The limits of explainability & human oversight in the EU Commission's proposal for the Regulation on AI- a critical approach focusing on medical diagnostic systems

Routledge
Taylor & Francis Group

Information & Communications Technology Law

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cict20

The limits of explainability & human oversight in the EU Commission's proposal for the Regulation on AI- a critical approach focusing on medical diagnostic systems

Daria Onitiu

To cite this article: Daria Onitiu (2022): The limits of explainability & human oversight in the EU Commission's proposal for the Regulation on AI- a critical approach focusing on medical diagnostic systems, Information & Communications Technology Law, DOI: 10.1080/13600834.2022.2116354

To link to this article: https://doi.org/10.1080/13600834.2022.2116354

Published online: 29 Aug 2022.

INFORMATION & COMMUNICATIONS TECHNOLOGY LAW
https://doi.org/10.1080/13600834.2022.2116354

Routledge
Taylor & Francis Group

The limits of explainability & human oversight in the EU Commission's proposal for the Regulation on AI- a critical approach focusing on medical diagnostic systems

Daria Onitiu

UKRI Governance & Regulation Node at Trustworthy Autonomous Systems Programme, Edinburgh Law School, Old College, South Bridge, Edinburgh EH8, UK

ABSTRACT

The EU Commission's proposal for the Regulation on Artificial Intelligence, whilst providing important specifications on the importance of transparency of high-risk systems, falls short in providing a nuanced picture of how technical safeguards in Articles 13 and 14 in the proposal should be translated to AI systems operating on the ground. This paper focusing on medical diagnostic systems offers a perspective on how transparency safeguards should be applied in practice, considering the role of post hoc explainability and Uncertainty Estimates in medical imaging. Medical diagnostic systems offer probabilistic judgements regarding disease classification tasks, having an impact on the interactive experience between the doctor and the patient. Accordingly, we need additional guidance regarding Articles 13 and 14 in the proposal, considering the role of shared decision-making, and patient autonomy in healthcare and to ensure that technical safeguards secure medical diagnostic systems that are a safe, reliable, and trustworthy.

KEYWORDS

AI Act; transparency; post hoc explainability; medical diagnostic systems

1. Introduction

We need a socio-technical basis for regulating medical diagnostic systems. Medical diagnostic systems intend to act as a decision-support for disease classification tasks in medical imaging, such as detecting breast cancer or stages in diabetic retinopathy.[1] However, an important question is what happens when those AI systems enter real-life situations, interfering with clinical decision-making on the ground.[2] This paper intends

CONTACT Daria Onitiu donitiu@ed.ac.uk @DariaOnitiu

[1]Greg Russell, 'First for Scotland as hospital patients treated with artificial intelligence' *The National* (Glasgow, 29 September 2021) <www.thenational.scot/news/19611349.first-scotland-hospital-patients-treated-artificial-intelligence/> accessed 9 June 2022.

[2]Brent Mittelstadt, 'The Impact of Artificial Intelligence on the doctor-patient relationship' (Steering Committee for Human Rights in the fields of Biomedicine and Health (CDBIO) December 2021) <https://rm.coe.int/inf-2022-5-report-impact-of-ai-on-doctor-patient-relations-e/1680a68859> accessed 9 July 2022; Will Douglas Heaven, 'Google's medical AI was super accurate in a lab. Real life was a different story' (*MIT Technology Review*, 27 April 2020) <www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/> accessed 9 June 2022.

to contribute to research on the role of transparency and accountability regarding the verification of medical diagnostic tools, focusing on the role of explainability and human oversight to secure a medical diagnostic systems safe, reliable, and trustworthy use in the medical domain.

The paper focuses on Articles 13 and 14 of the EU Commission's proposal for the Regulation on Artificial Intelligence (The AI Act proposal) and whether the technical safeguards in the provisions include AI system's interaction with the doctor and the patient.[3] The AI Act proposal, as well as the High-Level Expert Group on Artificial Intelligence (AI HLEG) consider that notions of explainability, human agency and oversight are fundamental values that need to underpin a high-risk system's design and deployment.[4] I am giving a practical viewpoint on how Articles 13 and 14 of the AI Act proposal apply to popular post hoc explainability including visualisation methods, as well as Uncertainty Estimates with regard to the deployment of medical diagnostic systems. I argue that an AI system's enhanced visualisation of disease classification tasks requires manufacturers including AI providers to consider how these technical safeguards can induce individuals to translate probabilistic judgements into prescriptive decisions, undermining patient autonomy and shared decision-making.[5]

AI providers and users should view the notion of transparency as a 'way of thinking' regarding the design and deployment of medical diagnostic systems.[6] However, the AI Act proposal gives the original provider a lot of leeway in how transparency goals are to be implemented in practice. First, Articles 13–14 limit the notion of transparency to the system's performance metrics and accuracy as the health care professional's ability for risk management.[7] Second, AI Act proposal provides a perspective of transparency and human oversight which is limited to a system's intended use, leaving out the tool's interactive experience with the doctor and the patient. What we need is not a definition of holistic transparency which can be adapted to specific needs and ends, but a different perspective of the role of probabilistic judgements suiting the individual's own decision-making when operating an AI for decision-support on the ground.

This paper aims to translate technical safeguards in Articles 13–14 of the AI Act proposal into actionable principles regarding the use of medical diagnostic systems. We need to have a broader perspective of the socio-legal implications of technical safeguards, such as post hoc explainability for decision-making and Uncertainty Estimates for human oversight, to articulate an approach for human-centric governance and regulation on AI.[8] I highlight that an AI system's verification should include the value of risk management and communication as a qualitative metric to ensure trust. Referring to Article 13, I

[3]Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Legislative Act [2021] COM(2021) 206 final (hereafter 'Artificial Intelligence Act proposal') art 13, art 14.

[4]Artificial Intelligence Act proposal, art 13, art 14; EU High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (8 April 2019) pages 15–16, 18.

[5]Indeed, a 'user' may become a 'provider'considering the AI Act proposal; see AI Act Proposal, Recital 66; art 3 (1–4); see also, Lilian Edwards, 'Expert Opinion: Regulating AI in Europe: four problems and four solutions' (Ada Lovelace Institute 2022) <www.adalovelaceinstitute.org/report/regulating-ai-in-europe/> accessed 7 July 2022, pages 6–7.

[6]Anastasiya Kiseleva, Dimitris Kotzinos and Paul De Hert, 'Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations' (2022) 5 Frontiers in Artificial Intelligence 1, 6.

[7]Artificial Intelligence Act proposal, art 13, art 14.

[8]EU High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (n 4) page 37; cf Alessandro Mantelero, *Beyond Data: Human Rights, Ethical and Social Impact Assessment of AI* (Asser Press, 2022) 99.

suggest that an individual's ability to communicate probabilistic judgement is an important aspect for providers (and the users) to consider with regard to ensuring the transparency of medical diagnostic tools. In addition, I submit that Article 14 currently ensures an individual's passive position to assess an AI system's confidence levels, based on a tool's foreseeable risks concerning performance.

2. The AI act proposal and transparency for high-risk systems

The EU Commission's proposal for the Regulation on Artificial Intelligence (AI Act proposal) is argued to illustrate one of the 'most influential regulatory steps taken so far internationally' with regard to the formal governance of AI systems.[9] The proposal is a culmination of years of work entailing national regulators, as well as the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG).[10] The EU Commission attempts is to codify the EU's vision of trustworthy and human-centric AI,[11] and sets out horizontal rules applicable to the design and deployment of AI products in the EU market.

An important aspect of the AI Act proposal is its risk-based approach, including a layered certification mechanism for AI products.[12] The proposal envisages prohibited practices, specific rules for the deployment of high-risk systems, as well as minimum transparency rules for AI systems interacting with natural persons.[13] Most medical AI systems are considered high-risk systems, being subject to enhanced transparency requirements and human oversight, amongst others, but except for those AI approaches falling through the cracks of Medical Device Regulation (MDR), including wearable technology for fitness and wellbeing purposes.[14]

The transparency and human oversight obligations in Articles 13–14 of the AI Act proposal are important safeguards to ensure the deployment of high-risk systems, including medical AI products on the ground.[15] As argued by the AI HLEG, '[i]f we are increasingly going to use the assistance of or delegate decisions to AI systems, we need to make sure

[9]Luciano Floridi, 'The European Legislation on AI: A Brief Analysis of its Philosophical Approach' (2021) 34 (2) Philosophy & Technology 215.

[10]Philippe Dambly and Axel Beelen, 'Europe: Analysis of the Proposal for an AI Regulation' (Montreal AI Ethics Institute 2022) <https://montrealethics.ai/europe-analysis-of-the-proposal-for-an-ai-regulation/> accessed 25 February 2021.

[11]EU High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (n 4) page 4; Commission (EC), 'On Artificial Intelligence – A European approach to excellence and trust' (White Paper) COM(2020) 65 final, 19 February 2020, page 3.

[12]Mauritz Kop, 'EU Artificial Intelligence Act: The European Approach to AI' (2021) Transatlantic Antitrust and IPR Developments <https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/> accessed 7 June 2022.

[13]Artificial Intelligence Act proposal, art 5, Title III, art 52.

[14]Hannah van Kolfschooten, 'Conspicuous by its absence: health in the European Commission's Artificial Intelligence Act' (*The BMJ Opinion*, 30 July 2021) <https://blogs.bmj.com/bmj/2021/07/30/conspicuous-by-its-absence-health-in-the-european-commissions-artificial-intelligence-act/> accessed 25 February 2022; see also, Jérôme De Cooman, 'Humpty Dumpty and High-Risk AI Systems: The Ratione Materiae Dimension of the Proposal for an EU Artificial Intelligence Act' (2022) VI (1) Market and Competition Law Review 49; Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC [2020] OJ 117/1 (hereafter 'Medical Device Regulation').

[15]Indeed, there are other transparency obligations relating to labelling with regard to AI systems that are of limited risk. Nevertheless, a detailed discussion would be out of scope of this paper; for more discussion about transparency obligations regarding limited risk systems in the AI Act proposal, see Gianclaudio Malgieri and Marcello Ienca, 'The EU regulates AI but forgets to protect our mind' (*European Law Blog*, 7 July 2021) <https://europeanlawblog.eu/2021/07/07/the-eu-regulates-ai-but-forgets-to-protect-our-mind/> accessed 7 June 2022.

these systems are fair in their impact on people's lives, that they are in line with values that should not be compromised and able to act accordingly, and that suitable accountability processes can ensure this'.[16] Take the example of the IBM Watson natural language processing tool for personalised cancer treatment recommendations, which under-delivered in a healthcare environment and caused inconsistent decisions and insights regarding treatment recommendations.[17] Focusing on the design and deployment of safety-critical applications, including AI in healthcare, we are increasingly interested in evaluating and verifying the claims 'about' and 'by' the algorithm.[18] Defining the AI system's interactive experience with users and relevant stakeholders including operators in the ground is a regulatory and technical challenge, as well as a normative proposition regarding the continuous, reliable, safe and trustworthy use of AI in healthcare.

2.1. Right time, but wrong turn for medical diagnostic tools

How does the regulatory intervention conceptualise the level of risk to safeguard human-centric values? Looking at the AI Act proposal, we find that the proposal intends to establish a link between its approach to product safety and EU human-centric values. On the one hand, most provisions, including Articles 13 and 14 of the proposal, focus on the manufacturer or so-called "provider" to establish the technical safeguards regarding the AI system's continuous use on the ground.[19] On the other hand, the EU Commission intends to place 'people at the centre of the development of AI' including the individuals impacted by the AI systems.[20] Medical AI systems do not simply reinforce information asymmetries including patient vulnerability,[21] but shape the doctor-patient relationship within the algorithms' contours of decision-making. The EU Commission's balancing between the need for innovation and formal, as well as enhanced regulation, intends to establish an 'ecosystem of trust' regarding high-risk systems.[22]

However, it is less clear how that balance between product safety and human values plays out in practice focusing on medical diagnostic tools. For instance, the AI HLEG's guidance issues a checklist which includes important questions on human agency and oversight, amongst others.[23] One question stipulates how '… the AI system [could] affect human autonomy by interfering with the end-user's decision-making process in any other unintended and undesirable way'.[24] We need to put this question into the perspective of medical diagnostic tools in a healthcare setting, and ask how can algorithmic claims fit with medical reasoning and the patient values and needs? More guidance

[16]EU High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (n 4) page 9.
[17]Casey Ross and Ike Swetlitz, 'IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close' *STAT* (Boston, 5 September 2017) <www.statnews.com/2017/09/05/watson-ibm-cancer/> accessed 25 February 2021; Eliza Strickland, 'IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care' (2019) 56 (4) IEEE Spectrum 24, 29.
[18]Taken from, David Spiegelhalter, 'Should we Trust Algorithms?' (*HDSR: MIT Press*, 31 January 2020) <https://hdsr.mitpress.mit.edu/pub/56lnenzj/release/3> accessed 7 June 2022.
[19]cf Aritificial Intelligence Act proposal, art 14 (4); Edwards (n 5) 6.
[20]Commission (EC), 'Building Trust in Human-Centric Artificial Intelligence' (Communication) COM(2019) 168 final, 8 April 2019, page 2.
[21]See Hannah van Kolfschooten, 'EU Regulation of Artificial Intelligence: Challenges for Patient Rights' (2022) 59 CMLR 81.
[22]Commission (EC), 'On Artificial Intelligence – A European approach to excellence and trust' (n 11) page 3.
[23]EU High-Level Expert Group on Artificial Intelligence, 'The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment' (17th July 2020) page 7.
[24]ibid.

needs to establish the EU values' relevance to medical diagnostic tools to avoid that EU AI Act's proposal to take a wrong turn to a rigorous product safety approach.

Against this background, it is important to investigate how rules on transparency regarding high-risk systems in Articles 13 and 14 can promote notions of autonomy, agency, and human oversight of safety-critical applications, including medical diagnostic systems. I intend to contribute to research focusing on the role of technical safeguards ensuring transparency in medical diagnostic tools and how to ensure transparency goals within these values.

3. Article 13 and transparency and information disclosure

Article 13 provides a notion of holistic transparency that provides opportunities as well as challenges for legally trustworthy AI. On the one hand, Article 13 does 'not explicitly provide examples of degrees of transparency'.[25] On the other hand, it has been argued that Article 13 unduly prescribes the design and deployment of high-risk systems to system's use, leaving out 'those individuals coming into contact with providers and users of AI systems'.[26] What follows is that we need to interpret this provision as a delicate balancing act between the role of transparency to fulfil goals including explainability of high-risk systems and the individual's perspective and the system's usability concerning these technical safeguards.

Manufacturers ensuring the transparency of AI products focus on a system's intended use.[27] Article 13 (3) (b) mentions that the AI provider, designing the technical safeguards in Article 13 (1)–(2), outline the system's capacities and limitations with regard to the AI tools' purpose, performance, as well as foreseeable risks.[28] Indeed, the AI Act proposal establishes a post-market monitoring system for manufacturers to proactively collect and document changes to the AI systems that may increase the risks on individuals.[29] However, transparency remains predominantly an ex ante measure and there is a lot of leeway for manufacturers to adopt a robust 'product lifecycle approach' for unanticipated risks of AI products when operating in a healthcare setting.[30] For instance, manufacturers should specify which technical specifications support a state of 'shared

[25] Anastasiya Kiseleva, 'Making AI's Transparency transparent: notes on the EU Proposal for the AI Act' (*European Law Blog*, 29 July 2021) <https://europeanlawblog.eu/2021/07/29/making-ais-transparency-transparent-notes-on-the-eu-proposal-for-the-ai-act/#:~:text=Transparency%20%E2%80%93%20Interpretability&text=This%20type%20of%20AI%20system,output%20and%20use%20it%20appropriately> accessed 26 February 2021; Artificial Intelligence Act proposal, art 13 (1); EU High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (n 4) page 29.

[26] Nathalie Smuha, Emma Ahmed-Rengers, Adam Harkens, Wenglong Li, James MacLaren, Riccardo Piselli and Karen Yeung, 'A Response to the European Commission's Proposal for an Artificial Intelligence Act' (2021) LEADS Lab @ University of Birmingham, 35 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3899991> accessed 26 February 2021; compare with Article 29 which stipulates that '[u]sers of high-risk AI systems shall use the information provided under Article 13 to comply with their obligation to carry out a data protection impact assessment under Article 35 of Regulation (EU) 2016/679 or Article 27 of Directive (EU) 2016/680, where applicable'; Artificial Intelligence Act proposal, art 29; see also, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ 119/1, art 35 (6).

[27] Artificial Intelligence Act proposal, art 13 (3) (b).

[28] ibid, art 13 (b) (i), (iii), (iv).

[29] ibid, art 61.

[30] This is indeed a problem that has been recognised by the Food & Drug Administration regarding AI as medical device including adaptive algorithms; U.S Food and Drug Administration (FDA), 'Artificial Intelligence and Machine Learning in Software as a Medical Device Action Plan' (January 2021) <www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device> accessed 7 June 2022.

information'[31] between the doctor and the patient when those safeguards are operative in a specific setting.

Article 13 provides one important guidance for the AI provider to ensure a system's transparency which is that '[h]igh-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable individuals to interpret the system's output and use it appropriately … '.[32] We could argue that Article 13 mandates an explicit requirement regarding the explainability of AI systems.[33] To clarify, there is no uniform definition of explainability; however, we could say that Article 13's transparency and information duties mandate a form of transparency that allows the individual to comprehend the algorithms' observed effects.[34] Nevertheless, technical safeguards including explainability, whilst promoting the individual to make good explanatory decisions, do not necessarily promote the individual to make good exploratory actions.

To elaborate on this point, we need to make an important distinction between the role of explainability methods and the information duties in Article 13 (1)–(2),[35] whereby the latter includes risk communication, and the former includes both risk communication and risk management. The AI Act proposal, focusing on a system's intended use, does not draw a clear dividing line supporting an AI system's safe and reliable on the ground. I am going to investigate this further in the next Section, focusing on popular methods of post hoc explainability in medical imaging.

3.1. Widening the parameters of transparency

Why do we need post hoc explainability methods to understand the parameters of transparency in Article 13 of the AI Act proposal? Post hoc explainability are methods intended to assist AI decision-making, such as by showing feature importance in medical imaging.[36] By way of illustration, imagine a medical diagnostic system which can classify diabetic retinopathy in patients' retina and that can show the contribution of the input feature to the prediction including output.[37] What this shows is that explainability methods can produce

[31] According to Jens Christian Bjerring and Jacob Busch establishing a 'state of shared information … .is not tenable in light of the black-box nature of the machine learning decision-making'; Jens Christian Bjerring and Jacob Busch, 'Artificial Intelligence and Patient-Centred Decision-making' (2020) 34 (2) Philosophy & Technology 349, 351.

[32] Artificial Intelligence Act proposal, art 13 (1).

[33] See also, EU High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (n 4) page 29; EU High-Level Expert Group on Artificial Intelligence, 'The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment' (n 23) page 27.

[34] I am not going into a discussion defining explainability and interpretability of AI approaches. I admit that I adopt the approach taken by Cynthia Rudin, who argues that explanations are 'approximations to model predictions', whereas interpretability is inherent in the model itself. What this shows that explainability and interpretability can promote similar goals (i.e. structured information for the human decision-maker) based on a different degree of benchmarking; see Cynthia Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (ArXiv, 22 September 2019) <https://arxiv.org/pdf/1811.10154.pdf> accessed 9 March 2022, page 4; Ricards Marxinkevics and Julia E Vogt, 'Interpretability and Explainability: A Machine Learning Zoo Mini-tour' (ArXiv, 3 December 2020) <https://arxiv.org/pdf/2012.01805.pdf> accessed 10 March 2022; cf Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal, 'Explaining Explanations: An Overview of Interpretability of Machine Learning' (ArXiv, 3 February 2019) <https://arxiv.org/pdf/1806.00069.pdf> accessed 6 March 2022.

[35] Artificial Intelligence Act proposal, art 13 (2).

[36] Aniek F Markus, Jan A Kors and Peter R Rijnbeek, 'The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies' (2021) 113 Journal of Biomedical Informatics 1, 4.

[37] Amitojdeep Singh, Sourya Sengupta and Vasudevan Lakshminarayan, 'Explainable deep learning models in medical image analysis' (ArXiv, 28 May 2020) <https://arxiv.org/pdf/2005.13799.pdf> accessed 8 June 2022, page 4.

'trains of thought' helping the individual to receive a 'structured set of information' from the algorithms' decision-making process.[38] Article 13 exemplifies this process of the individual interpreting and understanding the system's performance parameters.[39] For example, the individual could see how the medical diagnostic system visualises some areas of the patient's retina, considering the tool's classification of mild diabetic retinopathy in the image.

Nevertheless, Article 13 does not mention the individual's ability to verify individual decisions.[40] Imagine the medical diagnostic system has been validated and verified for the classification of stages of diabetic retinopathy, whereby the system suggests a re-screening appointment for patient's suffering from mild diabetic retinopathy or a reference to an ophthalmologist for advanced diabetic retinopathy.[41] Assume now that the tool might have a high sensitivity and specifity, but it does not adequately show how patient with mild diabetic retinopathy could be vulnerable to experience further complications. Cynthia Rudin convincingly highlights that whether the model is relying on the correct input feature for a prediction is based on the *individual's perception* of the algorithms' interpretation and explanation of a prediction.[42] Article 13 focusing on the technical specifications that are 'pre-determined' by the manufacturer, only gives an account of the user monitoring the system's performance and instructions of use,[43] rather than the algorithms' decision-making when operating on the patient.[44]

We shall not argue that the health care professional's ability to evaluate individual decision is implicit in the manufacturer's implementation of post hoc explainability methods in medical imaging. Popular post hoc explainability including visualisation methods for medical imaging allow the operator including healthcare professional to make correlations transparent,[45] understand how predictions change,[46] as well as visualise input pixels influencing prediction.[47] Nevertheless, do these methods, which aim verify predictive goals, promote reliable decisions in individual circumstances? There is a lot of

[38]I took this terms from a paper by Dafna Shahaf, Carlos Guestrin and Eric Horvitz who write about information maps in imaging; Dafna Shahaf, Carlos Guestrin and Eric Horvitz, 'Trains of Thought: Generating Information Maps' (WWW 2012 – Session: Web Mining, Lyon, France, 16–20 April).

[39]Artificial Intelligence Act proposal, art 13 (3).

[40]I purposely refer to individual decisions, rather than predictions as methods, such as Local Interpretable Model-agnostic Explanations (LIME) focus on explaining individual predictions; see Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, '"Why Should I Trust You?" Explaining the Predictions of Any Classifier' (ArXiv, 9 August 2016) <https://arxiv.org/pdf/1602.04938.pdf> accessed 8 June 2022.

[41]Example taken from, 'FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems' (U.S Food & Drug Administration: Press Release 2018) <www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye> accessed 8 June 2022.

[42]Aaron M Bornstein, 'Is Artificial Intelligence Permanently Inscrutable? Despite new biology-like tools, some insist interpretation is impossible' (*Nautilus*, 29 August 2019) <https://nautil.us/is-artificial-intelligence-permanently-inscrutable-5116/> accessed 8 June 2022; Marzyeh Ghassemi, Luke Oakden-Rayner and Andrew L Beam, 'The false hope of current approaches to explainable artificial intelligence in health care' (2021) 3 (11) The Lancet 745, 746.

[43]See also Article 29 (4) of the AI Act proposal which stipulates that '[u]sers shall monitor the operation of the high-risk AI system on the basis of the instructions of use', Artificial Intelligence Act proposal, art 29 (4).

[44]ibid, art 13 (3) (c).

[45]This would entail a method called Deep Shapley Additive explanations (SHAP); Singh Sourya Sengupta and Lakshminarayan (n 37) page 5.

[46]In this respect, the paper by Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin propose the method that is called Local Interpretable Model-agnostic Explanations (LIME); Ribeiro, Singh, Guestrin (n 40).

[47]Here, I refer to saliency maps; see Mariana da Silva, 'Interpretable Deep Learning Part II: Visual Interpretability with Attribution Methods' (GitHub 2020) <https://metrics-lab.github.io/2020/10/08/visual-interpretability-with-attribution-methods.html> accessed 8 June 2022.

literature suggesting that this is not the case.[48] By way of illustration, imagine a medical diagnostic system which uses a saliency map that highlights the important pixels or regions in an image concerning the output.[49] We can show the constraints of post hoc explainability methods including saliency maps with regard to multi-class predictions, whereby the 'explanation heat map for multiple classes may be same … and the correct image area may be highlighted in the heat map even if the prediction is wrong'.[50] This type of constraint is evident in the classification of diseases entailing several symptoms.[51] For instance, Marianna da Silva explains that heat maps or saliency maps when applied to medical imaging tasks regarding Alzheimer or a complex brain disease will pick up focused lesions but not less common features, notwithstanding the feature's importance to explain decisions.[52] What follows is that Article 13 (1)–(2) assumes that technical safeguards and explainability as such can help the operator to use the system 'appropriately' and have 'complete' information about and based on the tool's intended use. [53]

Accordingly, we need to engage into widening the parameters of transparency to produce notions of accountability regarding medical diagnostic tools. Article 13 (1)-(2) offers a view of transparency that concentrates on the functional revelations, rather than the performative notion of predictive algorithms within the doctor-patient relationship. Functional revelations of a medical diagnostic system make the data and the algorithmic "thought process" less ubiquitous to the average operator but still not transparent to the individual implementing the real-world objectives in healthcare. We risk equating the individual's risk management and communication with a process that is retroactive, rather than interactive with the doctor and patient.

3.2. Neglecting patient autonomy

I argued in Section 3 that Article 13 promotes an outlook envisaging holistic transparency, requiring that the system should be 'sufficiently transparent' to the user.[54] Nevertheless, we need to add that another layer of transparency is how the individual defends his or her views to the patient when managing and communicating the risks of the AI system. Accordingly, we are interested in the system's transparency to stimulate the individual's interactive experience on the ground and to produce a dialectic tendency concerning patient needs and values.

[48]Alex John London, 'Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability' (2019) 49 (1) The Hastings Center Report 15, 19-20; Zachary Lipton, 'The mythos of model interpretability' (2018) 61 (10) Communications of the ACM 36, 41; Thomas P Quinn, Stephan Jacobs, Manisha Senadeera, Vuong Le and Simon Coghlan, 'The three ghosts of medical AI: Can the black-box present deliver?' (2022) 124 Artificial Intelligence in Medicine 1, 3; Ghassemi, Oakden-Rayner and Beam (n 42) 745; see also, Maya Krishnan, 'Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning' (2019) 33 (3) Philosophy & Technology 487, 492–493.

[49]Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, Julius Adebayo, Matthew D Li and Jayashree Kalpathy-Cramer, 'Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging' (2021) 3 (6) Radiology 1.

[50]Zohaib Salahuddin, Henry C Woodruff, Avishek Chatterjee and Philippe Lambin, 'Transparency of deep neural networks for medical image analysis: A review of interpretability methods' (2022) 140 Computers in Biology and Medicine 1, 11.

[51]da Silva (n 47).

[52]ibid.

[53]Artificial Intelligence Act proposal, art 13 (1), art 13 (3), art 13 (3) (c), art 29 (1).

[54]ibid, art 13 (1).

How do we implement technical safeguards, such as post hoc explainability methods, regarding medical diagnostic systems interacting with the doctor and the patient? An important aspect is that the health professional and the patient should not only understand the basic functionality of the AI system,[55] but it is the grasp of the model's feature importance, being relevant for the doctor and the patient when deciding on further recommendations for treatment. In this respect, a healthcare professional needs to respect patient autonomy, amongst others,[56] which includes the patient's informed and deliberate choice.[57] What follows is that a medical diagnostic system may impact an individual's values to form a deliberate choice regarding the AI systems implications for disease classification. For instance, if post hoc explainability methods are not based on instructions of use of the AI system that 'are concise, complete, correct and clear' to the relevant user,[58] this might disturb risk communication and patient autonomy. [59]

However, I would like to highlight that an AI system considering patient preferences is not an enabling condition for the respect for patient autonomy. Suppose that the medical diagnostic system's instructions of use highlight both, benefits, and risks of the tool's utility, including 'any known or foreseeable circumstance, related to the use of the high-risk AI system … .which may lead to risks to the health and safety or fundamental rights'.[60] Accordingly, the individual communicates that the medical diagnostic tool has a specific sensitivity and specificity for performance and that the model's saliency map will pick up a visual explanation about the regions in the image for the output. Finally, the individual re-assures that he or she will revisit the AI system's decision to avoid any risks of confirmation bias with the medical diagnostic tool.[61] This example exemplifies that a patient's appreciation of risk is consequential to the system's performance metrics.[62] However, patient autonomy is based on the individual's action to increase a patient's wellbeing, as well as enabling patient to act on the beliefs and values they hold.[63] What this shows is that a system's probabilistic judgements become the defining feature for the individual to evaluate treatment recommendations including a patient's values. Once probabilistic judgements become prescriptions, then patient autonomy is negated.

Therefore, we need to identify how should a healthcare professional communicate probabilistic judgements to a patient, considering the role of transparency and explainability in Article 13 of the AI Act proposal? Article 13 makes an important distinction

[55]Mittelstadt (n 2) page 46.
[56]For a useful framework on ethical principles, see Tom L Beauchamp and James F Childress, *Principles of Biomedical Ethics* (8th edn, OUP, 2019) 13.
[57]Michael Beil, Ingo Proft, Daniel van Heerden, Sigal Sviri and Peter Vernon van Heerden, 'Ethical considerations about artificial intelligence for prognostication in intensive care' (2019) 7 (1) Intensive Care Medicine Experimental 1, 5.
[58]I refrain to say "average user" as the manufactuer needs to provide 'an appropiate degree of transparency', Artificial Intelligence Act proposal, art 13 (1).
[59]Artificial Intelligence Act proposal, art 13 (1); Mittelstadt (n 2) page 18.
[60]Artificial Intelligence Act proposal, art 13 (3) (b) (iii).
[61]I will elaborate on the role of human oversight in Section 4; Artificial Intelligence Act proposal, art 13 (3) (d); Ghassemi, Oakden-Rayner and Beam (n 42) 746.
[62]See also, Frederico Cabitza who argues that reducing a system's reliability and validity to 'quantitative metrics' including error rates 'ethically problematic as accuracy at the level of the single medical case can be appraised only in hindsight'; Frederico Cabitza, 'How to evaluate the performance of AI by the bedside of the patient?' (Linkedin 2018) <www.linkedin.com/pulse/how-evaluate-performance-ai-bedside-patient-federico-cabitza/> accessed 27 February 2022.
[63]This definition fits with an approach of autonomy based on consequentialist thought, see G.T Laurie, SHE Harmon and G Porter, *Law & Medical Ethics* (10th edn, OUP, 2016) 6-7; see also, Carina Prunkl, 'Human autonomy in the age of artificial intelligence' (2022) 4 (2) Nature Machine Intelligence 99.

between the individual interpreting the system's output and the 'characteristics, capabilities and limitations of performance of the high-risk AI system', which includes the system's 'performance as regards the persons or groups of persons on which the system is intended to be used'.[64] The latter is framed as a question of design regarding high-risk systems and medical diagnostic systems. However, the manufacturer establishing the instructions of use needs to consider the system's intended use, as well as intended impact for risk communication and management, as a requirement of ex ante transparency.[65] In other words, how do system's evidence-based solutions including its risks to undermine evidence-based outcomes,[66] interact with the health care professional's duty to engage with patient-centred outcomes? This is an important aspect of risk management that is missing in the equation of Article 13's transparency goals which do not specify the role of post hoc explainability to manage clinically relevant outcomes, as well as a patient's interests. Therefore, we need to note that the proposal does not highlight the link between risk management and risk communication, neglecting the impact of medical diagnostic systems on patient autonomy.

4. Article 14 and the benchmark of transparency for human oversight

Article 14 provides another notion of transparency and accountability, which deals with the human decision-maker exercising discretion over algorithmic decision-making. In particular, it establishes a benchmark regarding medical diagnostic systems, which is measured by the degree of human oversight.[67] However, Article 14 of the proposal does add to the conditions of high-risk systems as ensuring transparency, underlining that human oversight 'shall aim at preventing or minimising the risks to health, safety or fundamental rights … .in particular when such risks persist *notwithstanding* the application of other requirements set out in this Chapter' including the foreseeable risks in Article 13 (emphasis added).[68] Article 14 intends to offer an account how the individual ought to address the role of an AI for decision-support, as well as the manufacturer's added safeguards to enable continuous risk management.

The AI HLEG specifies the role of human oversight regarding AI systems, whereby algorithmic decision-making should not undermine human agency.[69] Accordingly, the individual needs to fully recognise 'the capacities and limitations of the high-risk AI system' and effectively intervene with the algorithm decision-making process.[70] In addition, Article 14 indicates that algorithms need to augment decision-making, considering the risk of automation bias with regard to high-risk systems intended 'to provide information or recommendations for decisions to be taken by natural persons'.[71] The provision's main

[64]Artificial Intelligence Act proposal, art 13 (1), art 13 (3) (b), art 13 (3) (b) (iv).
[65]Indeed, this could be a point relevant for quality management; however, the notion of ex ante transparency and risk communication is not mentioned; Artificial Intelligence Act proposal, art 17 (1).
[66]See for example, Douglas Heaven (n 2).
[67]Article 14 focuses on the system's characteristics and intended use to define the parameters of transparency that are identified by the manufacturer, similarly, to Article 13; Artificial Intelligence Act proposal, art 14; see also, Aritificial Intelligence Act proposal, art 13 (3) (d).
[68]Compare with Article 14 (3) (a) which stipulates that technical measures should be 'identified and built, when technically feasible … '; ibid, art 14 (2), art 14 (3) (b).
[69]EU High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (n 4) pages 12-14.
[70]Artificial Intelligence Act proposal, art 14 (4) (a), art 14 (4) (d).
[71]ibid, art 14 (4) (b).

idea of human oversight is to ensure the expert-in-the-loop and individual agency regarding human operator's operation of high-risk systems.[72] Article 14 envisages technical safeguards that shall illustrate 'the knowledge and tools [for decision-makers] to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system'.[73]

Manufacturers have a lot of leeway to implement the parameters of transparency with regard to high-risk systems, notwithstanding the provisions' requirements to ensure transparency, human oversight and agency. This is understandable, as technical safeguards need to correspond to a specific socio-technical environment that is shaped by the human-AI interaction. By way of illustration, we might advocate an account of human oversight that is close to human control and active intervention regarding autonomous vehicles,[74] whereas medical diagnostic systems stimulate manufacturers to deal with methods that predominantly allow healthcare professionals to recognise gaps in algorithmic disease classification and intervene with regard to the system's output.[75] Focusing on the latter, Article 14 provides an interesting account of how technical safeguards could shape our understanding of diagnostic uncertainty in medical decision-making.

4.1. Defining uncertainty

Imagine a scenario where a medical diagnostic tool classifies the patient's mild diabetic retinopathy, as well as maps out some aspects of the patient's retina that support the system's decision. Article 14 (4) (d)-(e) underlines that the medical professional can not simply follow the AI system's recommendation without further insight into how the tool defines the degree of its reliance on the individual features.[76] Rather, the manufacturer needs to ensure that the healthcare professional can revisit and define the assumptions made by the AI system, as well as the probability of an outcome, including the disease relating event.

The aspect of defining and discussing factors of reliability is indeed an inherent aspect of medical decision-making. This process, entailing the medical professional scrutinising user perception to manage and communicate risk to the patient, can be defined as the role of diagnostic and prognostic uncertainty in medical practice.[77] Referring back to our example, the medical professional must reflect what is the appropriate degree of evidence to rely on an AI diagnostic decision? Additionally, what is the suitable degree of knowledge to accept a certain level of risk when the health care professional is

[72]ibid, art 14 (4) (e); EU High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (n 4) page 16.
[73]ibid.
[74]For example, an autonomous vehicle's switch control to a human operator could illustrate a design constraint; Riikka Koulu, 'Proceduralizing control and discretion: Human oversight in artificial intelligence policy' (2020) 27 (6) Maastricht Journal of European and Comparative Law 720, 728.
[75]Indeed, technical safeguards needs to safeguard notions of human agency, discretion and intervention within Article 14 and notwithstanding the specific high-risk system in question.
[76]Artificial Intelligence Act proposal, art 14 (4) (d–e).
[77]See for example, Ashley Graham Kennedy, 'Managing uncertainty in diagnostic practice' (2017) 23 (5) Journal of Evaluation in Clinical Practice 959.

recommending further treatment?[78] These are indeed questions which induce the medical professional to engage with the inherent notion of uncertainty in diagnostic and prognostic decisions. Article 14 (4) (d) of the proposal provides an interesting contribution to the debate regarding the professionals' managing uncertainty in what it mentions the risks of 'automation bias' regarding high-risk systems used as decision-support.[79] Accordingly, AI systems may provide a probability of risk, but it is the healthcare professional who maintains the role to effectively reduce the uncertainty and manage the risks for the patient.

Uncertainty estimates (or quantification methods) are indeed a measure that can be implemented by manufacturers, being 'technically feasible' regarding medical diagnostic systems.[80] First, Uncertainty Estimates illustrate an added safeguard to post hoc explainability methods in that it highlights 'which cases require further inspection by [a] specialist'.[81] In particular, it supports the operator to understand the estimate regarding the probability of risk focusing on the AI system's confidence score. Second, it is argued that Uncertainty Estimates help the operator to tailor the system's classification to individual circumstances.[82] For instance, a medical diagnostic tool may provide a prediction, as well as an Uncertainty Estimate to underline a certain risk regarding the diagnostic decision. Furthermore, a medical diagnostic tool may include a so-called 'rejector', whereby the algorithms would 'abstain' from making a decision where there is a 'large amount of uncertainty for a given patient'.[83] Referring back to our example, the medical professional may revisit clinical research and patient encounters, ask for a second-opinion or conduct further medical examinations, based on the system's 'prediction of high uncertainty' which can be above a referenced threshold.[84]

Accordingly, Uncertainty Estimates are a way to show the system's limitations regarding computer-aided diagnosis, within the parameters of Articles 14 (1) and (4) (a) of the AI

[78]On the risks of uncertainty regarding prognosis see Alexander K Smith, Douglas B White and Robert M Arnold, 'Uncertainty: The Other Side of Prognosis' (2013) 368 (26) The New England Journal of Medicine 2448.

[79]Artificial Intelligence Act proposal, art 14 (4) (d).

[80]In general, there are downsides of using (traditional) Bayesian approaches over more advanced methods in a deep learning model, such as Monte Carlo Drop-Out (MCDO); however, a detailed discussion goes beyond the scope of this paper. Please consult; Murat Seckin Ayhan, Laura Kühlewein, Gulnar Aliyeva, Werner Inhoffen, Focke Ziemssen and Philipp Berens, 'Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection' (2020) 64 Medical Image Analysis 1.

[81]Teresa Araújo, Guilherme Aresta, Luís Mendonça, Susana Penas, Carolina Maia, Ângela Carneiro, Ana Maria Mendonça, and Aurélio Campilho, 'Dr|GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images' (2020) 63 Medical Image Analysis 1, 3; Thomas Grote, 'Trustworthy medical AI systems need to know when they don't know' (2021) 47 (5) BMJ Publishing Group 337.

[82]Marilia Barandas, Duarte Folgado, Ricardo Santos, Raquel Simão, Hugo Gamboa, 'Uncertainty-Based Rejection in Machine Learning: Implications for Model Development and Interpretability' (2022) 11 (3) Electronics 1, 7.

[83]These are 'selective prediction models', see Benjamin Kompa, Jasper Snoek and Andrew L Beam, 'Second opinion needed: communicating uncertainty in medical machine learning' (2021) 4 (1) NPJ digital medicine 1, 2; Mike Miliard, 'New AI diagnostic tool knows when to defer to a human, MIT researchers say' (Healthcare IT News 2020) <www.healthcareitnews.com/news/new-ai-diagnostic-tool-knows-when-defer-human-mit-researchers-say> accessed 2 March 2021; reference to, Hussein Mozannar and David Sontag, 'Consistent Estimators for Learning to Defer to an Expert' (ArXiv, 25 January 2021) <https://arxiv.org/pdf/2006.01862.pdf> accessed 2 March 2022; I will provide some thoughts about selective prediction models considering Article 14 of the AI Act proposal in the conclusion in Section 5.

[84]Marilia Barandas, Duarte Folgado, Ricardo Santos, Raquel Simão, Hugo Gamboa, 'Uncertainty-Based Rejection in Machine Learning: Implications for Model Development and Interpretability' (2022) 11 (3) Electronics 1, 7; see also, Viraj Bhise, Suja S Rajan, Dean F Sittig, Robert O Morgan, Pooja Cuadhary and Hardeep Singh, 'Defining and Measuring Diagnostic Uncertainty in Medicine: A Systematic Review' (2017) 33 (1) Journal of General Internal Medicine 103; Cf. Hendrik Kempt and Saskia K Nagel, 'Responsibility, second opinions and peerdisagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts' (2021) 48 (4) Journal of Medical Ethics 222, 226.

Act proposal.[85] For example, Uncertainty Estimates can specify so-called epistemic uncertainty, which is the model's uncertainty that can be addressed with more training data.[86] What this shows is that the role of Uncertainty Estimate shifts the individual's focus from the system's accuracy to statistical reliability, having a more nuanced picture of the role of AI in safety-critical applications. By way of illustration, a healthcare professional using a medical diagnostic system regarding the classification of stages of diabetic retinopathy needs to navigate between various treatment options, including addressing the patient's mild diabetic retinopathy, which can entail differing risks and benefits for a health-related outcome.[87]

4.2. Defining knowledge

One must issue a warning when using Uncertainty Estimates as a methodology, considering Article 14 of the AI Act proposal.[88] Technical safeguards *as such* do not offer users a targeted approach to maintain the human oversight regarding AI as decision-support.[89] What this shows is that indeed, manufacturers need to implement technical safeguards, which are supplemented by methods for the decision-maker to quantify the balance between benefit and risk to be an effective tool for risk management regarding high-risk systems.

As a first step, this requires us to elaborate on the extent medical diagnostic systems define probabilistic judgements regarding Uncertainty Estimates. An important consideration is that Uncertainty Estimates do *not* work descriptively and computer scientists in fact test new observations on an underlying dataset.[90] Moreover, computer scientists use Uncertainty Estimates, such as Uncertainty Quantification methods, to study the 'reliability of scientific inferences'.[91] Following this thought process, the computer scientist can observe a high (epistemic) uncertainty in those spaces 'where there are few or no observations for training'.[92] What this shows that Uncertainties Estimates are a way of

[85]Artificial Intelligence Act proposal, art 14 (1); art 14 (4) (a); However, in Section 4.ii and iii I am going to highlight that technical safeguards do not enable users to '*fully* understand the capacities and limitations of the high-risk AI system' (emphasis added), Artificial intelligence Act proposal, art 14 (4) (a).

[86]The ML community also enumarates a second type of uncertainty which is irreducible and can be caused by 'noise in the observations' (i.e. aleatoric uncertainty). Taken from, Seckin Ayhan, Kühlwein, Aliyeva, Inhoffen, Ziemssen and Berens (n 80) 2; Talha Siddique, Md Shaad Mahmud, Amy M Keesee, Chigomezyo M Ngwira and Hyunju Connor, 'A Survey of Uncertainty Quantification in Machine Learning for Space Weather Prediction' (2022) 12 (1) Geosciences (Basel) 1,5; see also, Philipp Seeböck, Jose Ignacio Orlando, Thomas Schlegl, Sebastian M Waldstein, Hrvoje Bogunovic, Sophie Klimscha, Georg Langs and Ursula Schmidt-Erfurth, 'Exploiting Epistemic Uncertainty of Anatomy Segmentation for Anomaly Detection in Retinal OCT' (2020) 39 (1) IEEE Transactions on Medical Imaging 87, 88.

[87]Michela Assale, Silvia Bordogna and Federico Cabitza, 'Vague Visualizations to Reduce Quantification Bias in Shared Medical Decision Making' (Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Valletta, Malta, February 2020).

[88]Artificial Intelligence Act proposal, art 14.

[89]cf Dae Y Kanf, Pamela N DeYoung, Justin Tantiongloc, Todd P Coleman, Robert L Owens, 'Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine' (2021) NPJ Digital Medicine 142.

[90]For instance, Michela Assale, Silvia Bordogna and Federico Cabitza argue that 'render[ing] signs or symptoms in terms of numbers on ordinal scales, or clear-cut categories, does not make them more objective or free from noise, error and uncertainty'; taken from Assale, Bordogna and Cabitza (n 86) page 3.

[91]Talha Siddique, Md Shaad Mahmud, Amy M Keesee, Chigomezyo M Ngwira and Hyunju Connor, 'A Survey of Uncertainty Quantification in Machine Learning for Space Weather Prediction' (2022) 12 (1) Geosciences (Basel) 1, 5.

[92]Michel Kana, 'Uncertainty in Deep Learning. How To Measure?' (*Towards Data Science*, 26 April 2020) <https://towardsdatascience.com/my-deep-learning-model-says-sorry-i-dont-know-the-answer-that-s-absolutely-ok-50ffa562cb0b> accessed 6 April 2021; see also Benjamin Kompa, Jasper Snoek and Andrew L Beam who highlight that '

getting closer to knowledge, whilst not exhausting all possibilities on the lack of knowledge in a diagnostic domain.

In other words, much of the exercise of what constitutes a reliable estimate of risk and uncertainty is done by the individual who is deciding upon the risk or uncertainty for positive action.[93] If we accept this view, then we conclude that all probabilistic prepositions are amenable to the context through which they emerge.[94] For instance, a statement, assuming that we need another health professional's opinion with regard to the model's a higher uncertainty in a diagnostic setting and based on a higher risk of a wrong classification,[95] still leaves a gap for us to define the role of medical decision-making to promote patient-centred outcomes. Hence, how we judge uncertainty does not equal our conclusion about what is justified in individual circumstances. Technical safeguards in Article 14 only include what could be a higher level of risk regarding a system's operation, rather than the individual's interpretation of uncertainty and risk. What follows that Uncertainty Estimates and human oversight as a technical specification do not adequately allow the human operator to 'abstain from a decision' which will maximise the patient's health and/or wellbeing.

Article 14 defines how the manufacturer needs to articulate "risk" and "uncertainty" using technical safeguards, leaving out the role of the individual to interpret the role of uncertainty as a boundary exercise entailing risk management in clinical practice. Therefore, I intend to show in the next Section that Article 14 promotes a rather superficial view of human oversight, based on the individual's ability to identify the system's intended use as a guarantee for medical diagnostic tool's confidence levels.

4.3. Quantifying shared decision-making

A related aspect I consider worth underlining regarding Article 14's application with regard to the use of medical diagnostic systems on the ground is based on the provision's construction of transparency and accountability. Quantifying the algorithms' observed effects, including uncertainty, does not replace shared decision-making in medical practice. Hence, Article 14 needs to stimulate broader discussion on a high-risk system's alignment with ethical principles, as well as fundamental rights in the medical domain.

A healthcare professional, when reading a medical image, needs to consider that several individuals can make a different hypothesis about a patient. Indeed, Article 14 of the AI Act proposal, highlighting technical specifications enabling the human operator to '*fully* understand the capacities and limitations of the system' (emphasis added), does not confront the individual with a full grasp of uncertainty. However, it is important to note that Article 14, whilst helping the individual to translate some probabilistic judgements for risk management, does not support decision-making, promoting patient-centred outcomes.

… a test point far from training data should result in a higher amount of predictive uncertainty'; Kompa, Snoek and Beam (n 83) 2.

[93] See also, Savas L. Tsohatzidis, *Interpreting J.L Austin* (CUP, 2017) 98-99.

[94] Cf. Karl Popper's concept of 'falsiability' regarding scientific statements; Karl Popper, *The Logic of Scientific Discovery* (1st edn, Routledge Classics 2002) 17-18.

[95] Example taken from Seckin Ayhan, Kühlwein, Aliyeva, Inhoffen, Ziemssen and Berens (n 80) 8.

We can elaborate on this statement based on the connection between patient autonomy and shared decision-making in medical ethics and practice. In this respect, we describe the doctor-patient relationship as a 'decisional process' regarding the most suitable treatment for the patient.[96] The notion of shared decision-making illustrates a move-away from paternalistic decision-making.[97] Accordingly, it is about the health care professional communicating the standard of evidence and acting upon the patient's best interests.[98] Focusing on Article 14 (2) of the proposal, the standard of deliberation how Uncertainty Estimate fit with the patient's understanding of aggregated evidence is only addressed based on the manufacturer's obligation to deploy AI products that minimise their impact on safety, fundamental rights within the system's intended use. However, we know that principles of shared decision-making and patient autonomy do not operate in a vacuum but need to be judged by the clinical expertise, practitioners' differing experience and patient interests defining the acceptable standard of evidence.[99]

The AI Act proposal seems to take a different approach to the significance of human agency and oversight, focusing on the individual who takes a passive role in communicating risks about the implications of high-risk systems. By way of illustration, Article 14 (2) describes the role of foreseeable incidents, which might lead to situations where there is a risk of error of which the individual is *not* aware. Conversely then, a healthcare professional can argue that he or she is aware of all foreseeable risks, when a system shows an estimate for a given disease. [100] However, this deductive approach to uncertainty, which suggests that lower uncertainty will lower foreseeable risks, is not effective in most instances where there is a risk of harm that has not a readily identifiable cause. One cannot witness any anomaly in the system's operation if the healthcare professional, based on the system's reporting of an uncertainty intends to conduct more tests and ask for a second opinion. Once we acknowledged this, then there is no room for the healthcare professional to engage in a process of introspection and acknowledge the risks of complication and management concerning a patient's well-being and treatment of an illness.

In other words, Article 14 does have a grasp of the value of the modalities of algorithms to create foreseeable risks based on the system's limitations. However, it does not elaborate on the role of technical specifications for the healthcare professional to communicate risks and uncertainty to the patient. Article 14 addresses a different aspect of transparency, which is not related to human agency but rather to the model's erroneous distribution of thresholds. This aspect of transparency covers the obligations of manufacturers ensuring the system's performance once these are actionable to users, using

[96]Lars Sandman and Christian Munthe, 'Shared Decision Making, Paternalism and Patient Choice' (2009) 18 (1) Health care analysis 60, 61.

[97]Stefano Triberti, Ilaria Durosini and Gabriela Pravettoni, 'A "Third Wheel" Effect in Health Decision Making Involving Artificial Entities: A Psychological Perspective' (2020) 8 Frontiers in Public health 1, 2.

[98]For the connection between patient autonomy, shared decision-making as well as benefecience regarding AI as decision-support, see Sune Holm, 'Handle with care: Assessing performance measures of medical AI for shared clinical decision-making' (2022) 36 (2) Bioethics 178, 183.

[99]As argued by Juan Manuel Durán and Karin Rolanda Jongsma, '[d]iagnostic and treatment decisions are fundamentally evaluative judgements for which risks and uncertainties have to be weighed against a backdrop of medical knowledge, expert knowledge and Intuitions'; Juan Manuel Durán and Karin Rolanda Jongsma, 'Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI' (2021) 47 (5) Journal of Medical Ethics 329, 333; see also, Paul J Christine and Lauris C Kaldjian, 'Communicating Evidence in Shared Decision Making' (2013) 15 (1) The Virtual Mentor 9.

[100]This example is taken from Thomas Grote's outlook on Uncertainty Quantification, Grote (n 81) 337.

validation methods, as well as usability studies.[101] Human oversight and agency illustrates a different facet of transparency, which includes the degree of permissive thresholds to manage risks that are translated into intangible risks for patient autonomy and safety.

In addition, we could argue that Article 14, unduly focusing on the technical specifications for ensuring human oversight, does not address the distribution of uncertainty and risk as taken *by the human decision-maker*. Article 14 (4) (a–e) lists several measures which intend to support 'individuals to whom human oversight is assigned' to monitor the system's performance. However, the provision does not elaborate an important distinction between the quantification of risks and the role of permissible risks in individual circumstances. By way of illustration, we might agree that the individual relying on certain decisions, such as the risk of death, requires a high-level of certainty from the AI system. Conversely, we also agree that the individual relying on certain decisions, such as promoting patient's wellbeing or limiting suffering, might engage with a different balance to manage uncertainty and risk with regard to two people suffering from an identical disease but with different values and needs.[102] Accordingly, the individual engaging with the AI needs to understand both, the probabilistic account of risk and uncertainty, as well as his or her own engagement with risk when intervening with the AI-decision. Nevertheless, Article 14 only lists measures that aim to highlight the system's capacities and limitation regarding its performance, rather its reliability to promote the human operator's deliberate choice to manage and communicate risk to the patient.

To summarise, it is this transition of the system's account of uncertainty to the individual's inferential judgement to define human agency regarding medical diagnostic systems that is not elaborated by Article 14. Indeed, it is not the role of the AI Act proposal to flesh out how medical diagnostic systems need to correspond to ethical principles, beyond the manufacturers' duty to ensure human oversight.[103] However, those parameters need to be discussed with a view of the AI system converting value judgements in human decision-making.[104]In other words, how do we translate generalisations on the value of autonomy, agency and oversight into actionable values is an important regulatory task for ensuring that high-risk systems are safe, reliable and trustworthy in a medical setting. We need more clarification how risk and uncertainty should be read by the individual engaging with a medical diagnostic system to inform the role of effective intervention with AI, rather than insisting for manufacturers to quantify the meaning of human oversight applied on the ground.

[101]As argued by David S Watson, interpretability needs to 'quantify expected error rates' and currently 'it is impossible to subject algorithmic explanations to severe tests, as it is required of any scientific hypothesis'. Taken from, David S Watson, 'Conceptual challenges for interpretable machine learning' [2020] Synthese 1; This is indeed an important question which; however, is relevant for robust clinical evaluation, rather than transparency goals in human agency and oversight. A detailed account of the role of performance metrics and interpretability methods to validate algorithmic decision-making is beyond the scope of this paper.

[102]Stephan Loftus highlights that 'e case-based nature of clinical reasoning means that two people with an identical biomedical diagnosis may have to be managed quite differently. For example, one patient may be a young adult who is otherwise healthy while the other patient may be a very elderly, frail person with many comorbidities. One may receive aggressive treatment and the other may get palliative care'; Stephan Loftus, 'Thinking Like a Scientist and Thinking Like a Doctor' (2017) 28 (1) Medical Science Educator 251, 252.

[103]Nevertheless, the AI Act does underline that AI products correspond to ethical norms and EU values; Commission (EC), 'On Artificial Intelligence – A European approach to excellence and trust' (n 11) page 8.

[104]Rosalind J McDougall, 'Computer knows best? The need for value-flexibility in medical AI' (2019) 45 (3) Journal of Medical Ethics 156.

5. Concluding thoughts

In praise of a new reality of safety-critical applications seamlessly integrated into a healthcare setting, we seem to ignore the complications of medical diagnostic systems to shape values of shared decision-making and patient autonomy. The first part of the discussion intends to identify what kind of knowledge is required for a healthcare professional to interpret the system's output. When we speak of post hoc explainability, we tend to assume that algorithmic processes are a replica of scientific knowledge building. By way of illustration, we can mention visualisation methods in medical imaging to acknowledge the algorithms' observed patterns of an illness. However, this understanding of transparency and accountability assumes a functional setting, including a close alignment between human and "machine" reasoning. Building explainability methods with the sole aim of being on par with human judgement is a way to conflate functional revelations with normative propositions in medical decision-making. Article 13 risks limiting the healthcare professional's duty of risk communication, equating the individual's quantification of knowledge with the patient's perception of risk.

There needs to be a shift in academic discourse debating the notion of transparency as proof that medical diagnostics systems provide real-world validation of user claims. In other words, Article 13 requirements need not only consider patient perspectives and needs. Rather, post hoc explainability methods need to focus on reconciling human expertise with patient-centred values. I underline that Article 13 needs to consider the role of healthcare professionals in interpreting the system's output, whilst making normative propositions about the nature of risk communication in a clinical setting.

How do providers and manufacturers engage with proactive explainability, considering Article 13? One important aspect is that the system's appropriate 'degree of transparency'[105] is not dependent on performance metrics but manufacturers need to further specify the medical diagnostic tool's within a specific setting. In other words, the provider needs to document the circumstances that allow a high-risk system to fulfil its goals, such as moving beyond the AI system's role as a crude disease classification task. For example, we could think of manufacturers engaging with 'qualitative assessment' of post hoc explainability methods, such as saliency maps, which would be an ex ante transparency measure.[106] In doing so; however, we must establish the qualitative criteria that include the notions of risk management and communication in medical practice, and which go beyond the individuals' training regarding the operation of an AI system as decision-support.[107]

New interpretative guidelines need to specify how a healthcare professional aligns normative propositions of medical reasoning with predictive reasoning, to inform the value of information duties in Article 13.[108] There has been some work in this direction done by the EU the Steering Committee for Human rights in the fields of Biomedicine and Health (CDBIO) who are advising on ethical questions and the human right implications connected to evolving technologies including AI in healthcare, as well as the AI HLEG who

[105]Artificial Intelligence Act proposal, art 13 (1).
[106]Xiaoxuan Liu, Ben Glocker, Melissa M McCradden, Marzyeh Ghassemi, Alastair K Denniston, Lauren Oakden-Rayner, 'The medical algorithmic audit' (2022) 4 (5) The Lancet 384, 390.
[107]Cf. Artificial Intelligence Act proposal, Recital 48.
[108]Artificial Intelligence Act proposal, art 13 (2).

wrote a checklist regarding their ethical guidelines on trustworthy AI.[109] Nevertheless, I believe that more multidisciplinary engagement on the education of AI technology in health, both within and outside key stakeholders, specialists including novices, nurses, patients, will drive the conversation on the role of human decision-making *with*, rather than *about* medical diagnostic systems.[110]

Another open question will be regulatory alignment with the New Legislative Framework (NLF) legislation, including sectoral legislation, such as the Medical Device Regulation.[111] Here, the AI Act proposal stipulates that 'while the safety risks specific to AI systems are meant to be covered by the requirements of this proposal, NLF legislation aims at ensuring the overall safety of the final product and therefore may contain specific requirements regarding the safe integration of an AI system into the final product'.[112] Accordingly, specific legislation could effectively support nuanced approaches regarding the verification of medical diagnostic tools.[113] However, what this discussion shows is that the AI Act proposal – including the idea of the fundamental alignment of ex-ante obligations to human-centric values – equally requires nuanced approaches that provide an effective benchmark regarding trustworthy AI, as well as, human-centric regulation concerning AI in healthcare.

The second part of the discussion dealt with the role of algorithms to quantify uncertainty. Manufacturers can use Uncertainty Estimates enabling better human oversight with regard to the system's output. However, a closer reading of Article 14 shows that the role of transparency does not add to the inherent risks of dealing with uncertainty in medical decision-making. Therefore, we need to issue further guidelines that tweak Article 14's wording to turn technical specifications, such as Uncertainty Estimates in medical imaging into an effective safeguard for the doctor-patient relationship.

As a first step, Article 14 needs to shift its focus from a system failure to ambiguity in decision-making. We need to focus on the way a healthcare professional appreciates patient risks, irrespective of the system's risks of failure. By way of illustration, when a healthcare professional oversees the system's reporting of a boundary case between mild and advanced diabetic retinopathy, and the classifier's associated uncertainty, he or she must know what threshold is required to re-visit the influence of inference and reinstate the assumed premises of disease pathogenesis. I suggest the threshold is not subject to the foreseeable risks but arises when the process of inductive inference requires the reconciliation of human-machine expertise.

Finally, I did not mention the role of selective prediction models and a model's use of a "rejector" considering Article 14 of the AI Act proposal.[114] Selective prediction methods are a valuable tool for securing that level of introspection in medical decision-making

[109]Mittelstadt (n 2); High-Level Expert Group on Artificial Intelligence, 'The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment' (n 23).
[110]For instance, see Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, Alastair K Denniston and the SPIRIT-AI and CONSORT-AI Working Group, 'Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension' (2020) 26 (9) Nature Medicine 1364; We would need a similar engagement with regard to the transparency goals of medical diagnostic systems.
[111]Artificial Intelligence Act proposal, [1.2]; Medical Device Regulation.
[112]Artificial Intelligence Act proposal, [1.2].
[113]Examples include, Medical Device Regulation, [23. 4(f) Annex I, Chapter III].
[114]See Section 4.i of the discussion.

when defining the degree of foreseeable harm.[115] The current wording of Article 14 only mentions that 'human oversight measures should be identified and built when technically feasible', whereby those technical safeguards only include the human operator's *active* intervention with the system's operation, such as using a 'stop button or similar procedure'.[116] Conversely, a system's rejector would impose the individual's passive intervention into the AI's decision-making. We need to determine first the level of the healthcare professionals' and patients' tolerance of uncertainty, as a first step before moving to the role of passive decision-making surrounding AI as decision-support. In addition, the individuals' inferential step to define evidence-based solutions based on patient-centred outcomes is a finite step within his or her own appreciation of risk.[117] This inferential step shall never be outsourced to the AI system, and we must make sure that human agency is an aspect that is inherent in clinical decision-making to maintain aspects of patient-centred care.

Acknowledgements

In addition, this work benefitted from my research stay at the Stanford Center for AI Safety. Finally, I presented this work at the SLSA 2022 conference, and I would like to extend my thanks to Professor Brian Simpson, and Dr Mark O'Brien, as well as the participants in the IT Law and Cyberspace stream, for their helpful feedback.

Disclosure statement

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

Funding

This work has been supported by the UKRI Research Node on Governance & Regulation Node within the Trustworthy Autonomous Systems (TAS) programme [grant nnumber EP/V026607/1].

[115]Adam Conner-Simons, 'AI systems that work w/doctors and know when to step in' (*MIT CSAIL*, 31 July 2020) <www.csail.mit.edu/news/ai-systems-work-wdoctors-and-know-when-step> accessed 11 June 2022.

[116]Artificial Intelligence Act proposal, art 14 (4) (e).

[117]Again, some qualitative assessments on the users' and stakeholders' engagement and oversight of AI-based solutions might be useful. In this respect, the work by Elizabeth Bondi, Raphael Koster, Hannah Sheahan et al is interesting, examining whether selective prediction has an impact on the accuracy of human judgements; Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, Krishnamurthy Dvijotham, 'Role of Human-AI Interaction in Selective Prediction' (ArXiv, 16 May 2022) <https://arxiv.org/abs/2112.06751> accessed 11 June 2022.