Journal of Ethics and Legal Technologies – Volume 4(1) – May 2022

Transparency in Artificial Intelligence: A Legal Perspective

Dr. Juris Vasiliki Papadouli

PhD in Private Law, Aristotle University of Thessaloniki,
LLm, Member of the Research Group "Digital Economy and Private Law" hosted by Faculty of Law, AUTh
Attorney at Law, Thessaloniki Bar Association
vickypapadouli@hotmail.com

Abstract: Autonomous Artificial Intelligence (henceforth AI) applications indicate extraordinary capabilities that completely alter our daily lives. Nonetheless, during and as a result of their operation, numerous incidents of human-rights violation have already been observed, thus, jeopardizing their public acceptance and further evolution. The main reason for this lies in the inherent opacity of autonomous AI systems, which constitutes the so-called black-box problem or black-box effect. To eliminate this effect, the scientific community often suggested that 'transparency' should be the appropriate tool to that end. Indeed, the volume of academic research conducted on the topic of AI transparency rapidly increased during the 21st century, urging even the European legislator to adopt harmonized rules of law regarding transparency in high-risk AI systems, among others. Nonetheless, neither is transparency's semantic context adequately defined, nor are its possible adverse effects exhaustingly explored. Consequently, concerns are raised regarding AI transparency's effectiveness. However, these concerns do not minimize the importance of transparency for the future of AI; they actually propose a different means of AI being perceived by the scientific community. Namely, taking into account that AI is per definition a multidisciplinary field, constituted both of computational and cognitive sciences, its transparency should accordingly have a dual meaning: first, a literal one, which would correspond to the technicalities of the decision-making procedure in an autonomous AI system; and second, a figurative one, which would refer to the necessity of fully comprehending the outcome of this procedure and, more importantly, of the human right to object to the decisions reached at by the autonomous AI system, ex ante or ex post. Subsequently, embedding transparency in AI should rather account for fostering a human-on-the-loop and a human-in-command approach than focusing only on a human-in-the-loop approach.

Keywords: autonomous AI systems, opacity, black-box effect, transparency

1. Introduction

Over the course of the 21st century, Artificial Intelligence has evolved into a scientific field of primary importance on a global scale since the constant contribution of its technological breakthroughs' to various fields of modern life. From finance, banking, monetary transactions, and the business enterprise to the medical sector, the justice system, and the entertainment industry, the effects of AI developments become more and more discernible, altering not only daily life, but also the citizens' overall attitudes and actions[1]. Despite the great attainments demonstrated by AI systems, concerns are raised regarding the possible negative effects they may provoke. More specifically, the incidents of human-rights violation that have been observed during, and as a result of, the operation of AI systems should be taken into account.

Indeed, the specific aspect of AI systems is deemed problematic, and therefore, a massive discussion is currently in effect regarding the means by which the issue of human-rights violation could be resolved without endangering the standards for the AI systems' outstanding performance. Surprisingly, the main underlying reason for the adverse effects induced by AI systems mirrors the main factor for their astonishing accomplishments, namely their inherent *opacity*, or, in other words, the humanly incomprehensible way AI systems operate.

Transparency is considered as the most appropriate tool for counteracting the inherent opacity of autonomous AI systems. Indeed, the rapidly increasing volume of research linking the application of AI technological developments to AI transparency reflects this matter. Based on this research, transparency is deemed a vital prerequisite for AI sustainability, namely for its further evolution and its acceptance by society[2]. As a consequence, this discussion triggers several concerns regarding the definition of transparency in AI and its significance for the future of the field.

The present paper aims at shedding light on the current debate on the meaning of transparency in AI, as well as presenting in more detail positive

[1] See more about AI applications in our lives *European Commission*, White Paper on AI – A European approach to excellence and trust, Com (2020) 65 final, 1; *European Commission*, Building trust in Human-Centric Artificial Intelligence, Com (2019) 168 final, 1; *Kemper/Kolkman*, "Transparent to whom? No algorithmic accountability without a critical audience" (2019) 22 Information, Communication and Society, 2082, who argue about an "*algorithmic life*".

[2] *European Commission*, Com (2019) 168 final, 2; *European Commission*, Com (2020) 65 final, 1; *Burt,* "The AI Transparency Paradox", Harvard Business Review (2019); *Wulf/Seizov,* "Artificial Intelligence And Transparency: A Blueprint For Improving The Regulation Of AI Applications In The EU" (2020) 31 European Business Law Review, 611.

and negative aspects AI transparency may entail. It further examines the form of transparency that should be implemented in autonomous AI systems and suggests possible (supporting) roles of the legal doctrine in this direction.

2. The significance of AI transparency

As previously mentioned, the discussion on AI transparency has recently received great interest, as it is directly linked to the discussion on AI sustainability. This connection is undoubtedly associated with the creation of *autonomous systems* or *autonomous machines* that are deemed the most significant AI breakthrough of the last decades and at the same time the most typical technological development[3] of the 4th Industrial Revolution era[4].

The most characteristic traits of the autonomous systems are the following, first, their *ability of learning,* and second, their *opacity*. The former feature, widely known as *machine learning*[5], refers to system's ability of improving itself at the execution of any assigned task, without having been explicitly programmed, thanks to its gained experience[6]. The latter feature, also known as the *black box problem* or *black box effect*, refers to the way an autonomous AI system operates incomprehensibly by human beings[7]. The black box problem[8] has three basic aspects. First, one uses the term in order to describe the complex and opaque way an autonomous AI system operates from a technical perspective. Second, the term is linked with the autonomous system's difficulty in providing a suitable explanation about the means and justification of a specific decision it has reached, using a language comprehensible for human beings[9]; third, it refers to the overall incapacity to

[3] See about the "ages" of Artificial Intelligence as scientific field *Smith* in The history of Artificial Intelligence, (2006), 6; *Bostrom*, Superintelligence, (2014), § 1.

[4] See more about the notion of the 4th Industrial Revolution *Braütigam/Klindt,"* Industrie 4.0, das Internet der Dinge und das Recht" NHW 2015, 1137.

[5] *Denicola, "*Ex machina: Copyright Protection For Computer-Generated Works" (2016) 69 Rutgers Univ. Law Review, 254-255; *Yanski-Ravid*, "Generating Rebrandt, Artificial Intelligence, Copyright, And Accountability In The 3A Era-The Human-like Authors Are Already Here- A New Model" (2017) Mich.St.L.Review, 676.

[6] *Russell/Norvig,* Artificial Intelligence, (2020), 39; *Wettig/Zehendner,* "A legal Analysis Of Human And Electronic Agents" (2004) 12 Artificial Intelligence and law, 111-135.

[7] *Wulf/Seizov* (2020), 619.

[8] Originally known as *qualification problem,* see *McCarthy/Hayes*, "Some Philosophical Problems From The Standpoint Of Artificial Intelligence" 1969.

[9] *Burrel, "*How The Machine "thinks": Understanding Opacity In Machine Learning Algorithms" 2016 Big Data & Society, 1; *Adadi/Berrada* "Peeking Inside The Black-Box" (2018) 6 IEEE Access, 52141; *Zednik*, "Solving The Black Box Problem", 2019; *European Commission*, Com (2020) 65 final, 12; *European Commission*, Annexes, SWD (2021) 84 final, 34.

a posteriori explain the AI system's reached outcomes, even by the system's designers and programmers.

However, if the autonomous AI system's reached decisions cannot be justified by the system itself nor inspected by humans, people's confidence in AI will decrease and AI sustainability will be endangered[10]. In order for such a situation to be avoided, the scientific community recommends that AI transparency be the necessary tool for counteracting the black box effect[11].

3. *Transparency* defined from the perspective of Artificial Intelligence

Defining transparency is an endeavor. The main reason for that lies in the fact that it is a multidisciplinary concept[12]. One can encounter this term in various scientific fields, such as in physics, social sciences, policy-making, etc. Nonetheless, its meaning differs among those sectors. For instance, in physics transparency means the physical capacity of a material to allow light to pass through it, permitting someone to see its interior. In social sciences, transparency is associated with a deeper knowledge and understanding of the means leading to a specific decision being drawn or an outcome being produced; it can be interpreted, in general, as displaying the infrastructure of a procedure, a fact that renders the person executing this procedure accountable for their actions/decisions and, thus, (morally) responsible for them[13]. In this framework, transparency in public government and policy making is usually connected with the attempt to decrease arbitrariness, fraud, and bribery in the political sector by divulging the procedure for reaching a specific decision[14]. In figurative terms, when a procedure is not transparent, it is *opaque*, leaving people in the dark regarding significant political or other issues and maintaining, in this way, the existing informational asymmetries[15]. This phenomenon causes several important inconveniences for the political

[10] Cf. *Licht/Licht*, "Artificial Intelligence, Transparency, And Public Decision-making" (2020) 35 AI & Society, 918.

[11] *European Commission*, Com (2020) 65 final, 15.

[12] *Larsson/Heintz*, "Transparency In Artificial Intelligence" (2020) 9 Issue 2 Internet Policy Review; *Felzmann et al.*, "Towards Transparency in By Design For Artificial Intelligence" (2020) 26 Science an Engineering Ethics, 3333.

[13] Cf. *Ananny/Crawford*, "Seeing Without Knowing: Limitations Of The Transparency Ideal And Its Application to Algorithmic Accountability" 2016 (13) New Media & Society, 3-4; *Licht/Licht* (2020), 918.

[14] See also *Doshi-Velez et al.*, "Accountability Of AI Under The Law: The Role of Explanation" working draft 2017, 5-7.

[15] Cf. *Lepri et al.*, "Fair, Transparent and Accountable Algorithmic Decision-making Process" (2018) 31 Philosophy and Technology, 611-627.

system, with the reduction of public confidence in political institutions being one of the most severe.

Regarding AI, in the current scholarship, transparency is translated as *explainable*, *interpretable*, *responsible* or *understandable* AI, while terms such as *traceability*[16], *understandability, inspectability*, *verifiability, explicability, interpretability*[17], *auditability*[18] or *accountability*[19] are often directly connected with it[20]. These terms, albeit not synonymous, are very often considered as equivalent to one another. In general, their meaning could be summarized as the AI systems' ability to reach decisions in a way interpretable and understandable by humans, who should have the right and the capacity to inspect and appeal them.

Given the fact that AI is a multidisciplinary scientific field, as it consists both of computational and cognitive sciences[21], the concept of AI transparency should be shaped accordingly[22]. Namely, the term transparency in AI should have a dual meaning, comprising both of a *physical* and a *cognitive* element. The physical element refers to the system's capacity to show its inner working processes and the origin of the training data *(literal transparency)*[23]. This physical element can be further analyzed in more parts, such as *functional transparency*, with respect to how the AI system functions as a whole (also referred to as *simulatability*) or the means by which its individual components operate (*decomposability*); *structural transparency* or *algorithmic transparency*, with respect to how the algorithm was realized in code and functions; and last, *run transparency*, regarding the way the program actually runs in a particular case[24]. The cognitive element refers on the one hand to the comprehensible explanations the AI system must be able to provide regarding its reached decisions *(figurative transparency)*[25],

[16] See more about this concept *European Commission*, COM (2019) 168 final, 5.
[17] See more about this concept *European Commission*, COM (2019) 168 final, 5.
[18] See more about this concept *European Commission,* COM (2019) 168 final, 6.
[19] See more about this concept *European Commission*, COM (2019), 168 final, 6.
[20] See *Floridi et al.,* "AI4People: An Ethical Framework For A Good AI Society: Opportunities, Risks, Principles, and Recommendations" (2018) 28 Minds and Machines, 689-700; *Kemper/Kolkman* (2019), 2083; *Schmidt/Biesmann/Teubner*, "Transparency And Trust In Artificial Intelligence Systems" 2020 Journal of Decision Systems, 2; *Felzmann et al.* (2020), 3337-3338.
[21] *Adadi/Berrada* (2018), 52145. See also *European Economic and Social Committee*, Opinion 2020/C 364/12, Recital Nr. 2.9.
[22] Cf. *Larsson/Heintz* (2020), 5.
[23] Cf. *Felzmann et al.* (2020), 2, who adopt the term "*prospective transparency*".
[24] *European Commission,* Annexes, SWD (2021) 84 final, 34.
[25] Cf. *Lepri et al.* (2018), 12; *Felzmann et al.* (2020), 3351, who adopt the term "*retrospective transparency*". See also for another perspective *Kizilcec*, "How much transparency? Effects of Transparency on Trust in an Algorithmic Interface", (2016) Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.

which may vary among the AI autonomous systems[26], and on the other hand, to the human right to participate in the decision-making process even *ex-post*. Literal transparency allows for human control over the AI system, whilst figurative transparency facilitates the system's accountability and, subsequently, the user's liability. Both types result in enhancing the public's confidence[27] in the AI inner procedures and in the aptness of its reached outcomes.

In light of all these it could be argued that literal transparency is an *ex-ante* method for monitoring the AI algorithms' inner workings and the interconnections behind the decision-making process before a final decision has been reached, while figurative transparency is an *ex-post* method for checking the systems' outcomes[28]. Ex-post transparency has occasionally been described in the literature as *transparency in rational,* in contrast to *transparency in process*, which corresponds to literal transparency[29]. However, also exists a third concept of transparency, namely the *transparency in policy*, which refers to the values AI should follow and/or to the goals it should try to attain[30]. Transparency in policy is another form of *figurative transparency* and at the same time an *ex-ante* method for controlling the system's reached decisions by tracing the basic values to which the AI system should adhere.

The before-mentioned three types of transparency (*in policy*, *in process*, *in rational*) are closely interconnected. One can envisage them as three different and consecutive steps[31]; the first step corresponds to transparency in policy, where the desirable values of AI should be selected and the desirable goals should be defined; the second step corresponds to transparency in process, where the AI system displays its inner decision-making process and the origin of its training data sets; finally, the third step corresponds to transparency in rational, where the system is able to provide specific justifications for its reached decisions[32] in a non-technical language and humans have the right to appeal them *ex post*[33]. This interconnection confirms that AI is an interdisciplinary scientific field which demands a close collabora-

[26] *Hacker et al.*, "Explainable AI under Contract and Tort Law: Legal Incentives And Technical Challenges" (2020) 28 Artificial Ingelligence and Law, 431.
[27] See also *Lipton* "The Mythos of Model Interpretability" 2017 acmqeue, 7.
[28] Cf. *Lepri et al.* (2018), 12; *Felzmann et al.* (2020), 3.
[29] Cf. *Licht/Licht* (2020), 918.
[30] Widely described as "*AI's ethical box"*, cf. *Floridi et al.* (2018).
[31] Cf. *Licht/Licht* (2020), 918, according to whom they should be understood as "degrees" of transparency.
[32] Cf. *Licht/Licht* (2020), 918.
[33] Cf. *Floridi et al.* (2018), 697-698.

tion among different scientific fields, in order for the best outcomes to be reached[34].

4. Transparency as a tool for counteracting the black box effect

In order for the black box effect to be eliminated, a specific subfield of AI called *explainable AI* (also known as *XAI*)[35] has emerged. This subfield aims at creating the appropriate techniques that will allow autonomous AI systems to be explicable, whilst maintaining high levels of autonomy[36]. To that aim, XAI fosters the request of embedding transparency in AI systems[37]. However, the question remains on which of the above-mentioned types of transparency should be implemented in AI, namely transparency in policy, transparency in process or transparency in rational?

The prevailing opinion in the literature seems not to take into account the already mentioned types of transparency. Rather, it focuses on transparency in terms of the inner processes of the autonomous system, as well as on the system's ability to justify its reached decisions. According to this approach, human beings retain the right of awareness regarding the way the autonomous system operates, while the system is accountable of 'accounting for' its decisions. Therefore, the most prevalent approach supports the necessity of embedding transparency in process and transparency in rational in AI.

However, this approach raises several concerns. First of all, from a practical standpoint, it is claimed that constructing more transparent autonomous systems is not an easy endeavor, since the black box effect is an inherent drawback of autonomous systems. Any attempt to make the systems' inner procedures more transparent can have adverse effects on their accuracy and efficiency. As mentioned in the literature[38], the more transparent the auton-

[34] Cf. *Lepri et al.*, (2018), 12; *Hacker et al.* (2020), 435-436; *Wulf/Seizov* (2020), 622.

[35] The term was first used by *Van Lent/Fisher/Mancuso* in the framework of a game simulation, "An Explainable Artificial Intelligence System For Small-unit Tactical Behavior" 2004 IAAI Emerging Application, 900. See also *Adadi/Berrada* (2018), 52139; *Zednik* (2019), 2; *Licht/Licht* (2020), 919; *Carabantes* "Black-Box Artificial Intelligence: An Epistemological And Critical Analysis" (2020) 25 AI & Society, 314; *Wulf/Seizov* (2020), 621-622.

[36] *Adadi/Berrada* (2018), 52138. However, the authors admit to the absence of a generally accepted definition of *XAI* (52140).

[37] *Adadi/Berrada* (2018), 52142. See also *European Parliament*, Civil Law Rules on Robotics, (2018/C 252/25), Ethical Principles, Recital Nr. 12.

[38] *Weller,* "Challenges for Transparency", 2017 Open Review, 1; *Adadi/Berrada* (2018), 52145; *Lepri et al.* (2018), 9-10; *Felzmann et al.* "Trasparency You Can Trust: Transparency Requirements For Artificial Intelligence Between Legal Norms and Contextual Concerns", 2019 Big Data & Society, 7; *Felzmann et. al.* (2020), 3339-3340; *Carabantes* (2020), 310; *Hacker et al.* (2020), 430; *Wulf/Seizov* (2020), 619.

omous system is, the less accurate it is. Moreover, it is claimed that more transparent algorithms could be less resistant and hence, more susceptible to 'attacks'[39], while their designing procedure is much steeper[40] and may have a negative environmental footprint[41]. Furthermore, it is argued that the disclosure of the system's inner proceedings and of the source of its training data may lead to companies losing a competitive advantage[42], to invasion of business secrets,[43] and to an infringement on sensitive data and/or specific rules of law established for their protection[44]. After all, there is always the danger of explanations being intentionally manipulated by the provider – typically by the large companies using AI[45]. The possibility of some sort of system's failure cannot be avoided through transparency, as well. For instance, existing biases in the training data are not hard-coded and so transparency in process cannot help to identify them[46]. Last but not least, it is widely supported that transparency in process does not actually serve the desirable goal of enhancing people's confidence in AI. The main reason for this lies in the fact that the provided information about the system's inner proceedings is too complicated to be edited by an audience, which is technologically illiterate[47]. In other words, even if a company reveals its programming codes or the exact algorithm it uses, it is very difficult for the AI users to understand how the AI system works since they lack specific knowledge about it. This is the so-called problem of *information overload* or *transparency paradox*[48] that actually leads to a counter effect: misleading rather than illuminating people about the system's reliability[49].

[39] *Burrel* (2016), 3; *Kemper/Kolkman* (2019), 2086; *Burt* (2019).

[40] *Adadi/Berrada* (2018), 52143; *Felzmann et al.* (2019), 7; *Burt* (2019). See an in-depth analysis about the cost of AI *European Commission*, Annexes, SWD (2021) 84 final, 3-6.

[41] See *European Parliament*, Civil Law Rules on Robotics, (2018/C 252/25), Environmental Impact, Recital 47. Cf. also *Woodstra*, "What Does Transparent AI Mean?" 2020 AI Police Exchange, 2.

[42] *Burrel* (2016), 3.

[43] *Kemper/Kolkman* (2019), 2086.

[44] *Wellller* (2017), 57-58; *De Laat,* "Algorithmic Decision-Making Based On Machine Learning From Big Data: Can Transparency Restore Accountability" (2018) 31 Philosophy and Technology, 527-528; *Lepri* et al. (2018), 9; *Burt* (2019); *Felzmann* et al. (2020), 3340; *Woodstra* (2020), 2.

[45] *Burt* (2019), 2; *Wulf/Seizov* (2020), 615.

[46] *Kemper/Kolkman* (2019), 2086.

[47] Cf. *Kemper/Kolkman* (2019), 2086.

[48] *Larsson/Heintz* (2020), 6-7; *Licht/Licht* (2020), 922. See also *Doshi-Velez et al.* (2017), 4; *Weller* (2017), 57, 58; *Felzmann et al.* (2019), 3, 7, 8; *Burt* (2019). Cf. *Richards/King*, "Three Paradoxs Of Big Data" (2013) 66 Stanford Law Review, 41.

[49] The phenomenon is also encountered in other fields, leading to similar results. For instance, in consumer protection, when too much information about a product or a service is given to a consumer, so that they cannot actually cope with it, their final decision about purchasing or not the specific product or service may be false. See for more *Wulf/Seizov*

At this point, the provisions of the European Commission at the Proposal for a *Regulation of the European Parliament and of the Council regarding harmonized rules on AI*[50], including AI transparency, are worth noting. Although much emphasis is placed on *transparency-standard* of high-risk AI systems[51] as a prerequisite for their acceptance, the provided information to the AI user seems in essence to be rather technical and ordinary than illuminating[52]. Indeed, the information provided in Art. 13 of COM (2021) 206 final, concerns data, such as the identity and the contact details of the provider, the characteristics, capabilities and limitations of performance of the high-risk AI systems, including its intended purpose, the expected "lifetime" of the AI system and any care measures, etc., but not their produced outcomes and how the AI system has reached them. For these reasons, it is claimed that transparency in process may not, in fact, boost public confidence in the autonomous AI systems and, hence, it may not contribute to AI sustainability[53]. Although technical information is highly important, especially for AI programmers, designers, and engineers, its disclosure may not be so crucial for the average person to rely on AI outcomes and to accept them[54].

Consequently, the scientific community should put more emphasis on figurative transparency (transparency in policy and transparency in rational) than on literate transparency[55]. More specifically, the main focus should be more on the values[56] the AI systems (must) follow and on the systems' provided explanations about their reached decisions in an accessible language, rather than on the technicalities about how and why the AI systems reached them. In other words, our attempts should be directed towards a more *human-on-the-loop (HOTL)* and *human-in-command (HIC)* approach than to a *human-in-the-loop (HITL)* approach[57]. The human-on-the-loop approach refers to the capability of human intervention during the *design cycle* of the system and the monitoring of the system's operation, while the hu-

(2020), 628.

[50] COM (2021) 206 final.

[51] See COM (2021) 206 final, Art. 13 "*High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately.* (…)".

[52] See the provision at Art. 13 par. 3 (COM (2021) 206 final).

[53] *Licht/Licht* (2020), 920-923.

[54] *Doshi-Velez et al.* (2017), 4. See also *European Commission*, COM (2019), 168 final, 4 (Reference 13). Cf. also *Zednik* (2019), 12, who highlights the difference between *what*-questions and *why*-questions for several stakeholders in the Machine Learning ecosystem.

[55] Cf. *Licht/Licht* (2020), 923-924.

[56] See *European Commission*, COM (2019) 168 final, 2, pointing out that "*The values on which our societies are based need to be fully integrated in the way AI develops.*".

[57] See *European Commission*, COM (2019), 168 final, 4 (Reference 13)· Opinion of the *European Economic and Social Committee*, COM (2020) 65 final 364/12), Recital Nr. 2.3.

man-in-command approach refers to the human capability to oversee the overall activity of the AI system and to their ability to decide when and how to use the system in any particular situation including the ability to override a decision made by the system. Finally, the human-in-the-loop approach refers to the human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable.

5. Embedding transparency in autonomous AI systems

A human-on-the-loop approach suggests the need to embed *transparency in policy* in autonomous AI systems; in other words, the necessity of constructing and embedding an *ethical box* in autonomous AI systems. Although this term has already been used in literature, its meaning seems to escape clarity in so far as there is no specific definition for it. According to one opinion, systems with an ethical box are *moral AI systems,* namely AI systems which can *display an acceptable behavior*, not causing any harm either to the people or to themselves[58]. A more progressive approach suggests that *moral systems* are the systems that can select the best action in case of ethical dilemmas[59] (e.g. a self-driving car in the case of an unavoidable collision). Despite the fact that the latter approach entails a truly revolutionary concept, if one examines it more thoroughly, they can realize that such an attempt will fail to succeed. The reason for that lies in the fact that there is no worldwide acceptable global ethical theory; there never was and maybe there will never be. Therefore, any attempt to create moral AI systems seems to be as much condemned as the attempt to form moral people. Although research has already been conducted on this subfield of AI, called *AI ethics*, and has already exhibited quite interesting ideas (e.g. the proposal of creating AI systems that will follow Kant's *categorical imperative* or Aristotle's theory on *ethics* or the principles of *utilitarianism*)[60], it cannot be held that we can ever construct AI systems capable of rendering ethical decisions, particularly in cases of ethical dilemmas; even humans are not able to execute such a task[61]. Basically, this is the reason why AI systems (will) fail to initiate such actions.

[58] See *Assaro*, "What Should We Want From A Robot Ethic?" (12/2006) 6 International Review of Information Ethics, 10.
[59] See *Assaro* (2006), 10.
[60] See *Allen et al.*, "Prolegomena To Any Further Artificial Moral Agent" (2000) 12 Journal of Experimental & Theoretical Artificial Intelligence, 251; *Anderson/Anderson*, "Machine Ethics: Creating An Ethical Intelligent Agent" (2007) 4 AI Magazine, 15.
[61] *Carabantes* (2020), 316.

For this reason, creating moral AI systems should not mean creating either autonomous systems capable of rendering ethical decisions in cases of ethical dilemmas or AI systems that can demonstrate an acceptable behavior. Instead, it should mean *designing autonomous AI systems that are able to follow specific rules encoded in their interior code*[62]. It is the latter which should be ethical, namely in accordance with the current legal social values, which - in turn - are reflected in the current legal framework. In order for this to happen, abstract ethical values and rules of law, such as justice, respect of human dignity, lack of discrimination, protection of privacy, democracy, etc.,[63] should be translated in (more) technical terms in order to be encoded in AI. A specific subfield of AI, called *Microethics*[64], has already been elaborating on this project. However, the abstract content of these concepts cannot easily be mathematically operationalized, a fact that inhibits their implementation in the AI system's *ethical box*[65]. Although steps have already been taken in this direction, it seems that the scientific community is still far away from achieving this goal.

A human-in-command approach fosters the embedding of transparency in rational in AI systems, asking both for interpretable explanations in an accessible language behind each decision the system has reached and for a human right to participate in the decision making process, even *ex-post*. For instance, in the case of self-driving cars, the human driver must have the capacity to override the system in deciding the course to be taken, when they realize the system's decision is wrong; similarly, in the case of autonomous AI systems in the job market that conduct recruitment/dismissal/promotion procedures, the employer must have the capacity to recall a system's decision; accordingly, in the case of electronic personal assistants, the AI system can provide the user with the right to consent to or deny concluding a specific contract, or the right to alter the face of their contractual partner and/or to modify the contractual content. This could easily apply, if, for example, an automated message (SMS or email) is sent to the user by the autonomous AI system, asking for their consent in order for the contract to be concluded.

[62] Cf. *Anderson/Anderson* (2007), 15, who distinguish between *implicit* and *explicit* intelligent agents; *implicit* are the intelligent agents that are able to follow specific ethical rules, which are already encoded, while *explicit* are the agents that are capable of rendering right ethical decisions in case of ethical dilemmas. According to the authors, priority must be given to the creation of *explicit* intelligent agents. See also *Assaro* (2006), 11, who characterizes an agent moral, when it is able to "*adhere to systems of ethics*".

[63] See *Floridi et al.* (2018); *European Commission,* Com (2019), 168 final, 2; *European Commission,* COM (2020), 65 final, 3.

[64] See *Hagendorff,* "The Ethics Of AI Ethics: An Evaluation Of Guidelines" (2020) 30 Minds and Machines, 111.

[65] Cf. *Allen et al.* (2000), 257; *Anderson/Anderson* (2007), 18; *Hagendorff* (2020), 111.

This last aspect seems to be of primary importance mainly due to the inherent black box effect that not only hampers the system's inner visibility, but also affects its ability to explain its produced outcomes[66]. Therefore, instead of asking every time for rational explanations behind each system's decision - after all a human is also unable of showing such a rational behavior[67] - or trying to decode the system's provided justifications with various mechanisms (such as natural language explanations, visualizations of learned representations or modes and explanations by example[68]), it is more important that the scientific community reinforce human interference in the decision-making process *a posteriori* by ensuring the human right to appeal the system's decisions, prior to (override) or after their being made (recall)[69]. In this way we ensure human *autonomy*. Although this concept is traditionally interpreted as human freedom to make decisions for one's self, in the new reality, where humans and autonomous AI systems collaborate in various fields to attain a specific goal, its semantic context should be broadened, covering the human's right to decide whether or not to *adopt* an autonomous system's reached decision, as well. In this *meta-autonomy era*[70] we already live in, human accountability is going to be based not only on the human will to take decisions for themselves, but also on their will to accept or to decline decisions made by AI systems for them.

6. Conclusions

Having said all this, this contribution supports the concept that transparency in AI is a multidisciplinary subfield of AI[71], which requires the collaboration of scholars working on AI more than ever before. We should perceive AI transparency in a dual sense: literal and figurative. As legal scholars, our emphasis shall be placed on figurative transparency, fostering the integration of transparency in policy and transparency in rational in autonomous AI systems. At this point, our contribution could be highly constructive in multiple ways[72]. First of all, we can contribute to the selection and definition of the rules and values AI systems should follow, in order to comply with the current legal framework and the current social reality[73]. In other words, we

[66] *Felzmann et al.* (2019), 4.
[67] *Zerilli et al.*, Transparency in Algorithimic And Human Decision-Making: Is There A Double Standard? (2018) 32 Philosophy and Technology.
[68] *Lipton* (2018), 15 ff.
[69] See *Woodstra* (2020), 3.
[70] See *Floridi et al.* (2018), 698.
[71] Cf. *Felzmann et al.* (2019), 3.
[72] Cf. *Burt* (2016), 3-4.
[73] Cf. *Larsson/Heintz* (2020), 7.

can facilitate the construction of an AI *legal and ethical box*. Furthermore, we can recommend ways for examining the rightness of the AI systems' justifications and their conformity to their embedded legal and ethical box[74]. In addition to this, we can considerably contribute to the development of ex-post mechanisms that will provide humans with the right to object to the AI system's decisions and recall or override them[75], enhancing in this way the *traceability of the emerging human responsibility* in case of damage[76].

References

Adadi Amina and Berrada Mohammed, "Peeking Inside The Black-Box" (2018) 6 (2018) 6 *IEEE Access* 52138-52160.

Allen Collin and Varner Gary and Zinser Jason, "Prolegomena To Any Further Artificial Moral Agent" (2000) 12 *Journal of Experimental & theoretical Artificial Intelligence* 251-261.

Ananny Mike and Crawford Kate, "Seeing Without Knowing: Limitations Of The Transparency Ideal And Its Application to Algorithmic Accountability" 2016 (13) *New Media & Society* 1-17.

Anderson Michael and Anderson Susan Leigh, "Machine Ethics: Creating An Ethical Intelligent Agent" (2007) 4 *AI Magazine* 15-26.

Assaro Peter, "What Should We Want From A Robot Ethic?" (12/2006) 6 *International Review of Information Ethics* 9-16.

Bostrom Nick, *Superintelligence* (Oxford: OUP, 2014).

Braütigam Peter and Klindt Thomas, „Industrie 4.0, Das Internet Der Dinge Und Das Recht" 2015 *NJW* 1137-1142.

Burrel Jenna, "How The Machine "thinks": Understanding Opacity In Machine Learning Algorithms" 2016 *Big Data & Society* 1-12.

Burt Andrew, "The AI Transparency Paradox" (2019), available at https://hbr.org/2019/12/the-ai-transparency-paradoxHarvard Business Review (accessed 27 December 2021).

Carabantes Manuel, "Black-Box Artificial Intelligence: An Epistemological And Critical Analysis" 2020 (35) *AI & Society*, 309-317.

[74] See *European Parliament*, Civil Law Rules on Robotics, (2018/C 252/25), Recital Q.

[75] See *European Parliament* Resolution on Civil Law Rules on Robotics (2018/C 252/25), Recital Q· *ICDPPC*, Declaration on ethics and data protection in artificial intelligence, (2018), 5.

[76] See *European Commission*, Com (2020) 64 final, 9).

De Laat Paul B., "Algorithmic Decision-Making Based On Machine Learning From Big Data: Can Transparency Restore Accountability" (2018) 31 *Philosophy and Technology* 525-541.

Denicola Robert, "Ex machina: Copyright Protection For Computer-Generated Works" (2016) 69 *Rutgers Univ. Law Review* 251-287.

Doshi-Velez Finale, Kortz Mason, Budish Ryan, Bavitz Chris, Gershman Sam, O'Brien David, Scott Kate, Shieber Stuart, Waldo James, Weinberger David, Weller Adrian and Wood Alexandra, "Accountability Of AI Under The Law: The Role of Explanation" working draft (2017), available online https://arxiv.org/ftp/arxiv/papers/1711/1711.01134.pdf (accessed 28 October 2021).

European Commission, Annexes, SWD (2021) 84 final.

European Commission, Building trust in Human-Centric Artificial Intelligence, Com (2019) 168 final.

European Commission, Digitalization of justice in the European Union, A toolbox of opportunities, COM (2021) 710 final.

European Commission, White Paper on AI – A European approach to excellence and trust, Com (2020) 65 final.

European Economic and Social Committee, Opinion of the European Economic and Social Committee on "White paper on Artificial Intelligence — A European approach to excellence and trust", 2020/C 364/12.

European Parliament, Civil Law Rules on Robotics, 2018/C 252/25.

Felzmann Heike, Villaronga Eduard Fosch, Lutz Christoph and Tamò-Larrieux Aurelia, "Transparency You Can Trust: Transparency Requirements For Artificial Intelligence Between Legal Norms And Contextual Concerns" 2019 *Big Data & Society* 1-19.

Felzmann Heike, Villaronga Eduard- Fosch, Lutz Christoph and TamòLarrieux Aurelia, "Towards Transparency By Design For Artificial Intelligence" (2020) 26 *Science and Engineering Ethics*, 3333-3361.

Floridi Luciano, Cowls Josh, Beltrametti Monica, Chatila Raja, Chazerand Patrice, Dignum Virginia, Luetge Christoph, Madelin Robert, Pagallo Ugo, Rossi Francesca, Schafer Burkhard, Valcke Peggy and Vayena Efy, "AI4People: An Ethical Framework For A Good AI Society: Opportunities, Risks, Principles, and Recommendations" (2018) 28 *Minds and Machines* 689-707.

Hacker Philipp, Krestel Ralf, Grundmann Stefan and Naumann Felix, "Explainable AI under Contract and Tort Law: Legal Incentives And Technical Challenges"2020 (28) *Artificial Intelligence and Law* 415-439.

Hagendorff Thilo, "The Ethics Ff AI Ethics: An Evaluation Of Guidelines" (2020) 30 *Minds and Machines* 99-120.

International Conference of Data Protection and Privacy Commissioners, "Declaration On Ethics And Data Protection In Artificial Intelligence (2018), available online https://globalprivacyassembly.org/author/icdppc-update/ (accessed 27 April 2021).

Kemper Jakko and Kolkman Daan, "Transparent to whom? No algorithmic accountability without a critical audience" (2019) 22 *Information, Communication and Society* 2081-2096.

Kizilcec Rene, "How Much Transparency? Effects Of Transparency On Trust In An Algorithmic Interface" (2016), available online https://rene.kizilcec.com/wp-content/uploads/2016/01/kizilcec2016information.pdf (accessed 13 May 2021).

Larsson Stefan and Heintz Fredrik, "Transparency In Artificial Intelligence" (2020) 9 Issue 2 *Internet Policy Review* 1-16.

Lepri Bruno, Oliver Nuria, Letouz´e Emmanuel, Pentland Alex and Vinck Patrick, "Fair, Transparent And Accountable Algorithmic Decision-making Processes" (2018) 31 *Philosophy and Technology* 611-627.

Licht Karl de Fine and Licht Jenny de Fine, "Artificial Intelligence, Transparency, And Public Decision-making" (2020) 35 *AI & Society* 915-926.

Lipton Zachary, "The Mythos Of Model Interpretability" 2017 *acmqeue*, 1-27.

McCarthy John and Hayes Patric, "Some Philosophical Problems From The Standpoint Of Artificial Intelligence" (1969), available online http://jmc.stanford.edu/articles/mcchay69/mcchay69.pdf (accessed 28 October 2021).

Richards Neil and King Jonathan, "Three Paradoxs Of Big Data" (2013) 66 *Stanford Law Review* 41-46.

Russell Stuart and Norvig Peter, "Artificial Intelligence: A modern approach" 4th Edition. (2020) Pearson Series for AI, Hoboken.

Schmidt Philipp, Biesmann Felix and Teubner Timm, "Transparency And Trust In Artificial Intelligence Systems" 2020 *Journal of Decision Systems,* also available online https://www.tandfonline.com/doi/full/10.1080/12460125.2020.1819094 (accessed 28 October 2021).

Smith Chris in "The History Of Artificial Intelligence" (2006), available online https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf (accessed 28 October 2021).

Uni global Union, "Top 10 Principles For Ethical Artificial Intelligence" available online http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf (accessed 11 December 2021).

Van Lent Michael, Fisher William and Mancuso Michael, "An Explainable Artificial Intelligence System For Small-unit Tactical Behavior" 2004 *IAAI Emerging Application* 900-907.

Weller Adrian, "Challenges For Transparency", (2017) 58 *Open Review* 55-62.

Wettig Steffen and Zehendner Eberhard, "A legal Analysis Of Human And Electronic Agents" (2004) 12 *Artificial Intelligence and law* 111-135.

Woodstra Fenna, "What Does Transparent AI Mean?" 2020 *AI Police Exchange,* also available online https://aipolicyexchange.org/2020/05/09/what-does-transparent-ai-mean/ (accessed 27 April 2021).

Wulf Alexander and Seizov Ognyan, "Artificial Intelligence And Transparency: A Blueprint For Improving The Regulation Of AI Applications In The EU" (2020) 31 *European Business Law Review* 611-640.

Yanski-Ravid Shlomit, "Generating Rebrandt, Artificial Intelligence, Copyright, And Accountability In The 3A Era-The Human-like Authors Are Already Here- A New Model" (2017) Mich.St.L.Review, 659-726.

Zednik Carlos, "Solving The Black Box Problem" (2019), available online https://arxiv.org/ftp/arxiv/papers/1903/1903.04361.pdf (accessed 26.04.2021).

Zerilli John, Knott Alistair, Maclaurin James and Gavaghan Colin, "Transparency In Algorithmic And Human Decision-Making: Is There A Double Standard?" (2018) 32 Philosophy and Technology 1-24