Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/CLSR

Computer Law & Security Review

Promises and limits of law for a human-centric artificial intelligence

*David Restrepo Amariles, Pablo Marcello Baquero**

HEC Paris, France

ARTICLE INFO

Keywords:
Human-centric artificial intelligence
Regulation of artificial intelligence
AI systems
Lifecycle and co-evolutionary approach
Human rights-based perspective
Proportionality test

ABSTRACT

While the concept of *human-centric artificial intelligence* (AI) has emerged as a key principle to govern AI systems, two obstacles for its implementation remain largely understated. First, the excessive focus on accountability at the *design* stage of AI systems, overshadowing the fact that human values can be affected at different stages across the AI life cycle. Second, the market-driven approach of current regulatory initiatives, limited in their ability to actively promote human values. In this article, we argue for a twofold approach to tackle these limitations. On one hand, we propose a co-evolutionary and life cycle approach to tackle the lack of accountability of AI systems, showing that this approach can help ensure accountability beyond the design stage by enabling meaningful human control and human-AI interaction across the entire lifecycle of the system. On the other hand, we propose that regulatory initiatives should balance the market-driven approach by giving a more predominant role to human rights and by introducing explicitly the notion of proportionality test. This *rebalancing* would serve to handle conflicts between the objectives pursued by AI systems circulating in the markets and the need for an effective protection of human rights.

Few expressions have been more echoed in regulatory initiatives in the field of artificial intelligence (AI) than *human-centric AI*.[1] The concept emerged as a key principle to govern

* Corresponding author.
E-mail address: baquero@hec.fr (P.M. Baquero).

[1] Most existing ethical principles and legal guidelines on AI consistently refer to the promotion of human values, international human rights or human control of technology. Jessica Fjeld and others, 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI' (2020) Berkman Klein Center Research Publication No 2020-1; The Art. 1.2 OECD AI Principles, for instance, refers to human-centred values imposing to "respect the rule of law, human rights and democratic values, throughout the AI system lifecycle" and, to than end, to "implement mechanisms and safeguards, such as capacity for human determination". 'OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449' <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449> accessed 3 March 2022; See also a more recent OECD report providing a comparative framework on the different tools that are being deployed in practice to implement trustworthy AI, 'Tools for Trustworthy AI: A Framework to Compare Implementation Tools for Trustworthy AI Systems', vol 312 (2021) OECD Digital Economy Papers 312 <https://www.oecd-ilibrary.org/science-and-technology/tools-for-trustworthy-ai_008232ec-en> accessed 18 April 2022; Similarly, the UNESCO Recommendation on Ethics of AI is abundant with references to humanity and the promotion of human rights as overarching principles; and particularly indicates the need for human oversight and determination over AI systems. 'UNESCO, Recommendation on the Ethics of Artificial Intelligence, SHS/BIO/REC-AIETHICS/2021' <https://unesdoc.unesco.org/ark:/48223/pf0000380455> accessed 3 March 2022; The relevant guidelines appearing in the context of European Union represent no exception to that. The Ethics Guidelines for Trustworthy AI refer to the need to ensure that AI systems are "human-centric, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare

https://doi.org/10.1016/j.clsr.2023.105795

AI systems,[2] with two main purposes – to ensure that human values are incorporated into the design of algorithms, and that humans do not lose control over automated systems.[3] Despite a promising potential, however, the principle has yet to overcome important limitations to effectively guarantee that AI systems will promote, or at least not degrade, humanity and human agency.[4]

Two crucial limitations remain notably understated if human-centric AI is meant to ensure trust in technology beyond safety and security concerns.[5] First, the excessive focus on accountability at the *design* stage of AI systems has overshadowed the fact that human values can be affected at different stages across the life cycle, including during the development and operational phases, when AI systems are not anymore under the control of designers. Second, current regulatory initiatives seeking to implement a human-centric approach to AI, such as the European proposals for an AI Regulation and an AI liability Directive, are market-driven rather than human rights-centric. Despite explicitly invoking the need to implementing human values into AI systems, the main objective pursued by these regulations is to set rules and standards to enable AI-based products and services to circulate in markets. They are therefore limited in their ability to actively promote human values, at most seeking to prevent blatant violations to them.

In this article, we argue for a twofold approach to tackle these limitations. First, we propose to adopt a co-evolutionary and life cycle approach to tackle the lack of accountability of AI systems. We show that this approach can help ensure accountability beyond the design stage by enabling meaningful human control and human-AI interaction across the entire lifecycle of the system. Second, we call for regulatory initiatives to balance the market-driven approach adopted so far by giving a more predominant role to human rights and by introducing explicitly the notion of proportionality test. This *rebalancing* would serve to handle conflicts between the objectives pursued by AI systems circulating in the markets and the need for an effective protection of human rights.

We divide this article in four parts. Part one examines the limitations of AI systems that single-handedly focus on promoting human values at the design stage. The promise beyond those limitations is a co-evolutionary approach that enables humans to constantly adapt and update AI systems in view of the challenges that may be identified in each of its different phases. Part two analyses the limitations of current regulatory initiatives promoting human-centric AI. We show that market-driven approaches seeking to facilitate the circulation of products and services that comply with minimal standards may be unsuitable to promote a human-centric perspective to AI. The promise beyond this limitation is a perspective where human rights become central for the regulation of artificial intelligence. Part three proposes to draw on the principle of proportionality to balance human values with the optimization objectives pursued by AI systems. Part four concludes.

1. The limits of design and promises of a life cycle and co-evolutionary approach

The promotion of human-centric AI has predominantly focused on the stage of *design* of AI systems to ensure compliance with human values,[6] somehow overlooking that threats to human values may occur at different stages across the life cycle. We show that an excessive emphasis on the *design* stage downplays the importance of promoting human values across the life cycle of AI systems. Luciano Floridi et al., examining how to ensure the compliance of AI systems with the requirements laid down by the proposed Artificial Intelligence Act of the European Union, argue that the life cycle of these systems should be envisioned as a process flow. This flow is composed of five key stages: the design, development, evaluation, operation and retirement of AI systems.[7] The Graph 1, below, illustrates these five stages.

In this paper we focus mostly on the first four stages, which are those related to the effective functioning of an AI system. In each stage, different aspects must be monitored to ensure that an AI system complies with the proposed EU AI Regulation. In the design stage, the model, data and variables to be employed by the AI system are defined, specifying the problem to be dealt with. The development stage involves different steps regarding the preparation of the data (sourcing, cleansing, analysis) and the training, validation and tuning of the model. In the evaluation stage, the dataset and models are analysed in detail against the testing dataset, examining its performance in relation to targets such as robustness

and freedom." 'Ethics Guidelines for Trustworthy AI' (2018) Text <https://ec.europa.eu/futurium/en/ai-alliance-consultation> accessed 3 March 2022 References along the same lines are to be found in the EU White Paper on Artificial Intelligence and in the EU Proposal for an AI Regulation.

[2] On the notion of AI systems, see David Restrepo-Amariles, 'Algorithmic Decision Systems: Automation and Machine Learning in the Public Administration', *The Cambridge Handbook of the Law of Algorithms* (CUP 2020).

[3] Joanna J Bryson and Andreas Theodorou, 'How Society Can Maintain Human-Centric Artificial Intelligence' in Marja Toivonen and Eveliina Saari (eds), *Human-Centered Digitalization and Services*, vol 19 (Springer Singapore 2019) 4.

[4] From one side, the danger comes from resorting to a short-term anthropocentric vision which, for instance, ignores environmental concerns relating to AI in favour of immediate benefits coming from its indiscriminate use. From the other side, the risk comes from employing a concept too broad to be workable in practice. Luciano Floridi, 'The European Legislation on AI: A Brief Analysis of Its Philosophical Approach' (2021) 34 Philosophy & Technology 215, 218.

[5] Roger Brownsword, *Law, Technology and Society: Reimagining the Regulatory Environment* (Routledge 2019) 11.

[6] Amitai Etzioni and Oren Etzioni, 'Designing AI Systems That Obey Our Laws and Values' (2016) 59 Communications of the ACM 29; Karen Yeung, Andrew Howes and Ganna Pogrebna, 'AI Governance by Human Rights–Centered Design, Deliberation, and Oversight: An End to Ethics Washing' in Markus D Dubber, Frank Pasquale and Sunit Das (eds), *The Oxford Handbook of Ethics of AI* (Oxford University Press 2020) <https://doi.org/10.1093/oxfordhb/9780190067397.013.5>.

[7] Luciano Floridi and others, 'CapAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act' (University of Oxford and University of Bologna 2022) 4064091 33 <https://papers.ssrn.com/abstract=4064091> accessed 13 April 2022.

Graph 1 – Lifecycle of AI Systems, adapted from Floridi et al. (2022).

and fairness. In the operation stage, the AI system is deployed in practice, and monitoring and feedback mechanisms are established, as well as procedures for problem-resolution and updates of the system. Finally, when AI systems are to be deactivated, in the retirement stage, it is important to observe the risks of deactivation and determine what to do with the residuals of the AI (such as the stored data).

The design stage has attracted significant interest from scholars and remains a central focus in policy discussions.[8] For instance, Ibo van de Poel provides a useful perspective by proposing an account of value embodiment to assess if designed AI systems embody a range of values such as those proposed by the EU High-Level Expert Group and the IEEE.[9] Similarly, Article 13 of the proposed EU AI Regulation establishes that 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". Hence, article 13 depicts the design stage as the crucial moment to ensure accountability of AI systems, including during the operational stage. However, this design-centric perspective to accountability has limitations, especially regarding AI systems based on machine learning. At the design stage, explainability of machine learning algorithms tends to focus on building constraints on the model-building process or approximating complex model to simpler ones.[10] Andrew Selbst and Solon Barocas propose to move beyond the design stage and implement interactive approaches that allow actual or potential users to interact with the system at the operational stage.[11] For instance, a person using an interactive interface would be able to see how each action, parameter and value of her file affect her credit score. Ideally this solution should not only be available to the potential beneficiary of a loan but also to those operating the system.

The limitations of design-centric approaches to ensure the accountability of AI systems can be illustrated by the discussions about algorithmic bias, which have often emphasized how AI systems can be *designed* to counteract algorithmic bias.[12] Bias can however be encountered in different stages of the lifecycle of AI systems, as illustrated in the Graph 2, below, in which we insert the different types of bias identified by Suresh & Guttag[13] in the life-cycle of AI systems proposed by Floridi et al.[14]

Suresh & Guttag identify seven types of bias and acknowledge they are found across different stages of the lifecycle of an AI system.[15] Some biases relate to the data collected itself – rather than to design of the algorithmic model. In that sense, historical bias exists when the data generated to feed a model presents a bias in itself –as when it reflects the fact that in society most people associate professions such as engineer to men and nurse to women. Representation bias can happen when the data sample does not adequately represent the targeted population to be assessed as, for instance, when certain subpopulations (e.g., women, or residents in a certain region of country) are not adequately statistically represented in the data sample. A measurement bias relates to the labels and features (often, proxies) used to associate individuals to broader categories or qualities (or constructs), such as when student success is only measured by GPA scores. Other biases can indeed relate to the design of the model and its training – such is the case of aggregations bias, learning bias or evaluation bias.[16] Finally, bias can also be found in the deployment phase of the AI system. This is the case when the system is potentially used for the purpose of solving a problem that it was not intended to address. For example, scoring systems to evaluate the risk of recidivism and aid a judge in the ambit of a parole sentence can generate biases when used for the purpose of measuring the length of a criminal sentence.

To a certain extent, regulatory initiatives seeking to govern AI systems or parts of it acknowledge that violations to human rights can occur across different stages of its life cycle. For instance, article 22 of the General Data Protection Regulation (GDPR) establishes the right of data subjects not to be subject to decision making based solely on automated processing if the resulting decision produces legal effects affecting the individual. In this sense, human control or human intervention on AI systems can be systemic or happen in the context of individual automated decisions.[17] Systemic control involves,

[8] Raja Chatila and John C Havens, 'The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems' in Maria Isabel Aldinhas Ferreira and others (eds), *Robotics and Well-Being*, vol 95 (Springer International Publishing 2019); PA Kroes and IR van de Poel, 'Can Technology Embody Values?' in P Kroes and PP Verbeek (eds), *The moral status of technical artefacts* (Springer Science 2014).

[9] Ibo van de Poel, 'Embedding Values in Artificial Intelligence (AI) Systems' (2020) 30 Minds and Machines 385.

[10] Andrew D Selbst and Solon Barocas, 'The Intuitive Appeal of Explainable Machines' (2018) 87 Fordham Law Review 1085.

[11] ibid 1115.

[12] See, for instance, Judith Simon, Pak-Hang Wong and Gernot Rieder, 'Algorithmic Bias and the Value Sensitive Design Approach' (2020) 9 Internet Policy Review <https://policyreview.info/concepts/algorithmic-bias> accessed 18 April 2022.

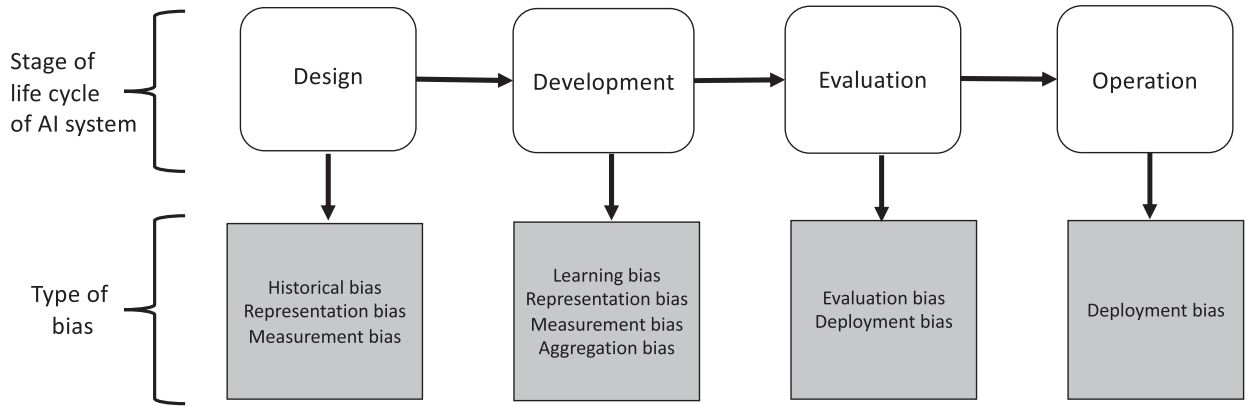
[13] Harini Suresh and John Guttag, 'A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle', *Equity and Access in Algorithms, Mechanisms, and Optimization* (ACM 2021) <https://dl.acm.org/doi/10.1145/3465416.3483305> accessed 14 April 2022.

[14] ibid.

[15] Suresh and Guttag (n 13).

[16] See ibid.

[17] Winston Maxwell, 'Comment Assurer l'efficacité Du Contrôle Humain Dans Les Systèmes de Décision Algorithmiques?', *Commission nationale consultative des droits de l'homme (CNCDH) 2022*



Graph 2 – Bias in the life cycle of AI systems.

for instance, determining the objectives to be optimized by a system using machine learning, selecting, cleansing and labeling the data, and prioritizing certain types of errors over others. Individual control involve humans supervising decisions or predictions generated by AI systems and potentially overturning decisions and recommendations.

Both types of control can take place at different stages of the life cycle of AI systems. Systemic control can be *ex ante*, before the system is deployed, in the stage of establishing the criteria and parameters of the algorithm. It can also be *ex post*, as in the right to question whether the recommendations generated by AI systems have used models, criteria and training data that are reliable and up to date – as determined by ECJ decisions on prediction of terrorist risks.[18]

Human control of individual AI decisions also takes place at different stages in the lifecycle of AI systems. It may be *ex ante* when the human has the ability to revise a recommendation or prediction generated by an AI system, and decide whether to follow the given recommendation or to diverge from it. The weight of the automated recommendation or prediction will depend to a great extent on the time and additional information available to the human when assessing the AI output. Individual control may be *ex post* as in the right of contesting purely automated AI decisions provided in Article 22 of the GDPR when the AI system processes personal data.

1.1. *Towards a co-evolutionary approach for meaningful human control*

Contemporary studies such as those of Joanna Bryson and Andreas Theodorou,[19] Lex Fridman et al. ,[20] and Filippo Santoni and Jeroen Van den Hoven[21] provide meaningful insights on how to consider human control of AI systems beyond the design stage and encompassing their entire life cycle. Bryson and Theodorou,[22] argue that, in addition to good design, humans can maintain control over AI systems by holding those who build, own, or operate them accountable through law and regulation. Fridman et al.[23] run a set of experiments to show that human supervision of AI systems, taken as blackboxes, can also be meaningful when it takes place at the operation and decision-making stages, and not only at the *design* stage. Finally, Filippo Santoni and Jeroen Van den Hoven,[24] reflecting about AI systems that take decisions autonomously, argue that to exercise meaningful human control over these specific type of AI systems two conditions needs to be satisfied. First, the systems should be *responsive* to the *human* moral reasons relevant in the circumstances in which it unfolds—independently of the number of levels, models, software, or devices that are part of it. Second, the actions of the system must be traceable to a proper moral understanding on the part of one or more humans involved in the design, programming, operation, and deployment of the system. All these studies highlight to some extent the need to monitor AI systems in a co-evolutionary perspective, creating revisable frameworks to allow human interaction and regulation at every stage of the life cycle of the system. The developments and implications of a co-evolutionary approach to AI systems, now widely discussed in different disciplines, still have not been closely considered from a regulatory perspective.

One of the most important challenges brough forth by the interdisciplinary literature regards how AI systems interact with the *human in the loop*[25] to explain complex aspects of how

(2022) <https://hal.archives-ouvertes.fr/hal-03544203> accessed 15 November 2022.

[18] See ECJ, Joint Cases C 511/18, C 512/18 and C 520/18. See also Winston Maxwell, 'La CJUE Dessine Le Noyau Dur d'une Future Regulation Des Algorithmes' [2020] Légipresse 671.

[19] Bryson and Theodorou (n 3) 4.

[20] Lex Fridman and others, 'Arguing Machines: Human Supervision of Black Box AI Systems That Make Life-Critical Decisions' (IEEE Computer Society 2019) <https://www.computer.org/csdl/proceedings-article/cvprw/2019/250600b335/1iTvoRJaiXe>.

[21] Filippo Santoni de Sio and Jeroen van den Hoven, 'Meaningful Human Control over Autonomous Systems: A Philosophical Account' (2018) 5 Frontiers in Robotics and AI 1.

[22] Bryson and Theodorou (n 3) 4.

[23] Fridman and others (n 20).

[24] Santoni de Sio and van den Hoven (n 21) 7–9.

[25] Eduardo Mosqueira-Rey and others, 'Human-in-the-Loop Machine Learning: A State of the Art' [2022] Artificial Intelligence Review <https://doi.org/10.1007/s10462-022-10246-w> accessed 14 November 2022; Yiwei Yang and others, 'A Study on Interaction in Human-In-The-Loop Machine Learning for Text Analytics', *Los Angeles* (2019).

automated outputs are generated.[26] There is concern that the human-AI interaction is reduced to a dynamic where the system dictates solutions to the human, who may have a role in implementing them in a particular environment. Instead, for an effective coupling between the AI system and the human to occur, there is a need for a joint human-AI cognitive system that provides the best reasons for action. The AI will not simply provide a solution to the human but constitute a resource to augment the cognitive capacity[27] of those tasked with the decision-making process[28] or exercising meaningful control over an automated systems.[29] Existing research mostly addresses this challenge by focusing on how to *design* joint human-machine cognitive systems that can truly support human decision-making.[30] However, in those studies there is a shift from a design that privileges the autonomy of system to one that focuses on human user experience with the objective of attaining this objective.[31]

Other disciplines are providing insights on how to establish a joint human-machine cognitive system beyond the design stage. In management studies, the frictions between design, business and engineering in complex situations, where problems are ill-defined, have led to the creation of a continuously revisable form of design, with the constant framing and reframing of particular problems.[32] The process of designing – not only products or services, but also the strategy of a company and the way people work[33] – becomes exploratory, with business, design and engineering decisions becoming intertwined.[34] More broadly, design studies increasingly come to value prototyping and reframing questions not with the objective of directly finding one solution, but to learn more about a problem and foster creativity.[35]

In computer sciences, co-evolutionary algorithms have been developed and applied in different fields to deal with complex problems that involve a significantly large number of variables.[36] The solution to these complex problems requires achieving different objectives simultaneously.[37] Since there is no single solution that can ideally optimize all objectives, the best trade-off between those must be found. To that end, the multiple objectives are decomposed into different lower-level components and assigned to different populations to be solved more easily according to the chosen criteria.[38] In a subsequent step, complete solutions to the problem are found by assembling the learning of representative individuals (in terms of their abilities to reach solutions) within each subpopulation.

In legal studies, similar exploratory perspectives can be found for instance in experimentalist approaches[39] to regulation[40] or contracts,[41] especially – but not exclusively – in the context of highly uncertain environments. In such scenarios, legal obligations are broadly defined and then specified jointly by the different involved actors locally and continuously revised considering the learning acquired. A co-evolutionary approach also recalls perspectives that view the law as a reflexive system, which is continuously influenced by different disciplines, and then reshaped internally according to its own logic.[42]

Regardless of the theoretical approach adopted, especially in the technological context, the practice sees that legal obligations are evermore specified according to an assessment of the level of risk found in each type of activity.[43] The recently proposed AI Regulation reflects this approach by regulating

[26] Tathagata Chakraborti, Sarath Sreedharan and Subbarao Kambhampati, 'The Emerging Landscape of Explainable Automated Planning & Decision Making', *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (2021); Víctor Rodríguez-Doncel and others, 'Introduction: A Hybrid Regulatory Framework and Technical Architecture for a Human-Centered and Explainable AI' in Víctor Rodríguez-Doncel and others (eds), *AI Approaches to the Complexity of Legal Systems XI-XII* (Springer International Publishing 2021).

[27] Nan-ning Zheng and others, 'Hybrid-Augmented Intelligence: Collaboration and Cognition' (2017) 18 Frontiers of Information Technology & Electronic Engineering 153.

[28] David D Woods, 'Cognitive Technologies: The Design of Joint Human-Machine Cognitive Systems' (1985) 6 AI Magazine 86.

[29] ibid.

[30] ibid 89; Ben Shneiderman, 'Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy' (2020) 36 International Journal of Human–Computer Interaction 495; Santoni de Sio and van den Hoven (n 21).

[31] Ben Shneiderman, 'Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems' (2020) 10 ACM Transactions on Interactive Intelligent Systems 26:1, 2.

[32] Tua Björklund and others, 'Integrating Design into Organizations: The Coevolution of Design Capabilities' (2020) 62 California Management Review 100, 103; Bec Paton and Kees Dorst, 'Briefing and Reframing: A Situated Practice' (2011) 32 Design Studies 573.

[33] Jon Kolko, 'Design Thinking Comes of Age' [2015] *Harvard Business Review* <https://hbr.org/2015/09/design-thinking-comes-of-age> accessed 4 March 2022.

[34] Björklund and others (n 32) 105–6.

[35] Mary Lou Maher and Josiah Poon, 'Modeling Design Exploration as Co-Evolution' (1996) 11 Computer-Aided Civil and Infrastructure Engineering 195; Masaki Suwa, John Gero and Terry Purcell, 'Unexpected Discoveries and S-Invention of Design Requirements: Important Vehicles for a Design Process' (2000) 21 Design Studies 539; Kees Dorst and Nigel Cross, 'Creativity in the Design Process: Co-Evolution of Problem–Solution' (2001) 22 Design Studies 425.

[36] Xiaoliang Ma and others, 'A Survey on Cooperative Co-Evolutionary Algorithms' (2019) 23 IEEE Transactions on Evolutionary Computation 421.

[37] Wen-Jing Hong, Peng Yang and Ke Tang, 'Evolutionary Computation for Large-Scale Multi-Objective Optimization: A Decade of Progresses' (2021) 18 International Journal of Automation and Computing 155.

[38] ibid 159.

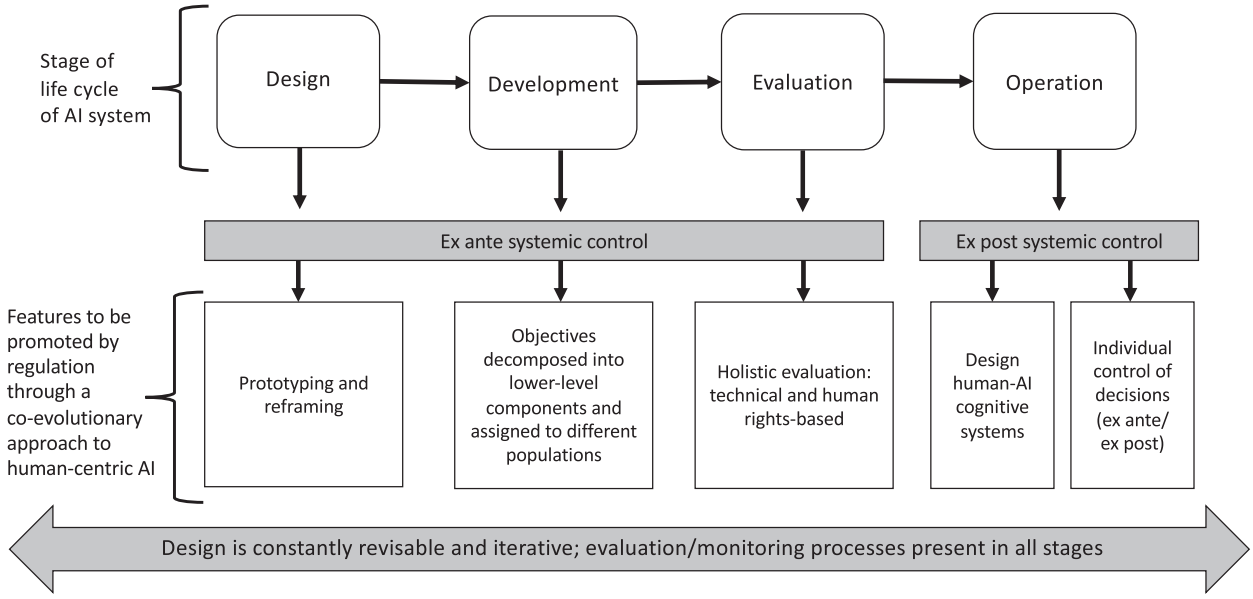
[39] Michael C Dorf and Charles F Sabel, 'A Constitution of Democratic Experimentalism' (1998) 98 Columbia Law Review 267.

[40] Charles Sabel and William Simon, 'Minimalism and Experimentalism in the Administrative State' (2011) 100 Geo. L. J. 53.

[41] Ronald J Gilson, Charles F Sabel and Robert E Scott, 'Contracting for Innovation: Vertical Disintegration and Interfirm Collaboration' (2009) 109 Columbia Law Review 431; Pablo Marcello Baquero, *Networks of Collaborative Contracts for Innovation* (Hart Publishing 2020).

[42] Gunther Teubner, 'Introduction to Autopoietic Law', *Autopoietic Law - A New Approach to Law and Society* (De Gruyter 1987).

[43] David Restrepo Amariles and Gregory Lewkowicz, 'Unpacking Smart Law: How Mathematics and Algorithms Are Reshaping the Legal Code in the Financial Sector' [2020] Lex Electronica 171, 177–8.



Graph 3 – Co-evolutionary approach to human-centric AI.

AI technologies according to the different levels of risk they present.[44] Under the Proposal, the definition of each level of risk, however, can be revised and regularly updated considering the state-of-art (except regarding prohibited AI practices).

In our perspective, current regulatory initiatives should capitalize on the insights from co-evolutionary approaches to make human-centric AI an operational concept in the regulation of AI systems. Graph 3, above, provides a comprehensive overview of the components of a co-evolutionary approach to human-centric AI and summarizes the ideas discussed above in this paper. This approach contains two main aspects. First, a framework to continuously revise the compliance of every stage of the life cycle of AI systems with laws and regulations. Second, a framework to enable human-AI interaction and equip humans with the tools to exercise meaningful control over the system or parts of it. Some of the features of this framework may include proto-typing and reframing at the design stage, collaborative development through disaggregation of objectives, holistic evaluation of AI system based on technical requirements and fundamental rights, and the implementation of human-AI cognitive systems to improve human control at the operational stage.

2. Limits of market-driven regulations and promises of human rights-based regulation

Most rules attempting to regulate AI systems stem out of market-driven regulations with the scope of facilitating trade by creating minimum standards for the circulation of products and services in a market. Even if they spell out a concept of human centric AI, their primary objective is not to preserve or enhance humanity, but at most trying to avoid extreme risks and affronts to human values. We propose that these limitations could be potentially overcome by regulations giving a more central role to fundamental rights.

The main purpose of current proposals to regulate AI in the European Union, such as the proposed AI Regulation and the AI liability Directive, is to promote the free trade of AI systems. These regulations seek to establish a set of rules and technical standards to which AI systems must abide to enter and circulate in the EU market and to be applied to assess non-contractual civil liability when a damage is caused with the involvement of AI systems. The legal basis of both proposals is Article 114 of the Treaty on the Functioning of the European Union which provides competence to legislate to the extent that it relates to the establishment and functioning of the European internal market. Principle 1 of the proposed AI Regulation determines that

> *The purpose of this Regulation is to improve the functioning of the internal market by laying down a uniform legal framework in particular for the development, marketing and use of artificial intelligence in conformity with Union values. This Regulation pursues a number of overriding reasons of public interest, such as a high level of protection of health, safety and fundamental rights, and it ensures the free movement of AI-based goods and services cross-border, thus preventing Member States from imposing restrictions on the development, marketing and use of AI systems, unless explicitly authorised by this Regulation.*

The initial draft of the Regulation follows a "risk-based approach and imposes regulatory burdens only when an AI system is likely to pose high risks to fundamental rights and safety".[45] AI systems are classified in different categories ac-

[44] Martin Ebers and others, 'The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)' (2021) 4 J Multidisciplinary Scientific Journal 589.

[45] European Commission, 'Laying Down Harmonised Rules on Artificial Intelligence - Explanatory Memorandum' (European Com-

cording to the risks they present, and certain requirement must be met by these systems and certain obligations followed by their operators. The content of these obligations, however, is meant to be specified by many technical standards issued by European standardization bodies.[46] The technical standards operationalize broader obligations concerning the implementation of AI such as human oversight, risk management and quality management systems. Even though these technical standards may often not be mandatory, they translate what in practice will be deemed sufficient to comply with the applicable legal principles and circulate in the EU market.

Risk regulation – the EU AI Regulation consists an example of that type of ordering – is a body of rules that "seeks to reduce the risks of harm to individuals and society stemming from all threats whether industrial or natural, voluntary or involuntary."[47] In the context of EU governance, decisions concerning risks are taken having in view scientific evidence along with social and ethical values.[48] The risk-based approach of the proposed AI Regulation seeks to contain risks to human rights and values by requiring compliance with standard requirements – however, it does not seek to mold AI systems to promote actively such values.[49]

The proposed Regulation evokes the functioning of the WTO Agreement, which establishes several rules and technical standards to facilitate the liberalization of trade. The WTO Agreement allows Member States to disregard these rules for overriding reasons of public interest (such as public health, environmental concerns or due process, for instance), only in very exceptional circumstances, listed in its Art XX ("General Exceptions").[50] In practice, the WTO case law has accepted measures restrictive to trade based on these exceptions in very few cases, construing the role of public interest in a limited way.[51] The proposed AI Regulation seems to adopt a similar design, focusing on establishing technical, market-driven standards that can only be disregarded for human rights and values in exceptional circumstances. The role of human rights becomes rather a defensive barrier against the abusive use of AI technologies rather than a way of conceptualizing and developing AI technologies in tune with human values.

The model of artificial intelligence regulation in China and the United States follows a same pattern, focusing on regulating the field through market-driven and technical standards.

In the United States, there is not a broad encompassing federal law to govern artificial intelligence in general in the same way that the EU Proposal for an AI Regulation does. Instead, different US agencies (e.g., the Federal Trade Commission and the Food and Drug Administration) and different states are currently pursuing the steps towards developing specific rules to govern AI in different domains.[52] Nevertheless, the White House Executive Order no. 13,859 – "Maintaining American Leadership in Artificial Intelligence"[53] makes an option to incentivize regulation through technical standards in its Section 1, (b), which states that:

> *(b) The United States must drive development of appropriate technical standards and reduce barriers to the safe testing and deployment of AI technologies in order to enable the creation of new AI-related industries and the adoption of AI by today's industries.*

Following that instruction, the US Congress, under the 2021 National Defense Authorization Act, has tasked the National Institute of Standards and Technology (NIST) to create a voluntary risk management framework for trustworthy AI systems. An initial draft of the Risk Management Framework, along with related documents, has been published in March 2022, providing guidance on the requirements that AI systems must fulfill across its life cycle to be compliant.[54] A specific example of the application of technical standards in the AI field concerns the "NISTIR 8292 supplement of the Face Recognition Vendor Test" used to assess the accuracy of automated face morph detection algorithms.[55]

A similar choice to pursue technical regulation of AI was made in China, where, in 2017, a "New Generation Artificial Intelligence Development Plan" has been published, establishing the goal for the country to become a "world leader in defining ethical norms and standards for AI".[56] While there has been a slow effort to attempt achieving that goal, the National New Generation Artificial Intelligence Governance Expert Committee in 2019 published 8 general principles that should be followed in the implementation of artificial intelligence (such as the common well-being, human rights, fairness and privacy). In order to establish more specifically how to operationalize those principles, however, the Standardization Administration of the People's Republic of China – an en-

mission 2021) COM(2021) 206 final 2.3; Martin Ebers, 'Standardizing AI: The Case of the European Commission's Proposal for an "Artificial Intelligence Act"' in Cristina Poncibò, Larry A DiMatteo and Michel Cannarsa (eds), *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics* (Cambridge University Press 2022).

[46] Arnaud Van Waeyenberge and David Restrepo Amariles, 'James Elliott Construction: A New (Ish) Approach to Judicial Review of Standardisation' (2017) 42 European Law Review 882; for an overview, see Stefano Nativi and Sarah De Nigris, 'AI Standardisation Landscape: State of Play and Link to the EC Proposal for an AI Regulatory Framework', (Publications Office of the European Union 2021) 17–21; Ebers (n 45).

[47] Alberto Alemanno, 'The Birth of the European Journal of Risk Regulation' (2010) 1 European Journal of Risk Regulation 1, 1.

[48] ibid 2.

[49] Pedro Rubim Borges Fortes, Pablo Marcello Baquero and David Restrepo Amariles, 'Artificial Intelligence Risks and Algorithmic Regulation' (2022) 13 European Journal of Risk Regulation 357.

[50] Petros Mavroidis, George Bermann and Mark Wu, *The Law of the World Trade Organization (WTO): Documents, Cases, and Analysis* (Thomson Reuters 2010) 684.

[51] ibid 687–8.

[52] For an overview, see Yoon Chae, 'U.S. AI Regulation Guide: Legislative Overview and Practical Considerations' (2020) 3 RAIL: The Journal of Robotics, Artificial Intelligence & Law 17.

[53] 'Maintaining American Leadership in Artificial Intelligence' (*Federal Register*, 14 February 2019) <https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence> accessed 18 April 2022.

[54] Elham (Fed) Tabassi, 'AI Risk Management Framework: Initial Draft - March 17, 2022' 23.

[55] Mei Ngan and others, 'NISTIR 8292 DRAFT SUPPLEMENT' 60.

[56] Huw Roberts and others, 'The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation' (2021) 36 AI & Society 59, 68.

tity responsible for laying out technical standards in China – published in 2019 a White Paper concerning Technical Standards for AI.[57]

The effort to develop technical standards to govern AI is not being only advanced from the side of national states. As previously mentioned, private and industry standardization associations have an important role in determining the applicable standards for the compliance of AI systems. A leading institution for the regulation of AI is the Institute of Electrical and Electronics Engineers (IEEE), a worldwide professional association, with branches in several countries, which has developed – and is currently developing rules for different domains, and providing certification for companies following the IEEE standards.[58]

In the end, technical, market-driven regulations are concerned with avoiding high-risk and safety issues potentially affecting humans, instructing engineers how to develop AI systems. Yet, it does not seem concerned with the development of a robust human-centric AI. Considering the new wave of computarization of society[59] and the key role AI plays in it, a truly human-centric perspective must go beyond the scope of risk regulation – based on preventing the occurrence of harm to humans – to develop a form of regulation that actively promotes fundamental rights in our digital society.

2.1. *Explainability and AI systems for legal decision-making*

The case of explainability of AI systems provides a good example of the limits of current regulatory approaches to promote fundamental rights. Take the case of the requirement of explainability applied to machine learning methods used to support legal decision making. Explainability provides that for an AI system to be suitable for human oversight they must provide explanations about how outcomes are generated.[60] For, without transparency about how machines operate,[61] humans cannot properly monitor the application of the technology and implement legal safeguards.[62]

There are two mainstream approaches regarding AI explainability, which focus on the technical means to provide explanations.[63] The first focuses on *designing* interpretable systems, which are simple or intuitive enough to be understood by humans without the need for further adaptation. The second approach is to adopt *post-hoc* methods to explain how complex systems, involving an extensive number of variables (such as deep learning methods), have generated specific outputs. Such *post-hoc* methods, however, provide only an approximated, simplified explanation of the functioning of AI methods, which are not fully explored nor understandable in detail. In the context of AI supporting legal decisions, however, both approaches neglect the importance of the legal basis upon which a particular legal prediction or recommendation has been generated.

The current discussion in legal studies about explainability has focused on the potential existence and scope of a requirement for explainability in the context of data protection regulation. In the context of the GDPR, the existence, scope and content of a potential requirement of explainability has been subject to controversy, with two main contrasting positions. One claims that the GDPR does not truly contain a general explainability requirement, but simply a right to have access to certain information connected to particular types of algorithmic decision-making.[64] The other conversely defends that the GDPR embodies a broader requirement of explainability, based on a holistic interpretation of the Regulation which takes into consideration its Recital 71, along with other provisions such as Arts. 13, 14, 15 and 22 GDPR.[65]

Underlying those discussions about the existence and scope of a requirement of explainability, however, there is a deeper and often overlooked limitation regarding this potential right under the GDPR: the type of explanations required focus on the technical methods of how a particular decision/prediction was made, rather than on providing an explanation about how the algorithmic system makes decisions in accordance with legal reasoning and sustained by an adequate legal basis. Without this type of "legal explainability", the technical explanations given are insufficient to assess whether a legal decision is in conformity with fundamental rights such as due process.

This gap could be filled by making fundamental rights central to the requirement of explainability. The duty to motivate judicial decisions has long been a largely accepted principle in the modern era, one that has been enshrined in the constitution of most national states and consistently been enforced by the highest courts of different countries.[66] In regard to the European Court of Justice, this principle is expressed in Article 36 of the Protocol no. 3 to the Treaty on the Functioning of the European Union (TFEU), establishing that judgments shall state the reasons on which they are based. The existence of a duty to state reasons has further been discussed in the context of the European Convention on Human Rights. The European

[57] 'Artificial Intelligence Standardization White Paper' (*Center for Security and Emerging Technology*) <https://cset.georgetown.edu/publication/artificial-intelligence-standardization-white-paper/> accessed 17 April 2022.

[58] 'IEEE CertifAIEd' <https://engagestandards.ieee.org/ieeecertifaied.html?utm_source=ieeesa&utm_medium=aem&utm_campaign=ais-2021> accessed 17 April 2022.

[59] Simon Nora and Alain Minc, *The Computerization of Society: A Report to the President of France* (MIT Press 1980).

[60] Selbst and Barocas (n 10).

[61] Frank Pasquale, *The Black Box Society* (HUP, 2016).

[62] Therese Enarsson, Lena Enqvist and Markus Naarttijarvi, 'Approaching the Human in the Loop - Legal Perspectives on Hybrid Human/ Algorithmic Decision-Making in Three Contexts' (2022) 31 Information & Communications Technology Law 123; Frank Pasquale, 'A Rule of Persons, Not Machines: The Limits of Legal Automation' (2019) 87 The George Washington Law Review 1.

[63] Selbst and Barocas (n 10) 110, 113. Other type of methods to provide explainability refer to interactive methods, see p. 117.

[64] Sandra Wachter, Brent Mittelstadt and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 1 International Data Privacy Law 76.

[65] Margot Kaminski, 'The Right to Explanation, Explained' (2019) 34 Berkeley Technology Law Journal 189.

[66] Ingrid Opdebeek and Stéphanie De Somer, 'The Duty to Give Reasons in the European Legal Area: a Mechanism for Transparent and Accountable Administrative Decision-Making? A Comparison of Belgian, Dutch, French and EU Administrative Law' (2016) 2016 Rocznik Administracji Publicznej 97, 97–8.

Court of Human Rights has interpreted different of its provisions as giving rise to a duty to give reasons whenever particular human rights are violated.[67] A number of cases have applied and analyzed the extent of this principle. In particular, the guarantee of access to justice has been emphasized in different ECJ cases.[68]

The resort to fundamental rights such as the right to due process or the right to have motivated decisions is an example of how human-centric regulation of AI could be developed based on a human rights framework. The challenge, however, concerns how to balance this enhanced human-centered perspective with the prescription of market-driven regulations and the industry needs for technical standards. That is where the law kicks back in.

3. The proportionality test: incorporating human rights in human-centric AI

To determine whether an AI system is human-centric, it should be required to pass a test that assesses the relation between what the system seeks to optimize and the impact such system can have on fundamental rights. We propose to draw on the principle of proportionality to examine, in each circumstance, whether a certain AI system fulfills this test.

To a certain extent, proportionality is already taken into consideration by the EU Proposal for an AI Regulation, when AI systems are classified according to different levels of risk and broadly regulated accordingly. In a more granular level, however, it would be necessary to consider the background and purposes of each AI system against its impact on fundamental rights to consider whether they are acceptable.

As noted by Smuha et al., who recognize the importance of necessity and proportionality tests in the context of the proposal of an AI Regulation in the European Union,

> *The Proposal offers little guidance on how proportionality is considered in the development, deployment and use of AI systems, other than the responsibilities of 'notified bodies' in relation to the certification procedure (Article 44(3)). While much of the enforcement of the Proposal relies on self-assessment by AI providers, little information is given as to their responsibilities regarding fundamental rights proportionality and necessity tests.*[69]

The principle of proportionality, however, has been more widely explored and elaborated in different legal domains, including in international law and WTO law.[70] It involves three requirements :[71] 1) adequacy – whether a certain measure attains its professed purpose; 2) necessity – whether there is an alternative less restrictive measure that could achieve that same objective; and 3) proportionality *strictu sensu* – whether the damage caused by the measure is not higher than the benefits achieved overall.

For instance, an AI system with a certain accuracy rate (for instance, 70%) may be justifiable or acceptable in certain environments, but not in others. Consider the example of a software making predictions concerning risks of credit default versus one used in the context of criminal justice, to predict risks of recidivism. In a context involving a significantly large amount of data available, both tools may provide to decision-makers an overview or guideline about what the data is conveying, with a more or less accurate rate. Whether a particular accuracy rate is *adequate*, however, will depend on the context; probably, a higher rate will be expected in the context of criminal justice, given the higher stakes involved in a criminal sentence. The test of adequacy will hinge upon the sensitivity of the context. Another example is black-box, non-explainable AI systems providing recommendations and predictions. Should human intervention be required? More sensitive decisions should require more human control and explainability – even at the cost of less accuracy – whereas in other fields it may be reasonable to pursue deep learning and opaque technical systems due to their technical accuracy.

Whether a particular AI system is *necessary* will depend on the other alternatives available – are there other non-automated and non-machine learning methods that could be used by the criminal justice, providing more reliable results? There may be, for instance, other algorithms available integrating fairness criteria or facilitating monitoring and compliance regarding non-discrimination. Finally, whether the AI system is proportional *strictu sensu* will depend on whether the overall outcome of applying the algorithm presents a better picture rather than not applying it at all. Would we have less conflict without its implementation? Are the time and costs saved worth of the challenges that have been generated by the automated system?

A proportionality test should become one of the fundamental pillars of any regulation seeking to implement a human-centric approach to AI. Not only it will ensure fundamental rights will remain at the center of our technological innovations but provide a reasonable framework to allow the co-existence of market and technical concerns with human values.

4. Conclusion

This article develops an approach through which a human-centric perspective to artificial intelligence can overcome two

[67] ibid 655.

[68] Case C 362/14, Maximillian Schrems v. Data Protection Commissioner, 2015 E.C.R. I 650, para. 95; Case C 682/15, Berlioz Investment Fund SA v. Directeur de l'administration des contributions directes, 2017 E.C.R. I 373, paras. 44 59; Opinion of Advocate General Cruz Villaln, supra note 31, at para. 43; Opinion of Advocate General Wathelet, supra note 27, at para. 67.

[69] Nathalie A Smuha and others, 'How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act' (5 August 2021) 38 <https://papers.ssrn.com/abstract=3899991> accessed 15 November 2022.

[70] Mads Andenas and Stefan Zleptnig, 'Proportionality: WTO Law: In Comparative Perspective International Academy of Commercial and Consumer Law Changing Law for Changing Times, 13th Biennial Meeting' (2006) 42 Texas International Law Journal 371.

[71] Lord Hoffman, 'The Influence of the European Principle of Proportionality upon UK Law', *The Principle of Proportionality in the Laws of Europe* (Hart Publishing 1999) <http://www.bloomsburycollections.com/book/the-principle-of-proportionality-in-the-laws-of-europe> accessed 18 April 2022.

crucial limitations to its implementation: one related to the excessive focus on accountability at the design stage – neglecting the need to promote accountability in other phases of the AI life cycle; and the other connected to the prevailing market-driven approach of AI regulations, which do not actively promote human rights, but at most prevent blatant violations against them.

In that sense, it provides a comprehensive overview of the components of a co-evolutionary approach to human-centric AI, composed of two main aspects. First, a framework to continuously revise the compliance of every stage of the life cycle of AI systems with laws and regulations. Second, a framework to enable human-AI interaction and equip humans with the tools to exercise meaningful control over the system or parts of it. This framework may include elements such as prototyping and reframing at the design stage, collaborative development through disaggregation of objectives, holistic evaluation of the AI system based on technical requirements and fundamental rights, and the implementation of human-AI cognitive systems to improve human control at the operational stage.

Regarding the limitations of market-driven regulations, it is proposed that they could be potentially overcome by regulations giving a more central role to fundamental rights. Despite some relevant distinctions, the model of artificial intelligence regulations in Europe, China and the United States focus on regulating the field through market-driven and technical standards, which, in the end, are concerned with avoiding high-risk and safety issues potentially affecting humans, instructing engineers how to develop AI systems. They do not seem primarily concerned with the development of a robust human-centric AI.

We elaborate on how this more robust perspective regarding human rights could, for example, be implemented in relation to the requirement of explainability of AI systems. While there are ongoing discussions about the existence and scope of a requirement of explainability under the GDPR, for instance, there is a deeper and often overlooked limitation: the type of explanations required focus on the technical methods of how a particular decision/prediction was made, rather than on providing an explanation about how the algorithmic system makes decisions in accordance with legal reasoning and sustained by an adequate legal basis. Without this type of "legal explainability", the technical explanations given are insufficient to assess whether a legal decision is in conformity with fundamental rights such as due process. This gap could be filled by making fundamental rights – such as a duty to motivate judicial decisions – central to the requirement of explainability.

Finally, we indicate that the feasibility of this human rights-based perspective hinges upon further elaboration on how a test of proportionality can be applied in this field, balancing market and human rights considerations to determine whether AI products and services can enter markets. Drawing on the literature on the test of proportionality in other fields, we analysed how the dimensions of adequacy, necessity and proportionality *strictu sensu* could be assessed in different scenarios involving algorithmic systems.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.