On the Legal Aspects of Responsible AI: Adaptive Change, Human Oversight, and Societal Outcomes

Daria Onitiu[1], Vahid Yazdanpanah[2], Adriane Chapman[2], Enrico Gerding[2], Stuart E. Middleton[2], Jennifer Williams[2]
[1]Oxford Internet Institute, 1 St Giles', Oxford OX1 3JS, United Kingdom
[2] University of Southampton: University Road, Southampton SO17 1BJ, United Kingdom
daria.onitiu@oii.ox.ac.uk

Abstract. This paper discusses the ways in which complexity and degrees of autonomy in AI-based medical devices (AIaMD) may challenge the safety and performance of software for EU regulatory alignment and responsible AI regarding AI-induced harms. It examines the EU Commission proposals for an AI Liability Directive and a revised Product Liability Directive to identify two research challenges that must be addressed for tracing and assigning legal responsibility of AI-induced harms during the products lifecyle. These challenges relate to identifications of "defects" arising from algorithmic change and degrees of human oversight. Some suggestions will be made in how they can be addressed through causal modelling, counterfactuals, and responsibility reasoning.

Keywords: EU liability, product liability, AI- based medical devices, AI

1 Introduction

Advances in Artificial Intelligence (AI) techniques in healthcare – from intelligent computer vision methods and deep learning in medical imaging to machine learning techniques for personalised medicine – raise important questions regarding who and to what extent could be liable for the consequences of their harms during product life cycle [1, pp. 14–15], [2]. Of particular importance are some special characteristics of AI-based medical devices (AIaMD); from concerns surrounding bias, opacity, and human oversight, to some models being able to "continuously learn" on real-world data [3]. In this article, we briefly introduce how the "AI component" in AIaMD produces some regulatory tension with the strict liability and fault- based liability regime currently proposed in the revised Product Liability Directive (PLD) and the AI Liability Directive (ALD) [4]; [5]. We discuss two research challenges illuminating on a tension of implementing safety and performance requirements within a regulatory framework surrounding responsible AI and that is currently ongoing. Our recommendations concentrate on technical safeguards for traceability and modelling human-agent reasoning intending. These recommendations intend to supplement future policy efforts as well as direct on identifying "design defects" and "non-compliance" with some duties in the "high-risk" obligations in the upcoming Regulation on Artificial Intelligence (AI Act).

EU governance for AI is currently at the cross-roads; with the EU Commission's proposal for an AI Act progressing and entering trilogue negotiations, the ambitious efforts within the EU to engage into technical standard setting for "high-risk systems" and the revision of the product liability framework, as well as the introduction of a new framework on procedural harmonization of fault-based torts [4]–[8]. Within this context, we already see the proliferation of AIaMD, such as the use of AI medical imaging applications for diagnosis [9]. Assurances relating to their performance and safety are inevitably intertwined with the extent that a liability framework can ensure an equilibrium between a device's intended use and unforeseeable consequences.

The proposals for a revised PLD and the ALD intend to illustrate a stepping stone for a more comprehensive AI regulation focusing on the consequences of AI- induced harms [10]. Whilst both frameworks are still at the early stages of development, their inherent link to the AI Act proposal as well as fundamental concepts on the role of software-related harms are an important source for us to issue further comment [11], [12]. We intend to focus on two specific aspects within this legislative net on EU liability; one is the notion of "defect" in the PLD, as well as the "presumption of causality" focusing on the provisions of human oversight in the AILD and referring to the AI Act [4, Art. 6]; [5, Art. 4 (1), 4 (2)].

In this work, we refer to AI and autonomous systems to describe liability issues for special to continuously updated systems. Indeed, we recognise different degrees of "change" in that a medical AI system may either be "locked", allow for "incremental learning" or are "adaptive" [13, p. 2]; [14], [15, p. 30], [16, p. 6]. Focusing on "adaptive systems", what is crucial here is that these decisions are made with little to no direct human intervention [13, p. 2], [17]. Following this narrative, we identify two research challenges which are relevant to the way future regulatory responses will ensure alignment between EU (sectoral) legislation, the AI Act and responsible AI. These are (i) the way a manufacturer may intervene with dynamic learning including adaptive change when AIaMDs exhibit a degree of autonomy for the identification of "defects" in the PLD; and (ii) the role of human-AI interaction in usability and risk management when distributing fault-based claims using the ALD's "presumption of causality" for human oversight.

Moving forward, we make two recommendations in how causal modelling, counterfactuals, and responsibility reasoning, can supplement performance and safety assurance of AIaMD with a view to addressing these research challenges. The paper recommends human-agent reasoning approaches, which are just a subset of approaches for AI assurance that are being actively researched and explored today [18]. Assurance methods can be applied to any part of the AI lifecycle, including data management, model learning, model verification and model deployment. For model verification prior to deployment approaches include probabilistic verification methods, generative adversarial networks, and combinatorial testing. For model assurance of post-deployment behavior monitoring approaches are very important, specifically focusing on model inputs, outputs, and the environment in which the model operates. Techniques include built-in tests, fallback model safe states and well-defined tolerance thresholds within which model outputs must stay.

2 Why contours of EU Liability do not address challenges pertaining to AIaMD

Imagine an AIaMD which employs deep learning algorithms for the interpretation of Chest X-rays. This device is “intended to be used” for the purpose of assistive diagnosis of pneumonia [19, Art. 2]. Developers validating the device based on a significant number of labelled medical images for training and evaluation of the algorithm were able to show the system’s impressive accuracy to distinguish between features in an image, which would outperform the judgement of a radiologist [20]. Nevertheless, this system when deployed into a healthcare environment could have serious implications for patient safety: from possible risks based on the opacity of the model leading up to “user errors” [21, Sec. 5.2], degrading in performance once exposed to real-world data, to learning from spurious correlations in the training data [23]; [24]. Especially relevant to AIaMD is that such risks could result in adverse outcomes that affect patient safety, including misdiagnosis and consequently mistreatment of the patient.

These problems on the EU regulation of AIaMD for patient safety, touching upon the Medical Device Regulation, incorporate important liability questions. By way of illustration, if we assume that the system incorporates a degree of “autonomy”, such as adaptive learning from real-world data in real-time and without explicit user input, then these characteristics pose a regulatory challenge for the developer to ensure continuous safety and functionality, to monitor changes that are a “significant change” to the device’s intended use, as well as to maintain the device’s benefit-risk profile [15], [16], [24]. Questions arise to what extent developers can be liable for any possible “defects” arising from “differential performance”; i.e., deviations of the algorithm from (pre-defined) parameters [1, p. 5], [13, p. 4], [25, p. 28];. The question arises: to what extent does the manufacturer including developers maintain control for these specific type of adaptive changes causing harm including “death or personal injury, including medically recognised harm to “medically recognised” psychological health”? [4, Art. 4 (6)], [26], [27, N. Amendment 5 Recital 17].

In addition, if we assume that the AIaMD assists healthcare professionals regarding the diagnosis and treatment of different stages of pneumonia, then questions for minimizing use errors and ensuring risk management on the one hand and providing for human oversight on the other hand, are important elements producing liability gaps. For example, if a human agent overrides the output of a fully and opaque autonomous system, does that provide for the liability of the user of the “high-risk system”?

The latest proposals for two EU Liability Directives – the PLD and ALD- seek to resolve some tensions arising from the challenge to allocate liability regarding software-related failures, as well as tensions arising from burden of proof for fault-based claims. The revised PLD includes in its material scope medical device software and AI systems [26], [28, p. 826]; [29, Art. Recital 12]; see also, [30]. Another objective of the proposed EU Liability framework is to incorporate the special characteristics of AI-algorithmic complexity, autonomy, and opacity- which would make it substantially difficult for victims to succeed in a liability claim based on a “lack of compliance under Union or national law” ([5, Art. 3 (1); Explanatory memorandum]. In this respect, The ALD, extending on the provisions in about "high-risk" systems in the AI Act, allows the victim to claim damages regarding harm caused by medical AI systems to safety and/or fundamental rights.[31, p. 4].

The extent to which these two legislative frameworks successfully close "liability gaps" still leaves some questions open [10], [13], [32]–[36]. Focusing on our example of AIaMD used for the detection of pneumonia, it could be argued that a "defect" assumes that there is a well-defined standard from which errors in design and performance can be judged [10, p. 24]. But such well-defined standards are not currently under the remit of AI regulation, and instead are currently defined based on academic research or in-house by manufacturers within their own product safety departments, if at all [37]. Similarly, if a user (over)-relies on the AI system's opaque output then it is unclear to what extent the ALD would provide any victims with redress under this framework, including the ability to know if there have been any victims.

Our analysis proceeds with the identification of two research challenges on the safety and performance of AIaMD stimulating future discussion on a comprehensive the EU Liability regime. First, we argue that making the developer strictly liable for software-related failures and hazardous situations arising from the use of AIaMD after deployment, needs to be accompanied with further specifications for developers to effectively intervene with adaptive learning. Second, we argue that the ALD currently provides the wrong incentives to "prove fault" only on the basis of incorrect specifications for human oversight, leaving out an insight into the elements for "providers" to justify the AIaMD's benefit-risk profile and "users" to follow the necessary competence for intervention [5, Art. 4 (2), 4 (3)].

2.1. Research Challenge 1: Lack of specifications for developers justifying an intervention with adaptive change

Our first statement of why the proposed EU Liability framework does not fully capture risks surrounding design and deployment of AIaMD is based on the lack of guidance for directing contours of continuous and adaptive change when the system exhibits a degree of autonomy. As a starting point, enabling change through updates is an important aspect of "software maintenance" being a necessary component to ensure patient safety ([38, Sec. 6]; see also, [4, Art. Recital 38], [33]). Many currently approved AIaMDs employ models that are "locked", providing the same "output on a given set of inputs"[17, p. 3], [39, p. 3]. "Batch-learning" refers to instances where the algorithm does retrain itself incrementally after seeing a new batch of training data [40, p. 30]. In these given instances we can argue that the manufacturer retains some form of control regarding the extent software updates are pushed through the lifecyle. This reasoning of control extends to "defects" arising from the lack or improper controls including software updates altering performance, safety and functionality after deployment [4, Art. 10 (2) (b)-(c)], [10, p. 46].

However, there is a lot of potential in algorithmic models that are programmed to internally modify their algorithms for a new output based on real-world data [16, p. 8]. These models are adaptive in that these algorithms continuously learn and change their performance [41]. Whilst these models could be useful to provide more "timely recommendations" based on real-world experience [42, p. 678], these characteristics could pose additional issues for validation and oversight [16, p. 71].

One issue is that the adaptive nature of algorithms poses a challenge for manufacturers to recognise "significant changes" and/or performance degradation for patient safety [43], [44, pp. 32–33]). This in turn can have significant implications for the extent to which "unpredictable" changes would be classified as a "defect" under the revised PLD leading to "material damage" [13, p. 4], [25, p. 28];. These "defects" as a basis to trigger strict liability are based on the failure to disclose relevant evidence about the product or comply with mandatory safety legislation, such as the Medical Device Regulation, or by virtue of an "obvious malfunction" [4, Art. 9 (2) (a)-(c); 8 (1); 9 (3)].

As highlighted by Borges [32, p. 4], once the manufacturer's control effectively relinquishes with regard to the operation of the system, then "defectiveness" has to be inferred from the system's behaviour [32, pp. 4–5]. To establish liability then would "require defining property as the ability to behave in a certain way in a certain situation or not to show a certain behaviour" [4, Art. Recital 37], [45, p. 35]. In this respect, Article 6 (1) (c) of the revised PLD establishes that a product could be "considered defective when it does not provide the safety which the public at large is entitled to expect, taking all circumstances into account, including…the *effect* on the product of other products that can reasonably be expected to be used together with the product" (emphasis added, [4, p. 6 (1) (c)]).

A specific hurdle for regulators to effectively adapt this notion of "defectiveness" for AIaMD concerns those exact specifications tracing adaptive change. The evaluation of AIaMD needs to be subject to pre-defined parameters whilst the variability of risk occurs *within* those dynamic changes. This is effectively a problem that lies at the heart of the verification and validation of AIaMD; once the manufacturer evaluates and confirms the medical device's specific and intended use, the next step is monitoring the extent that the system is operating within "a set of underlying assumptions" [46, p. 442].

Hence, an important challenge in a comprehensive legal responsibility framework of AI is determining the manufacturer's formalisation of the system's safety and reliability when interacting with various stakeholders on the ground, such as users, healthcare professionals, and patients. Indeed, regulatory developments are currently dealing with the problems associated with the regulation of adaptive algorithms in medicine. This can be seen in guidance by the U.S Food & Drug Administration [47], current efforts by the UK Medicines & Healthcare products Regulatory Agency (MHRA) [48], as well as a draft reflection piece by the European Medicines Agency (EMA) [49, Sec. 2.4.6]. In this regard, the EMA thought-piece enumerates "thresholds for model performance" that are needed for manufacturers to monitor degradation and failure modes of algorithms [49, Sec. 2.4.6].

Moreover, the presidency draft report makes an important first step clarifying the economic operator's, such as the manufacturer's responsibilities with regard to AIaMD entailing adaptive algorithms [4, Art. Recital 37]; [49, Art. Recital 29]. Following this thought process, a manufacturer intending to deploy an adaptive AIaMD causing "unexpected behaviour" still retains that level of control for damages that arise after deployment [49, Art. Recital 23], see also, [50, Art. Recital 23].

What is missing; however, is placing these considerations into a wider context on how manufacturers justified any residual risks before and after deployment in the event of design defects. This is because the notion of "defect" in the revised PLD assumes the manufacturer's level of control regarding *anticipated* modifications on the one hand, and "substantial modifications" based on the Medical Device Regulation or considering Article 4 (10a) in the presidency compromise text on the other [29, Art. 4 (10a) (a)-(b)], [50, Art. 4 (10a) (a)-(b)]. The question, therefore, is which level of modifications justify deviations within a system's acceptable risk policy after deployment and recognizing an extent of diminishing control. What follows is that, in addition to the extent that a system re-trains dynamically, we need further specifications in *how* manufacturers can justify an intervention with underlying assumptions surrounding adaptive change. That is, most specifications that are "pre-determined" by the manufacturer from the outset are likely to be either inconsistent or incomplete with adaptive algorithms [41, p. 1203]. In addition, if we assume that the manufacturer producing a risk management plan with "model performance thresholds" for adaptive systems, this has to be based on clear criteria for manufacturers to intervene with a pre-determined policy. Therefore, what matters is for manufacturers to be able to justify any residual risks during the product life cycle and base these on the way risk control and mitigation maintain the manufacturer's control.

2.1. a. The role of causal reasoning for dynamic and adaptive change

A promising way forward for addressing challenges around intervening with change for continuously learning AIaMD is to leverage the potentials of computational causal models and reasoning. Causal modelling and reasoning methods, such as those developed by Judea Pearl [51] and Joseph Halpern [52], can provide useful frameworks for modelling and anticipating the impacts of algorithmic changes to AIaMD. These methods allow developers to map out the potential chains of cause-effect relationships stemming from modifications to the AI system's algorithm. By forecasting how various causal factors may lead to harmful outcomes, developers can take steps to avoid or mitigate these risks proactively. Causal models that accurately represent the dynamics of the AI system, its real-world deployment context, and interactions with human users can enable explanatory insights about how and why unintended consequences may arise. Specifically, formal causal reasoning can supplement standard verification and validation protocols by identifying probable failure points or risks. In this way, causal modelling supports the design of safer AI systems and responsibility frameworks that account for complex sociotechnical interactions. Integrating these techniques can strengthen technical specifications for dynamic changes and manufacturer oversight of medical AI.

Furthermore, causal models can assist stakeholders in determining when and how to appropriately intervene whilst tracking risks of performance related harms. By modelling the dynamics of the AI system, developers can use causal models to identify specific adaptive changes and significant changes in continuously learning algorithms. This enables them to focus interventions on the most impactful areas, whether through altering technical specifications, adjusting training data, or implementing human oversight mechanisms. Causal reasoning allows for targeted interventions that avoid unnecessary restrictions on beneficial adaptations of the AI, while still providing safeguards against harmful impacts on health, safety. In this way, causal models support nuanced and selective interventions on high-risk changes, guided by causal understanding of how different modifications contribute to various intended and unintended consequences.

2.2. Research Challenge 2: Identifying the parameters for risk management and usability of AIaMD

The second research challenge, focusing on the interpretation of the ALD, is aligned with the AI Act's parameters for human oversight regarding AI used for decision-support. An important aspect of the proposal for an ALD is that it establishes a framework that "harmonise[s] non-contractual civil liability rules" [31, p. 5], [35]. In doing so, it ensures that the victim, such as the individual patient harmed by the output of the AIaMD detecting pneumonia, to have recourse to the developer, the provider or user's "fault" during the use of the AIaMD and that caused "damage" under national rules [31, p. 5], [35].

Article 4 entails a "presumption of causality" [5, Art. 4], [31, p. 6]. For example, if the provider (i.e., manufacturer or a person placing the product on the market) or "user" (i.e., a person using the AI system under the provider's authority) do not comply with the relevant provision of "human oversight" in the AI Act, then Article 4 implies that fault can be inferred from the fault on the provider or user based on non-compliance of provision relating to "high-risk" systems under the AI Act ([5, Art. 4 (2), 4 (3)], however, compare with [5, Art. Recital 25]). Further, it could be argued that the defendant's non-compliance with ensuring "human oversight" likely influenced the "output of the AI system" [5, Art. 4 (1) (b)], by providing an outcome that does not allow the user to recognise the limitations of the system [5, Art. 4 (1) (c); Recital 25]. Returning to the earlier example of AI-based pneumonia detection, this scenario would be exemplified by a qualified medical professional who received the relevant training for instructions of use whilst using the tool in a manner that later causes the AI algorithm to adaptively learn from incorrect or incomplete inputs during real-time learning updates. Other examples that attract a presumption of non-compliance under the AI Act relate to data quality requirements, as well as the transparency requirements based on the system's intended use [5, Art. 4 (2)], [53, p. 5]

Additionally, the ALD consists of another presumption which concerns the disclosure of evidence regarding a "high-risk" system based on a court order [5, Art. 3]. Article 3 (5) highlights that the defendant's failure to comply with a court order will lead to a presumption of non-compliance of duty of care under Union or national, which includes the example of non-compliance with "human oversight" in the AI Act above [5, Art. 3 (5)], [10, p. 34].

The defendant may rebut the presumption in Article 4 (1) "by showing that its fault could not have caused the damage" [5, Art. Recital 20, 3 (5), 4 (1), 4 (7)]. In addition, Article 4 (4) argues that the presumption does not apply in instances where the defendant successfully argues that "sufficient evidence and expertise is accessible for the claimant to prove the causal link" [5, p. 4 (4)]. Furthermore, the fact that the presumption is rebuttable highlights that it must have been “reasonably likely” on a case-by-case basis that the fault occurred based on non-compliance with a duty that had an impact on the system’s output [5, Art. 4 (1) (b)].

Non-compliance with the duty of human oversight opens an interesting discussion on how liability is distributed between the provider and user of a “high-risk” system including AIaMD. This is because human oversight is an aspect of usability for risk management as well as following instructions of use. To effectively implement this requirement for AIaMD requires more specifications of the limits of the system’s intended use [54, p. 171]. A consultation by the BSI and AAMI further highlight that device complexity and autonomy require human-in-the-loop testing as well as evaluations in how user interface design can introduce “automation bias” [55, Sec. 5.9.1-5.9.5]. Having said that; however, the lack of specifications for human oversight of AIaMD on the EU level currently leaves manufacturers considerable leeway to justify the limitations of a “high-risk” system interacting with stakeholders on the ground[54, p. 180].

The ALD currently precludes a multileveled assessment of how risk management and usability informs claims against the provider and user. Article 4 (2) of ALD whilst highlighting the “results of the risk management system” as a factor for constructing the “rebuttable presumption of causality” poses issues for interpreting usability at a level of complexity of the intended environment and user of AIaMD. This reasoning is based on its construction of “fault” as a “failure of the AI system to produce an output” which focus in deviations of specifications based on inadequate risk controls that lead to an erroneous assessment of the “overall residual risk” [55, p. 8.1]. A clear example of an incorrect specification is for developers to not include “warnings” or “specific training” for users when these elements would be a necessary component for using a AIaMD for assistive diagnosis and maintaining the benefit-risk profile [56, Sec. 7].

In other instances, however, it is less clear how the effects of specifications, influencing over-reliance and possibly producing “automation bias” would trigger the “presumption of causality” for aspects relating to the safety and performance of AIaMD under the AI Act and the MDR. This would require the identification of inconsistencies on the evaluation of usability and risk management, such as for the victim showing an imbalance in the overall benefit-risk ratio. A claim against the provider of the AIaMD for non-compliance with human oversight is currently narrowed to the developer’s specifications of risk controls for usability. Article 4 (2) only allows for proving fault on the basis of incorrect specifications; not how limitations of specifications of usability caused *wrong incentives*, when users are interacting with the system. This is because Article 4 (2) elaborates on the connection of fault without necessarily tapping into the manufacturer’s justification of the overall residual risk.

Moreover, Hacker makes an important point capturing the limits of Article 4 (3) ALD concerning claims against the user in that it presumed a strong link “between the fault and the output… [not the extent of human oversight] after the AI output is produced” [10, p. 36]. This is further seen in Recital 25 of the ALD illuminates how “fault” is usually tied to the boundary specifications such as the breach of “the perimeter of operation of the AI system” [5, Sec. Recital 25]. Additionally, Recital 15 highlights that claims against the user for human omissions implementing the AI output means that responsibility for the damage needs to be traced back to the user’s actions rather than the AI system’s output[5, Art. 15].

Without guidance on the appropriate level of interaction between the human and AI, a victim will have great difficulty to show the consistency between usability specifications on the intended user and the oversight shaped by the AI output (see also [36, p. 4]). In particular, it opens up a tension for the interpretation of Articles 4 (2) and 4 (3) that is limited to the user simply “implementing the AI system’s output” and the risks of overreliance detached from the safety and effectiveness of a “high-risk system”.

2.1.a. Counterfactual scenarios illuminating on degrees of human-AI interaction

Modelling counterfactual scenarios can support stakeholders in reasoning about the avoidance potential of different agents and whether they could have prevented outcomes producing a clear link between fault and non-compliance that caused harm [57], [58]. By examining alternate possibilities along the timeline, not just future risks, these models elucidate the range of actions manufacturers, providers, and users could have taken at each decision point. Or it could simulate how increased human monitoring and overrides at the deployment stage could have prevented improper implementation of the system’s output. Evaluating these counterfactuals helps determine if and where human oversight failed or could be enhanced. This analysis enables clearer delineation of responsibility by revealing who had the knowledge and capacity to avoid harms at various stages [59].

Building on counterfactual models, computational techniques for responsibility assignment also hold promise [60]–[62]. These approaches formally analyse the responsibilities of human and AI agents given counterfactual trajectories. The modelling considers agents' available actions, knowledge of likely outcomes. By simulating adherence to these norms, the model can assess and oversight. Encoding factors like safety rules, computational models can verify whether following the requirements could have changed outcomes. Verifying responsibility through this computational approach accounts for complex sociotechnical dynamics between manufacturers, health providers, and end users.

Importantly, formal computational models allow distinguishing graded levels of responsibility, from causal contribution to harm [63]. Whereas causal models capture roles in producing an outcome, computational responsibility modelling also verifies awareness of avoidability. For instance, these techniques could determine if a harmful AI outcome was due to reasonably foreseeable misuse. This enables moving to nuanced designations from noncompliance to full culpability. In this way, computational techniques could strengthen assessment of human oversight and clarify ambiguity around legal liability for harms involving AI systems.

3. Concluding thoughts

This paper illustrates a snapshot on some problems shaping EU regulators’ effort for responsible AI and governance. It provides an understanding of the regulation of

AIaMD which is based on their special characteristics and its specific benefit-risk profile to produce AI-induced harms. Whilst our discussion of the proposed EU Liability framework is not comprehensive, it picks up two important research challenges that evolve around the ongoing challenges for regulatory reform on the safety and performance of AIaMD. These are (1) problems of monitoring dynamic and adaptive change including continuously learning algorithms and (2) issues to specify human oversight of AIaMD used for decision-support.

An important limitation of our work is that our comments are clearly focused on the EU regulatory landscape and currently ongoing efforts by EU regulators. In this regard, future standard-setting could evolve horizontally within the AI Act and following an approach that is "assurance-based", as well as entail sector-specific standards for AIaMD whilst considering the Medical Device Regulation [8], [24, p. 43], [64, p. 65]. Our findings, whilst not providing an exhaustive picture to the design and use of AIaMD, indicates that these issues for responsible AI are cross-sectoral, being relevant for the interpretation of "defects", the notion of procedural "fault" and finally, regulatory alignment within a legislative climate that grapples with providing technical standard-setting, legal certainty and keeping up with technological developments.

Acknowledgements: This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the Trustworthy Autonomous Systems Hub (EP/V00784X/1), a Turing AI Acceleration Fellowship on Citizen-Centric AI Systems (EP/V022067/1) and the UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation (EP/V026607/1). In addition, this work was supported by the Natural Environment Research Council (NE/S015604/1) and Economic and Social Research Council (ES/V011278/1), as well as the National Institute for Health and Care Research Southampton Biomedical Research Center (IS-BRC-1215-20004). Finally, this work was supported through research funding by the Welcome Trust (223765/Z/21/Z) and Solan Foundation (G-2021-16779) within the Trustworthiness Auditing for AI project at the Oxford Internet Institute.

References

[1] Regulatory Horizons Council (RHC), 'RHC report on the regulation of Artificial Intelligence as a Medical Device', Nov. 2022. Accessed: Apr. 20, 2023. [Online]. Available: https://www.gov.uk/government/publications/regulatory-horizons-council-the-regulation-of-artificial-intelligence-as-a-medical-device.

[2] Paula Margolis, Samantha Silver, and Jenny Yu, 'Artificial Intelligence in Life Sciences: Regulating AI Technologies and the Product Liability Implications'. Accessed: Aug. 03, 2023. [Online]. Available: https://www.techuk.org/resource/artificial-intelligence-in-life-sciences-regulating-ai-technologies-and-the-product-liability-implications.html.

[3] IMDRF AIMD Working Group, ‘Machine Learning-enabled Medical Devices-A subset of Artificial Intelligence-enabled Medical Devices: Key Terms and Definitions’. Sep. 16, 2021. [Online]. Available: https://www.imdrf.org/documents/machine-learning-enabled-medical-devices-key-terms-and-definitions.

[4] *Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on liability for defective products*. 2022. Accessed: Jul. 31, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0495.

[5] *Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)*. 2022. Accessed: Jul. 30, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0496.

[6] *DRAFT Compromise Amendments on the Draft Report Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 –C9 0146/2021 –2021/0106(COD))*. 2023. Accessed: Oct. 16, 2023. [Online]. Available: chrome-extension://efaihttps://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf.

[7] Julia Tar and Luca Bertuzzi, ‘AI Act’s trilogue preparation, the EU submarine cable agenda’, www.euractiv.com. Accessed: Aug. 04, 2023. [Online]. Available: https://www.euractiv.com/section/digital/news/ai-acts-trilogue-preparation-the-eu-submarine-cable-agenda/.

[8] EU Commission, ‘Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence’. Accessed: Jul. 28, 2023. [Online]. Available: https://ec.europa.eu/docsroom/documents/52376.

[9] Will Douglas Heaven, ‘Google’s medical AI was super accurate in a lab. Real life was a different story. | MIT Technology Review’, MIT Technology Review. Accessed: Aug. 04, 2023. [Online]. Available: https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/.

[10] P. Hacker, ‘The European AI Liability Directives -- Critique of a Half-Hearted Approach and Lessons for the Future’. arXiv, Jul. 28, 2023. doi: 10.48550/arXiv.2211.13960.

[11] Luca Bertuzzi, ‘EU Council clarifies liability rules for software updates, machine learning’, www.euractiv.com. Accessed: Aug. 03, 2023. [Online]. Available: https://www.euractiv.com/section/digital/news/eu-council-clarifies-liability-rules-for-software-updates-machine-learning/.

[12] Luca Bertuzzi, ‘Has software industry missed the train on EU’s new liability rules?’, www.euractiv.com. Accessed: Aug. 03, 2023. [Online]. Available: https://www.euractiv.com/section/digital/news/has-software-industry-missed-the-train-on-eus-new-liability-rules/.

[13] M. N. Duffourc and S. Gerke, ‘The proposed EU Directives for AI liability leave worrying gaps likely to impact medical AI’, *npj Digit. Med.*, vol. 6, no. 1, Art. no. 1, Apr. 2023, doi: 10.1038/s41746-023-00823-w.

[14] A. Gepperth and B. Hammer, ‘Incremental learning algorithms and applications’, in *European Symposium on Artificial Neural Networks (ESANN)*, Bruges, Belgium, 2016. Accessed: Aug. 21, 2023. [Online]. Available: https://hal.science/hal-01418129

[15] Johan Ordish, Hannah Murfet, and Alison Hall, ‘Algorithms as medical devices’. Accessed: Apr. 16, 2023. [Online]. Available: https://www.phgfoundation.org/report/algorithms-as-medical-devices.

[16] BSI and AAMI, ‘MACHINE LEARNING AI IN MEDICAL DEVICE: Adapting Regulatory Frameworks and Standards to Ensure Safty and Performance’. Accessed: Apr. 21, 2023. [Online]. Available: https://www.bsigroup.com/en-US/medical-devices/resources/Whitepapers-and-articles/machine-learning-ai-in-medical-devices/.

[17] S. Gerke, B. Babic, T. Evgeniou, and I. G. Cohen, ‘The need for a system view to regulate artificial intelligence/machine learning-based software as medical device’, *npj Digit. Med.*, vol. 3, no. 1, Art. no. 1, Apr. 2020, doi: 10.1038/s41746-020-0262-2.

[18] R. Ashmore, R. Calinescu, and C. Paterson, ‘Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges’, *ACM Comput. Surv.*, vol. 54, no. 5, p. 111:1-111:39, May 2021, doi: 10.1145/3453444.

[19] *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 (Text with EEA relevance.)*, vol. 117. 2017. Accessed: Jul. 30, 2023. [Online]. Available: http://data.europa.eu/eli/reg/2017/745/oj/eng.

[20] E. J. Topol, ‘High-performance medicine: the convergence of human and artificial intelligence’, *Nat Med*, vol. 25, no. 1, Art. no. 1, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[21] ‘BS EN 62366-1:2015+A1:2020 Medical devices. Application of usability engineering to medical devices’. Accessed: Apr. 27, 2023. [Online]. Available: https://knowledge.bsigroup.com/products/medical-devices-application-of-usability-engineering-to-medical-devices-1?version=standard.

[22] ‘A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy | Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems’. Accessed: Apr. 22, 2023. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3313831.3376718.

[23] J. W. Gichoya *et al.*, ‘AI recognition of patient race in medical imaging: a modelling study’, *The Lancet Digital Health*, vol. 4, no. 6, pp. e406–e414, Jun. 2022, doi: 10.1016/S2589-7500(22)00063-2.

[24] the European Coordination Committee of the Radiological, Electromedical and Healthcare IT Industry (COCIR), ‘ARTIFICIAL INTELLIGENCE IN EU MEDICAL DEVICE LEGISLATION’, May 2021. Accessed: Apr. 20, 2023.

[Online]. Available: https://www.cocir.org/media-centre/publications/article/cocir-analysis-on-ai-in-medical-device-legislation-september-2020.html.

[25] Directorate-General for Justice and Consumers (European Commission), *Liability for artificial intelligence and other emerging digital technologies*. LU: Publications Office of the European Union, 2019. Accessed: Aug. 01, 2023. [Online]. Available: https://data.europa.eu/doi/10.2838/573689.

[26] 'Q&As on the revision of the Product Liability Directive', European Commission - European Commission. Accessed: Aug. 01, 2023. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/qanda_22_5791.

[27] *Draft report on the proposal for a directive of the European Parliament and of the Council on Liability for defective products (COM(2022)0495 – C9-0322/2022 – 2022/0302(COD))*. 2023. [Online]. Available: https://www.europarl.europa.eu/doceo/document/CJ24-PR-745537_EN.pdf.

[28] T. de Graaf and G. Veldt, 'The AI Act and Its Impact on Product Safety, Contracts and Liability', *European Review of Private Law*, vol. 30, no. 5, Oct. 2022, Accessed: Aug. 01, 2023. [Online]. Available: https://kluwerlawonline-com.ezproxy-prd.bodleian.ox.ac.uk/api/Product/CitationPDFURL?file=Journals\ERPL\ERPL2022038.pdf.

[29] *Mandate for negotiations with the European Parliament - Proposal for a Directive of the European Parliament and of the Council on liability for defective products*. 2023. Accessed: Oct. 16, 2023. [Online]. Available: https://data.consilium.europa.eu/doc/document/ST-10694-2023-INIT/en/pdf.

[30] Luca Bertuzzi, 'European Parliament tries to accelerate on product liability rulebook', *EURACTIV*, Jul. 03, 2023. Accessed: Oct. 16, 2023. [Online]. Available: https://www.euractiv.com/section/digital/news/european-parliament-tries-to-accelerate-on-product-liability-rulebook/.

[31] Tambiama Madiega and European Parliamentary Research Service, 'Artificial intelligence liability directive', PE 739.342, Feb. 2023.

[32] G. Borges, 'Liability for AI Systems Under Current and Future Law: An overview of the key changes envisioned by the proposal of an EU-directive on liability for AI', *Computer Law Review International*, vol. 24, no. 1, pp. 1–8, Feb. 2023, doi: 10.9785/cri-2023-240102.

[33] T. S. Cabral, 'Liability and artificial intelligence in the EU: Assessing the adequacy of the current Product Liability Directive', *Maastricht Journal of European and Comparative Law*, vol. 27, no. 5, pp. 615–635, Oct. 2020, doi: 10.1177/1023263X20948689.

[34] D. Schneeberger, K. Stöger, and A. Holzinger, 'The European Legal Framework for Medical AI', in *Machine Learning and Knowledge Extraction*, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 209–226. doi: 10.1007/978-3-030-57321-8_12.

[35] Samar Abbas Nawaz, 'The Proposed EU AI Liability Rules: Ease or Burden?', European Law Blog. Accessed: Aug. 01, 2023. [Online]. Available: https://europeanlawblog.eu/2022/11/07/the-proposed-eu-ai-liability-rules-ease-or-burden/

[36] Future of Life Institute, ‘FLI Position Paper on AI Liability - FLI position on the proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)’, Nov. 2022. [Online]. Available: https://futureoflife.org/wp-content/uploads/2022/11/FLI_AI_Liability_Position_Paper.pdf.

[37] J. Williams, K. Pizzi, S. Das, and P.-G. Noe, ‘New Challenges for Content Privacy in Speech and Audio’, in *2nd Symposium on Security and Privacy in Speech Communication*, Sep. 2022, pp. 1–6. doi: 10.21437/SPSC.2022-1.

[38] ‘BS EN 62304:2006+A1:2015 Medical device software. Software life-cycle processes’. Nov. 30, 2006. Accessed: May 01, 2023. [Online]. Available: https://knowledge.bsigroup.com/products/medical-device-software-software-life-cycle-processes/standard.

[39] U.S Food & Drug Administration, ‘Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback’, FDA, Apr. 2019. Accessed: Aug. 01, 2023. [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.

[40] Johan Ordish, ‘Large Language Models and software as a medical device - MedRegs’. Accessed: Jul. 27, 2023. [Online]. Available: https://medregs.blog.gov.uk/2023/03/03/large-language-models-and-software-as-a-medical-device/.

[41] B. Babic, S. Gerke, T. Evgeniou, and I. G. Cohen, ‘Algorithms on regulatory lockdown in medicine’, *Science*, vol. 366, no. 6470, pp. 1202–1204, Dec. 2019, doi: 10.1126/science.aay9547.

[42] C. Petersen *et al.*, ‘Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper’, *J Am Med Inform Assoc*, vol. 28, no. 4, pp. 677–684, Jan. 2021, doi: 10.1093/jamia/ocaa319.

[43] UK Medicines & Healthcare products Regulatory Agency (MHRA), ‘Consultation outcome: Chapter 10- Software as a Medical Device’. [Online]. Available: www.gov.uk/government/consultations/consultation-on-the-future-regulation-of-medical-devices-in-the-united-kingdom/outcome/chapter-10-software-as-a-medical-device#section-58---scope-and-definition.

[44] Karim Lekadir, Gianluca Quaglio, Anna Tselioudis, and Catherine Gallin, ‘Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts | Digital Skills & Jobs Platform’. Accessed: Apr. 20, 2023. [Online]. Available: https://digital-skills-jobs.europa.eu/en/inspiration/research/artificial-intelligence-healthcare-applications-risks-and-ethical-and-societal.

[45] G. Borges, ‘AI systems and product liability’, in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, in ICAIL ’21. New York, NY, USA: Association for Computing Machinery, Jul. 2021, pp. 32–39. doi: 10.1145/3462757.3466099.

[46] D. P. M. Doorn Neelke, Ed., ‘Francien Dechesne and Tijn Borghuis On Verification and Validation in Engineering’, in *The Routledge Handbook of the*

Philosophy of Engineering, New York: Routledge, 2020. doi: 10.4324/9781315276502.

[47] U.S Food & Drug Administration, ‘Artificial Intelligence and Machine Learning in Software as a Medical Device’, FDA. Accessed: Apr. 30, 2023. [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.

[48] Medicines & Healthcare products Regulatory Agency, ‘Software and Artificial Intelligence (AI) as a Medical Device’, GOV.UK. Accessed: Aug. 21, 2023. [Online]. Available: https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device.

[49] EMA, ‘Reflection paper on the use of artificial intelligence in lifecycle medicines’, European Medicines Agency. Accessed: Aug. 21, 2023. [Online]. Available: https://www.ema.europa.eu/en/news/reflection-paper-use-artificial-intelligence-lifecycle-medicines.

[50] *Presidency draft compromise proposal - Proposal for a Directive of the European Parliament and of the Council on liability for defective products*. 2023. [Online]. Available: https://data.consilium.europa.eu/doc/document/ST-7255-2023-REV-1/en/pdf.

[51] Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[52] Joseph Y Halpern, *Actual Causality*. MIT Press, 2019. Accessed: Aug. 01, 2023. [Online]. Available: https://mitpress.mit.edu/9780262537131/actual-causality/

[53] M. Ziosi, J. Mökander, C. Novelli, F. Casolari, M. Taddeo, and L. Floridi, ‘The EU AI Liability Directive: shifting the burden from proof to evidence’. Rochester, NY, Jun. 06, 2023. Accessed: Jun. 09, 2023. [Online]. Available: https://papers.ssrn.com/abstract=4470725.

[54] D. Onitiu, ‘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*, vol. 32, no. 2, pp. 170–188, May 2023, doi: 10.1080/13600834.2022.2116354.

[55] UK BSI and AAMI, ‘21/30428107 DC BS 34971/AAMI CR 34971. Guidance on the Application of ISO 14971 to Artificial Intelligence and Machine Learning’. Apr. 2022.

[56] ‘BS EN ISO 14971:2019+A11:2021 Medical devices. Application of risk management to medical devices’.

[57] T. Wei, F. Feng, J. Chen, Z. Wu, J. Yi, and X. He, ‘Model-Agnostic Counterfactual Reasoning for Eliminating Popularity Bias in Recommender System’, in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Aug. 2021, pp. 1791–1800. doi: 10.1145/3447548.3467289.

[58] S. R. Pfohl, T. Duan, D. Y. Ding, and N. H. Shah, ‘Counterfactual Reasoning for Fair Clinical Risk Prediction’, in *Proceedings of the 4th Machine Learning for Healthcare Conference*, PMLR, Oct. 2019, pp. 325–358. Accessed: Aug. 01, 2023. [Online]. Available: https://proceedings.mlr.press/v106/pfohl19a.html.

[59] M. Braham and M. van Hees, ‘An Anatomy of Moral Responsibility’, *Mind*, vol. 121, no. 483, pp. 601–634, Jul. 2012, doi: 10.1093/mind/fzs081.

[60] V. Yazdanpanah and M. Dastani, ‘Quantified Degrees of Group Responsibility’, in *Coordination, Organizations, Institutions, and Norms in Agent Systems XI*, V. Dignum, P. Noriega, M. Sensoy, and J. S. Sichman, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 418–436. doi: 10.1007/978-3-319-42691-4_23.

[61] M. Dastani and V. Yazdanpanah, ‘Responsibility of AI Systems’, *AI & Soc*, vol. 38, no. 2, pp. 843–852, Apr. 2023, doi: 10.1007/s00146-022-01481-4.

[62] V. Dignum, ‘Responsibility and Artificial Intelligence’, in *The Oxford Handbook of Ethics of AI*, M. D. Dubber, F. Pasquale, and S. Das, Eds., Oxford University Press, 2020, p. 0. doi: 10.1093/oxfordhb/9780190067397.013.12.

[63] V. Yazdanpanah *et al.*, ‘Different Forms of Responsibility in Multiagent Systems: Sociotechnical Characteristics and Requirements’, *IEEE Internet Computing*, vol. 25, no. 6, pp. 15–22, Nov. 2021, doi: 10.1109/MIC.2021.3107334.

[64] Stuart E Middleton, Emmanuel Letouzé, Ali Hossaini, and Adriane Chapman, ‘Trust, Regulation, and Human-in-the-Loop AI: within the European region’, *Communications of the ACM*, vol. 65, no. 4, pp. 64–68, 2022.