union (CJEU) ruled in this matter stating that *"the duty to give reasons which is justified in particular by the need for the Court to be able to exercise judicial review, must apply to all acts which may be the subject of an action for annulment"*.[37]

The CJEU here at once provided the logic behind the emphasis on explicability and/or justification. The crucial aspect, when addressing AADM, is the ability of any subject to contest such decision in an administrative or judicial proceedings.[38] Thus, mandating for dismantling of the façade of an explanation and for the provision of logical and legal justification in a manner that may be subject to challenge, as prescribed by the law. The comprehensive regulation governing "legal acts making", as per the cited CJEU case law, including administrative decisions, incorporates safeguards to necessitate a justification rather than a superficial explanation. Needless to say, that while this regulation

[34] DOSHI-VELEZ, F. et al. *Accountability of AI Under the Law: the Role of Explanation*. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working, 2019.

[35] HILDEBRANDT, *c. d.*

[36] PASQUALE, *c. d.*, p. 142.

[37] Judgment of the Court of the CJEU (Second Chamber), Case C-370/07, *Commission of the European Communities v. Council of the European Union*, ECR I-08917, recital 42.

[38] Judgment of the General Court of the CJEU (Eight Chamber), Case T-181/08, *Pye Phyo Tay Za v. Council of the European Union*, ECR II-01965, recital 94.

(e.g., CFR or TFEU) doesn't guarantee justification for every automated decision-making process, it does ensure it for the AADM.[39]

While justification allows a subject of an automated administrative decision to contest it, it does not provide a clear explanation of the accountability question. Pasquale claims that *"without transparency, accountability is impossible"*,[40] however full transparency does not guarantee accountability in all cases.[41] When talking about AADM, the discussion has to acknowledge the implications a decision (especially a wrong one) might have on its subjects. What if the AADM model adopted a wrong decision causing harm to its subject? Who is accountable? Is it the developer or the administrator who was overseeing the administrative process? What if the AADM is without human intervention, the so called "human outside of the loop AADM"? As Motzfeld summarised succinctly *"you cannot put R2D2*[42] *in a jail"*, addressing that an inhuman subject, an algorithm in this case, does not suffer from consequences if accountable and fear of sanction is therefore futile.[43] In a given case, it is possible to hold the public administration fully accountable, possibly with a right of recourse against the administrator responsible for the decision-making process, perhaps against the administrator who was present in the office. However, I share Pasquale's view[44] that this approach is impractical and would place unreasonable demands on the administrators involved.

The question of AADM accountability is even clearer considering cases of AADM either trained on "bad data" or progressively exhibiting biased or otherwise flawed decisions. Instances of automated decisions that exhibited bias against women[45] or Afro-Americans[46] leading to unfair outcomes, underscore the pressing need for transparency, at least in terms of *implementation transparency*. Even though it is fair to acknowledge the remark of Henman, that treating cases differently does not necessarily have to be a discrimination but rather a form of personalization,[47] such considerations should be a subject of broader ethical and human rights considerations, which are beyond the scope of this article.

As I argued above, an undue emphasis on transparency may in some cases compromise other legally protected interests such as the protection of sensitive information

[39] At least in the European Union.

[40] PASQUALE, *c. d.*, p. 175.

[41] LEPRI et al., *c. d.*

[42] R2D2 is a fictional character (droid or a robot and therefore an artificial existence with developed intelligence) created by George Lucas for the STAR WARSTM franchise.

[43] Hanne Marie Motzfeld is a professor at Faculty of Law Research Centres at the University of Copenhagen in Denmark, said rephrased statement is derived from our discussion, and the ultimate phrasing was contingent upon her endorsement.

[44] "[…] *full transparency of federal agency actions— let alone the actions of private firms— is far off. Too many regulators are underfunded, overworked, or angling for lucrative jobs from the very firms they are supposed to be regulating*." (see PASQUALE, *c. d.*, p. 175).

[45] DASTIN, J. Amazon scraps secret AI recruiting tool that showed bias against women. In: *Reuters* [online]. 10. 10. 2018 [cit. 2024-02-26]. Available at: https://www.reuters.com/article/amazoncom-jobs-automation/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSL2N1VB1FQ/?-feedType=RSS%26feedName=companyNews.

[46] ALLEN, J. A. The color of algorithms. *Fordham Urban Law Journal*. 2019, Vol. 46, No. 2, pp. 219–270.

[47] HENMAN, P. Improving public services using artificial intelligence: possibilities, pitfalls, governance. *Asia Pacific Journal of Public Administration* [online]. 2020, Vol. 42, No. 4, p. 216 [cit. 2024-02-26]. Available at: https://doi.org/10.1080/23276665.2020.1816188.

including intellectual property, personal data, or risks of exposing weaknesses of an AADM model allowing some subjects to exploit them. However, at the same time complete "black box-ness" of AADM might lead to unfair biased decisions and with no accountability measures put in place might lead to a "Computer Says No" paradox as was described by Wihlborg et al.[48] This paradox is a reference to a sketch from a British TV comedy show called "Little Britain" where an "a sort of" administrator unwaveringly insists on the decision made by "a sort of" automated decision-making model because the "computer" said so, even though it was clearly wrong.[49]

The issue of transparency is not a straightforward yes-or-no matter; rather, it involves several intricate questions that regulators must address if they intend to integrate AADM into their public administration. However, if the regulator wishes to secure explicability or justification of AADM and to establish a model of AADM's accountability that is equitable and not overly burdensome on the administrators, I contend that, building upon the abovementioned considerations, it is imperative to commence with some degree of transparency.

4. CHALLENGES IN THE INTEGRATION OF AADM

The European Commission's high expert group on AI published an Ethics Guideline for Trustworthy AI claiming, that a trustworthy AI or its use, should be lawful, ethical, and robust.[50] As the question of technical and social robustness is beyond the aim of my analysis, this part of the article further explores solutions fulfilling the lawfulness and ethicality of a potentially applicable and trustworthy AADM. The ethicality of AI in a sense of the said guideline is composed by four principles specified as ethical imperatives i.e., (i) respect for human autonomy; (ii) prevention of harm; (iii) fairness; and (iv) explicability. The first two principles are related to the robustness and operational details of AI either calling for design respecting the human autonomy or for providing safeguards preventing any harm to subjects of AI and/or AADM. The latter two principles are reflected by imperatives regarding specifically machine learning models advocated by Lepri et al. who argue that such models should feature transparency, accountability, and fairness.[51]

According to my analysis transparency is a prerequisite for the explicability of AI and justification in case of AADM and the same holds true for ensuring fairness. Even though Pasquale calls for transparency in order to secure accountability of a black box,

[48] WIHLBLORG, E. et al. "The Computer Says No!": a Case Study on Automated Decision-Making in Public Authorities. In: *2016 49th Hawaii International Conference on System Sciences (HICSS)*. Piscataway, NJ: IEEE, 2016, pp. 2903–2912.

[49] In this episode a little girl goes to hospital to get her tonsils removed, however the hospital clerk insists, that she is going for a double hip replacement operation, because "The computer said no!" (see *Little Britain USA*. Episode 1 [Episode of a TV Show]. HBO. 28. 10. 2008).

[50] Directorate-General for Communications Networks, Content and Technology (European Commission). Ethics Guideline for Trustworthy AI. In: *European Commission: Shaping Europe's digital future* [online]. 8. 4. 2019 [cit. 2024-02-26]. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[51] LEPRI et al., *c. d.*

at the same time he underlines that full transparency would *"be a nightmare of privacy invasion, voyeurism, and intellectual property theft"*. Instead, he proposes something he calls *"qualified transparency"*, a model of transparency where a trusted and possibly regulated subjects have full access to transparency while at the same time being able to assess a black box's functioning. Also, as some authors argue transparency does not necessarily lead to an increase of intelligibility among average citizens, or that due to complexity full transparency can be overwhelming.[52]

In my interpretation qualified transparency is not only in need of a trusted authority but also of an educated one and thus overcoming even the *objective* and possibly even the *emerging opacity*. It is my belief, that such models guarantee safety, fairness and provide grounds for the accountability.[53, 54] I argue that this approach makes a path for full transparency but in case of AADM on a sort off "security clearance" and "expertise" based approach.[55]

As was addressed in the first part of this article, Olsen et al. suggest, that introduction of AI based decision-making models (i.e., AADM) should not be prevented by a regulation requiring full transparency and thus creating different legal standards. The proposed question is that a human decision-making regulation does not require a background check or the neurological screening of a particular administrator, so why should the regulation treat AADM any differently?[56] The nuances are where challenges emerge.

Motzfeldt explains, that any public administration[57] is a "fine-tuned system" with checks and balances developed over its history. The term "disruptive technology" is aptly applied to AI, signifying its potential to disturb the established order. The automation of public administration serves as just one example of this potential disruption. Therefore, with the introduction of AI there is a need to apply different standards at least.[58]

An administrative decision is still an individual act of public administration affecting the rights and obligations of a subject. Automated or not, the basic standards need to be met in order to secure legality of such decision. To this point I agree with Olsen et al. However, what Motzfeldt indicated, is that the technology indeed has a great potential but at the same time with automatization comes a lot of emerging features which may have serious impacts. Some of which were presented in this article.

Qualified transparency together with provided justification to each decision might fulfil the aspect of legality of decision made by AADM model. However, the question of fairness may start long before a decision is made. As Ettore suggests, ensuring fairness including equality of arms when contesting an automated decision is also a matter

[52] INNERARITY, *c. d.*

[53] Particular models of accountability will differ based on nations' approaches, however models to consider are with some division of accountability between the "qualified" authority assessing a fully transparent AADM model and the human decision-maker.

[54] PASQUALE, *c. d.*, p. 142.

[55] This model ensures that only trusted and educated authorities have a full access to analyse details of an AADM model thus ensuring lawfulness, ethicality, and robustness.

[56] OLSEN et al., *c. d.*, p. 6.

[57] For example, the public administration in Czech Republic has its roots from Austrian-Hungarian empire long before AI was invested, whereas the current model is not that different.

[58] Said rephrased statement is derived from a discussion with prof. Motzfeldt, and the ultimate phrasing was contingent upon her endorsement.

of models' design.[59] As previously noted, the technical specifications of an AADM model are not a subject of this article, nevertheless it is important to highlight that the abovementioned imperatives can be integrated into the AADM's model itself without the need for additional mechanisms to supply them. Provided that these models are designed accordingly.[60]

Naturally machine learning models, even if designed in fairness, transparency and accountability share an inherent disadvantage and that is, that they are by nature constantly evolving. Even a fair AADM model might come to point where it starts to produce biased or unfair decisions. For this purposes Olsen et al. advocates for two models that ensure supply of fresh inputs and keep human-in-the-loop while maximizing efficiency at the same time.

I call theirs first model a "80:20 split", whereas the logic behind it is that a respective administrative authority process decisions which are randomly split in two loads between an AADM model producing 80% of all drafts and a human administrator producing the rest 20%. All drafts are subsequently reviewed and signed off by a human administrator. The AADM model is continuously updated by all final decision with a human touch.[61]

The second model is called by the collective of authors is an "Administrative Turing Test". Inspired by Alan Turing's test determining whether a machine can think, this test is composed of a set-up in which a particular percentage of entire case load is given to a human administrator and to an AADM model. Both drafts are then reviewed by a "judge"[62] not knowing which draft was made by human and which by an AADM model. At the end, the most convincing draft is issued and used to update the AADM model.[63]

Both proposed models are not from my standpoint sufficient to ensure a fair and accountable AADM model, but at same time generate some thought-provoking suggestions about possibilities how to implement AADM and not to lose human touch.

My final proposal, considering the severity and the outcomes of a possibly wrong AADM, is that the AADM model should be a subject to a regulatory sandbox before its implementation into day-day administrative decision-making. This environment used for testing of innovative technologies, facilitating development under direct supervision of competent authorities[64] can together with all other mentioned proposals ensure

[59] ETTORRE, F. P. The Right to Contest Automated Decision. In: *The Digital Constitutionalist* [online]. 2022 [cit. 2024-02-20]. Available at: https://digi-con.org/the-right-to-contest-automated-decisions/.

[60] Some approaches include Fairness in Design proposed by Zhang et al. proposes ensuring ethical AI (See ZHANG, J. et al. Fairness in Design: a Framework for Facilitating Ethical Artificial Intelligence Designs. *International Journal of Crowd Sciences* [online]. 2023, Vol. 7, No. 1, pp. 32–39 [cit. 2024-02-20]. Available at: https://doi.org/10.26599/IJCS.2022.9100033; or Accountability in Design defined by VASSILAKOPOULOU, P. et al. Sociotechnical Approach for Accountability by Design in AI Systems. In: *Twenty-Eighth European Conference on Information Systems: Research-in-Progress Papers* [online]. 2020, pp. 1–8 [cit. 2024-02-20]. Available at: https://aisel.aisnet.org/ecis2020_rip/12/?utm_source=aisel.aisnet.org%2Fecis2020_rip%2F12&utm_medium=PDF&utm_campaign=PDFCoverPages.

[61] OLSEN et al., *c. d.*, p. 24.

[62] Human administrator who signs off the final decision.

[63] OLSEN et al., *c. d.*, p. 25.

[64] HANDRLICA, J. et al. Forum shopping in regulatory sandboxes and the perils of experimental law-making. *Juridical Tribune* [online]. 2023, Vol. 3, No. 3, pp. 408–426 [cit. 2024-02-26]. Available at: https://www.tribunajuridica.eu/arhiva/An13v3/5.%20Handrlica,%20Sharp,%20Nespor.pdf.

compliance of an AADM's essential requirements i.e., transparency and explicability, fairness, and potentially even accountability. This is also stipulated in the proposed regulation on AI as explained below.

As of the publication date of this article, the status of the European regulation on AI (i.e., the regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence) (AI Act) remains uncertain regarding its adoption by the European Parliament and what the specific wording of it will entail in its final form. However, according to the available information, the AI Act plans to adopt some of the above said proposals.

The general idea behind AI act is to divide different use cases of AI on a risk-based approach into 4 categories (i) unacceptable risk;[65] (ii) high risk; (iii) limited risk; and (iv) minimal risk., whereas the said proposals are associated mainly with the high-risk AI systems (the HRAIS).

The AI Act proposes mechanisms ensuring the mitigation of risk associated with HRAIS, such as the obligation to establish a risk management system, ensuring quality of testing data, accuracy, robustness, security, and of course transparency and human oversight. As the proposal of the AI Act stipulates, any HRAIS should be designed by transparency ensuring the ability of users to easily interpret systems' outputs thus justifying its decisions. Qualified transparency is further ensured, as proposed above, by the obligation to design and develop HRAIS ensuring effective human oversight in order to asses risk to health, safety, or fundamental rights.[66] Furthermore the AI Act imposes obligations on the Member States to develop a regulatory sandbox for the purpose of testing development, testing and validation of innovative AI models together with safeguards in place when processing personal data for such development.

In answering the question "When is an AI system considered to be high-risk?", Article 6 of AI Act stipulates that, all AI systems referred to in Annex III are to be considered as HRAIS. But a closer look to the proposed AI Act shows, that not all AADM models are considered as HRAIS.

In a proposed draft, the said annex outlines various use cases including the evaluation of eligibility of public assistance benefits and services[67] or the use for migration and asylum proceedings,[68] all falling under the umbrella of administrative decision-making. However, there is notable absence of a general provision recognizing AADM models as HRAIS as opposed to every judicial decision-making. The AI Act states that *"researching and interpreting facts and the law and in applying the law to a concrete set of facts"* by judicial authority is considered as HRAIS,[69] yet this designation doesn't extend to public authorities.[70]

[65] See Title II of the AI Act; This includes techniques unconsciously distorting humans' behaviour or exploiting vulnerabilities of specific groups, social scoring techniques or real-time remote biometric identification.

[66] See Chapter 2 of the ibid.

[67] See Annex III, para. 5(a) of the ibid.

[68] See Annex III, para. 7 of the ibid.

[69] See Annex III, para. 8 of the ibid.

[70] According to the Amendments to the AI Act adopted by the European Parliament on 14 June 2023, said provision was subject to an amendment extending its impact to administrative bodies. But as the AI Act is currently subject to negotiation in a difficult legislative process, the final wording remains unclear.

This approach inadvertently overlooks other administrative procedures that significantly impact fundamental human rights, such as administrative misdemeanour proceedings. Consequently, it's imperative to expand the aforementioned general provision concerning judicial decision-making to encompass public administration bodies as well.

It is my standpoint, that maintaining human involvement in the decision-making process, with humans making the final decision before the issuance, helps alleviate accountability concerns, especially in a situation where the requirements for transparency and justification are met. However, the more decision-making relies on AI, the more challenging it becomes to evaluate the accountability of individual administrators. The future could usher in a broader understanding of AI, potentially reaching a level of familiarity akin to today's use of smartphones or computers. This evolution would guarantee that the accountability of individual administrators aligns with their knowledge, without imposing excessively burdensome pressure on them. Until that point, overcoming the current challenges in the landscape remains a formidable obstacle for the full implementation of automated models in public administration.

5. CONCLUDING REMAKRS

This article explored the black box phenomenon in the context of automated decision-making in public administration and challenges stemming thereof. In terms of law, my research indicated that the essential challenge is associated with the necessity to provide an AADM model capable of providing justification as prescribed by the law.

As was described in the first part, the existence of opacity within an AADM does not necessarily need to be perceived as a drawback. This article argues that the lack of comprehension is not an anomaly since we experience black boxes in our daily lives. However, this fact is not a cause for consolation, as there are important principles at stake.

With that in mind, this paper concludes that full transparency with no provided safeguards is an unstable trajectory. Such an approach might result in exposing weaknesses of any AADM model and thus simplifying bypassing its safeguards. Instead, it advocates for a model of qualified transparency ensuring its comprehensibility without compromising legitimate interests in every AADM model. This type of transparency also meets with the legal imperatives of a decision-making requiring accountability of a public administration and a qualified form of explicability of decision that is its justification. Unfortunately, the European legislation on AI in development shows, that the said is ensured only to some degree.

The question of accountability remains a challenge as an AADM model might issue an unfair biased decision and thus discriminating one individual over another. I argue that even a meticulously designed AADM does not ensure that the model won't evolve in a discriminatory decision-maker. Therefore, I advocate for mechanisms ensuring fresh human inputs into such model and careful approach when implementing. As an example, might work models proposed by Olsen et al. At the same time, I argue that the emphasis on subjecting an AADM model to a regulatory sandbox has the potential to ensure its safety before integration.

It is without a doubt that the range and the complexity of AI in decision-making associates with different levels of inherent dilemmas. While certain problems are more straightforward to address, others pose significantly greater challenges. As the current model is a "fine-tuned system" with checks and balances in play the introduction of AI changes the game dramatically.

Addressing immediate challenges without considering their underlying causes provides only a temporary resolution. The suggested discourse on transitioning from the current human-centric public administration model to a fully automated system needs to be comprehensive, especially given the uncharted nature of AI.

Mgr. Jan Nešpor
Charles University, Faculty of Law
nesporjan@prf.cuni.cz
ORCID: 0009-0009-1500-5158