Responsible Agency Through Answerability: Cultivating the Moral Ecology of Trustworthy Autonomous Systems

Responsible Agency Through Answerability

Cultivating the Moral Ecology of Trustworthy Autonomous Systems

Louise Hatherall*
The University of Edinburgh
lhathera@ed.ac.uk

Dilara Keküllüoğlu
The University of Edinburgh
d.kekulluoglu@ed.ac.uk

Nadin Kokciyan
The University of Edinburgh
nadin.kokciyan@ed.ac.uk

Michael Rovatsos
The University of Edinburgh
michael.rovatsos@ed.ac.uk

Nayha Sethi
The University of Edinburgh
nayha.sethi@ed.ac.uk

Tillmann Vierkant
The University of Edinburgh
t.vierkant@ed.ac.uk

Shannon Vallor
The University of Edinburgh
svallor@ed.ac.uk

ABSTRACT

The decades-old debate over so-called 'responsibility gaps' in intelligent systems has recently been reinvigorated by rapid advances in machine learning techniques that are delivering many of the capabilities of machine autonomy that Matthias [1] originally anticipated. The emerging capabilities of intelligent learning systems highlight and exacerbate existing challenges with meaningful human control of, and accountability for, the actions and effects of such systems. The related challenge of human 'answerability' for system actions and harms has come into focus in recent literature on responsibility gaps [2, 3]. We describe a proposed interdisciplinary approach to designing for answerability in autonomous systems, grounded in an instrumentalist framework of 'responsible agency cultivation' drawn from moral philosophy and cognitive sciences as well as empirical results from structured interviews and focus groups in the application domains of health, finance and government. We outline a prototype dialogue agent informed by these emerging results and designed to help bridge the structural gaps in organisations that typically impede the human agents responsible for an autonomous sociotechnical system from answering to vulnerable patients of responsibility.

*Authors addresses: Louise Hatherall, Usher Institute, The University of Edinburgh, Edinburgh, UK; email: lhathera@ed.ac.uk; Dilara Keküllüoğlu, School of Informatics, The University of Edinburgh, Edinburgh, UK; email: d.kekulluoglu@ed.ac.uk; Nadin Kokciyan, School of Informatics, The University of Edinburgh, Edinburgh, UK; email: nadin.kokciyan@ed.ac.uk; Michael Rovatsos, Bayes Centre, School of Informatics, The University of Edinburgh, Edinburgh, UK; email: michael.rovatsos@ed.ac.uk; Nayha Sethi, Usher Institute, The University of Edinburgh, Edinburgh, UK; email: nayha.sethi@ed.ac.uk; Tillmann Vierkant, Department of Philosophy, The University of Edinburgh, Edinburgh, UK; email: t.vierkant@ed.ac.uk; Shannon Vallor, Centre for Technomoral Futures, Department of Philosophy, The University of Edinburgh, Edinburgh, UK; email: svallor@ed.ac.uk.

TAS '23, July 11, 12, 2023, Edinburgh, United Kingdom

ACM ISBN 979-8-4007-0734-6/23/07.
https://doi.org/10.1145/3597512.3597529

KEYWORDS

Responsibility gaps, Agency, Answerability, Dialogue agents, AI ethics, Sociotechnical Systems Design

ACM Reference Format:
Louise Hatherall, Dilara Keküllüoğlu, Nadin Kokciyan, Michael Rovatsos, Nayha Sethi, Tillmann Vierkant, and Shannon Vallor. 2023. Responsible Agency Through Answerability: Cultivating the Moral Ecology of Trustworthy Autonomous Systems. In *First International Symposium on Trustworthy Autonomous Systems (TAS '23), July 11, 12, 2023, Edinburgh, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3597512.3597529

1 RESPONSIBILITY GAPS AS ANSWERABILITY GAPS

A widely discussed obstacle to the trustworthiness of autonomous systems is known as the responsibility gap problem [1–3]. Responsibility gaps emerge when an action is taken for which there are no agents clearly identifiable as morally responsible, but the act is, or is widely perceived by society as, requiring accountability; that is, someone to hold responsible. This is distinct from the problem of responsibility evasion, where an entity who clearly does meet the conditions for moral responsibility seeks to avoid being held responsible. However, like responsibility evasions, responsibility gaps endanger public welfare, trust and confidence, and pose an obstacle to AS adoption and innovation.

Today, organisations are rapidly developing and deploying a range of AS that in a growing number of cases, take actions for which: (1) due to their high moral stakes, society demands a clear assignment of moral responsibility, and yet: (2) no stakeholder in the system's actions meets the standard conditions for rightful attributions of moral responsibility: namely, adequate knowledge and control of the action. Applications where new responsibility gaps are emerging include autonomous vehicles and weapons systems, credit decision and trading algorithms, recommender algorithms and learning chatbots, and AI health and diagnostic tools.

Causes of responsibility gaps in AS include: 1) uninterpretability of 'black-box' deep learning models, 2) stochastic uncertainty regarding the distribution and types of learning algorithm errors; 3) intrinsic unpredictability of adaptive systems that learn in real-world environments; 4) model brittleness due to unexpected data drift; 5) speed or complexity of AS operations that impede 'human

in the loop' control or efficacious auditing; 6) unpredictable and/or untraceable interactions or influences on AS involving a widely distributed multitude of components, systems and agents, including many human actors who may be anonymous, opaquely involved, or otherwise unspecifiable.

While legal responsibility for AS can be assigned even in the absence of a *morally* responsible agent, e.g., via strict tort liability, this does not eliminate the gap. Liability untethered to plausible moral responsibility may be seen as unjust, if assigned to agents who lack the power to foresee or prevent AS harms, or if the remedies merely impose externalities of AS upon publics, for example, by placing costly legal burdens of accountability upon taxpayer-funded institutions.

Responsibility gaps are, however, not unique to the AI/AS context. As we outline in our own account, drawing from both the cognitive science literature and the literature on corporate responsibility, both individual human persons and organisations routinely fail to be able to give a satisfactory account of what they have done or why, in the manner we would expect from an ideally responsible agent with knowledge and control of their actions. However, even those who deny such gaps are novel [3–6] tend to admit to strong "normative mismatches" [7] between AS and public attitudes about responsibility. Until these are bridged, even safe and beneficial AS will operate outside the bounds of social trust.

Responsibility gaps thus undermine AS trustworthiness in principle and practice, not only by creating challenges for AS governance, but by violating shared social expectations of justice, namely that serious harms resulting from agent decisions or deliberate actions (vs. *force majeure* events) are *answerable* by said agent(s). Yet today's autonomous machines, on any conventional understanding, lack the ability to be morally answerable to a human being. They lack the moral agency and understanding to meaningfully acknowledge and respond appropriately to the moral concerns or demands expressed by a person who they have harmed or endangered. They also lack the moral vulnerability to what philosophers call the 'reactive attitudes' – sentiments like blame, indignation, resentment, forgiveness, guilt. These serve in traditional responsibility practices to both equalize the vulnerability between agent and the harmed person [or *patient of responsibility*], and to motivate the cultivation of more responsibility and moral competence in the agent going forward.

Indeed, recent literature on responsibility gaps has begun to converge around this theme of answerability [2, 8, 9]. Answerability is a dimension of responsibility practices that scholars have long distinguished from related responsibility concepts such as attribution, accountability, and liability [10]. Our work is grounded in a philosophical and empirical account of answerability as a distinctive responsibility practice that is broader and richer in meaning than what in the AI ethics literature often is labeled 'explainability' or even 'accountability.' Our aim is to provide the theoretical framework, empirical evidence, and computational techniques that demonstrate how to enable AS (including the wider "system" of developers, owners, users, etc.) to supply the kinds of answers that people seek from trustworthy agents.

We draw initial inspiration from philosophical and cognitive science literature showing that responsibility gaps are not a unique AS problem, as humans often don't understand or control the causal etiologies of their behaviour. Yet humans routinely bridge such responsibility gaps through moral dialogue with other impacted parties, even when viable explanations are unavailable. An approach to responsibility as answerability offers invaluable lessons in the AS context, which can help mitigate the persistent explainability challenges to current research efforts in AS responsibility. This offers a constructive way forward in bridging responsibility gaps in AS. We show how answerability promotes the cultivation of responsibility for autonomous system actions in our wider 'moral ecology' [11], even when the morally responsible agent(s) remains underspecified or uncertain.

For example, answerability practices give patients of responsibility a channel by which to probe the trustworthiness of a system over the course of a dialogue, and conversely, a channel for human and machine agents that compose an autonomous system to demonstrate the system's trustworthiness to patients and wider publics over time. Notably, answering for past acts can improve a human agent's future trustworthiness; we are investigating this potential for AS as well. Our project unfreezes the responsibility gap problem by shifting its focus from the properties of the agent (whether machine, human or a distributed/hybrid network), to the relationship between systems and patients (including broader publics) who are owed answers for AS actions and consequences. When responsibility is framed as the fulfilment of a relationship rather than a fixed fact, it can be seen as a practice to be refined, supplemented, negotiated, and reconstructed by the relevant parties to the relationship. We will show how a body of existing and constructible answerability practices can be drawn upon by AS designers, developers, users, and regulators to strengthen responsibility relationships in dialogue with patients of responsibility, as part of the wider moral ecology containing autonomous systems.

We need not only social and regulatory but computational means of strengthening AS answerability. To that end, we outline below a prototype for a dialogical agent that can function as an effective intermediary for translating and transporting answerability concerns and responses, by serving as a dialogical bridge between the different human and machine components of a sociotechnical system and the impacted patients of responsibility to whom that system – including its human elements – must answer. In order for our prototype agent to be effective, however, our design principles must be informed by empirical evidence of the kinds of answers that patients of responsibility are likely to expect, demand, or accept from trustworthy autonomous systems. The currently active phase of our project is gathering this evidence by means of structured interviews, focus groups and stakeholder workshops designed to elicit the answerability expectations likely to emerge in the application domains of health, finance and government use of AS.

2 EXPLORING STAKEHOLDER PERSPECTIVES ON AUTONOMOUS SYSTEMS & ANSWERABILITY

The proliferation of the development and application of AS poses legal and regulatory challenges for a range of reasons: from the need to regulate new forms of conduct, to issues of 'fit' between existing regimes and new technologies [12]. Scholars and practitioners are grappling with questions about who is legally responsible when

AS causes harm [13], clarifications of how new technology fits existing regulatory frameworks (such as large language models and medical software applications) [14], and who should bear the responsibilities of education and communicating decisions made with AS [15]. Underpinning these discussions is a recognition that different stakeholders (including organisations, employees, and broader publics) will need to trust such systems to ensure effective adoption and implementation.

Bhatt [16] and Esmaeilzadeh [17] have shown that the integration of AI into public sector and healthcare settings respectively require trust, and that this is built through stakeholder engagement – although there is limited work on identifying who the stakeholders are, how they might engage, and what the relationships between these stakeholders look like in navigating the developing AS landscape. Trust, gained through transparency, robust governance, choice, and the ability to dispute has also been highlighted as central to deploying medical AI [18, 19] but there remain questions of what knowledge is adequate to understand such transparency and make informed choices, and who should be responsible for ensuring this knowledge. Whilst many studies explore accountability or explainability to resolve the questions identified, what is underexplored is the concept of answerability and how an answerability approach can be implemented in developing regulatory frameworks. This project aims to fill this gap by identifying the answers that a range of stakeholders need to be able to trust AS, and how an answerability-focused approach can be implemented within regulatory approaches.

The goals of Workstream 2 (WS2) of this project are to:
Contribute towards the development of a novel, empirically informed, community-embedded framework of trustworthiness as answerability to bridge responsibility gaps caused by autonomous systems;
Explore how to embed this framework of answerability in regulatory frameworks and understand the implications of doing so; and
Provide empirical evidence to demonstrate how to enable autonomous systems (including the wider 'systems' of humans) to supply the kinds of answers that people seek from trustworthy agents.

The first step in developing the answerability framework is to understand what answers are needed in different contexts. We use socio-legal methods to identify the perspectives on answerability of different patients of responsibility. Our research focuses on three key areas where use of autonomous systems is rapidly progressing: health, finance, and government. We use a variety of methods outlined in Figure 1. Semi-structured schedules were developed for both the scoping conversations and the interviews. These were designed in collaboration with the wider TAS team. Key themes emerging from the initial scoping conversations also informed the questions asked in the interviews. The data was analysed using Braun and Clarke's [20] approach to reflexive thematic analysis. Notes and transcripts were read and re-read to identify salient segments of data (or codes) which were then grouped into themes. This process was iterative – moving back and forth between the data, the codes, and themes to generate insights.

Alongside gathering empirical data on answerability expectations, our work seeks to identity and critically evaluate current and developing regulatory frameworks for autonomous systems. We aim to identify current and anticipated challenges posed by autonomous systems for actors navigating this dynamic regulatory space. This will allow us to explore how to embed the answerability practices identified through our empirical research within regulatory and systems design, providing community driven mechanisms for doing so. This shift in responsibility practices is likely to have regulatory implications and, as such, the final aim of this workstream is to assess this impact through focus groups and deliberative workshops with key stakeholders including clinicians, regulators, developers, policy makers, and lay publics. A multipronged empirical data collection is proposed (Figure 1) to achieve the goals.

2.1 Initial Findings from Scoping Conversations and Interviews

2.1.1 Trustworthiness. Trust in AS is gained (and maintained) through **relations** (e.g., with the organization, AS developers, regulators, regulation) **characteristics** (knowledge, understanding, robustness, resilience, transparency) and **standards** (ethics, explainability, data protection)

2.1.2 Answers. The answers different patients of responsibility need can tentatively be split into different high-level categories of 'practical' (how exactly do I use this system?), 'reasons' (why am I using this system?), and 'relational' (who else uses it?).

2.1.3 Assurances. Some patients of responsibility need assurances about themselves or their abilities to rely on AS – and these assurances are often dialogic. For example, someone might want to know: am I using this system correctly? If something unexpected occurs, is it my fault due to how I interacted with the system?

2.1.4 Human-Centred Approaches. There is evidence of human-centred approaches (that is, focusing on the needs of different patients of responsibility who interact with AS) but these are informed assumptions based on previous experiences. Further empirical evidence about what a human-centred approach looks like in practice is needed.

2.1.5 Regulation. One of the key relations for trustworthy systems is a trusted regulator with a clear regulatory framework. This regulator – regardless of sector – should be able to provide robust answers concerning the systems they regulate and know what questions to ask to get those answers (e.g., in-depth evidentiary analysis, details of certification, clear details of investigatory processes). A regulatory 'light-touch' approach is less desirable than regulatory clarity.

This raises important questions for the future regulation of AS including: what is the best approach for regulating the socio-technical nature of AS (guidelines, standards, laws)? Where does the regulation of actors who design and develop AS fit within the framework of building trust and providing answers? How can regulation facilitate dialogue between groups with different levels of technical understanding and different language? And what are the different stakeholder understandings of regulatory concepts (such as 'transparency') in these contexts and how are these navigated in practice?

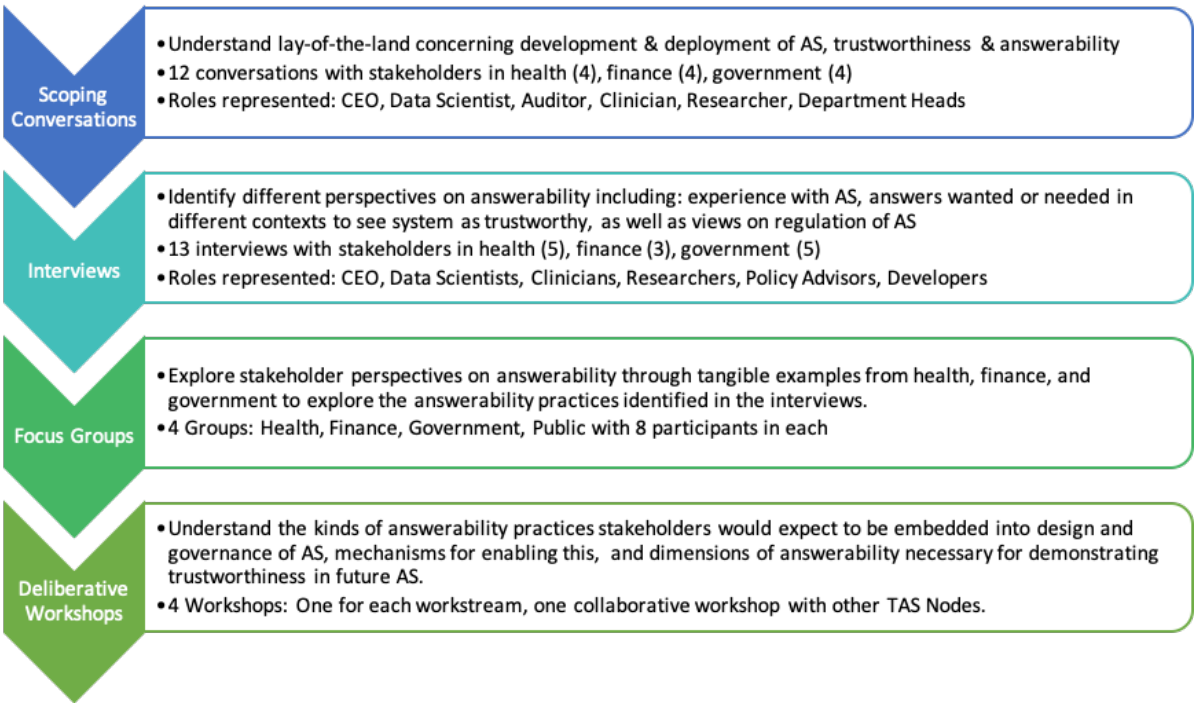

Figure 1: The Proposed Empirical Data Collection Pipeline

2.2 Future Outputs

We will communicate future findings from this research through academic journal articles, policy briefings, conference presentations, and through contributions to the practitioner handbook. You can keep up to date here: https://tas.ac.uk/making-systems-answer/

3 ENABLING ANSWERABILITY IN SOCIOTECHNICAL SYSTEMS

From the standpoint of developing future AI technologies that can, at least to some extent, learn to adopt the practices that help bridge responsibility gaps, the question is what such technologies might look like, and how we would go about designing and testing them.

To address this research question, we aim to create a mediator agent framework capable of exhibiting functionalities that improve the answerability of the sociotechnical systems. This framework will provide a blueprint for implementing specific mediator agents that can provide, together with human actors who can assume responsibilities for the target AI system's behaviour in the wider human-machine sociotechnical system, the answers that patients of responsibility are looking for. At the most basic level, the idea is to initiate a conversation between people who were harmed by autonomous decisions and the organisations employing these systems. The developed mediator agent will speed up dealing with the questions and complaints of users, identifying common harmful or inappropriate actions of the autonomous system, and help the organisation using it develop and exhibit enhanced responsibility practices. Over time, we expect that, by learning to handle responsibility-related interactions with users affected by the system, the mediator agent will learn to propose and enact methods for resolving situations in line with the expectations of users harmed by the system and the organisation responsible for it.

Developing this functionality will require (1) Creating a formal dialogue representation to enable communication between people and organisations; (2) Providing an answerability framework that could be employed by organisations; and (3) Developing a prototype to show the applicability of the developed framework to enable answerability in sociotechnical systems.

3.1 Dialogue Representation

To facilitate the conversation between different actors, we propose a formal representation of a dialogue that could be automatically analyzed by the mediator agent. Our proposed dialogue representation consists of three main components: (1) Explanation, (2) Action Update, and (3) Remedy. The explanation stage aims to help actors reach a common understanding regarding the reasons for which an automated decision was taken by the system, and the factors that contributed to it. If there is a disparity between the user's view and the actual operation of the system, false information should be corrected or deleted to update the action being recommended.

Unfortunately, the automated decision might have caused harm that cannot be remedied by only updating the user's records. Even with corrected information, automated decisions may still cause harm to people. Hence, we propose to have a remedy component at the final stage of a dialogue. This is to ensure the people harmed can seek restitution and the harm can be mitigated by a remedy agreed upon mutually. The remedy can be in different forms. For example, it can be a monetary compensation or an offer of providing a discount for future services. They can be also used to take forward-looking

responsibility such as fixing the automated system to make better decisions for other users. To implement these proposed components, we specify the steps each actor in the dialogue can make and the corresponding dialogue acts. We track the state of the dialogue with our dialogue state tracker, which helps the mediator agent to steer the conversation.

3.2 Answerable Sociotechnical Systems

We define a common terminology (i.e., an ontology) to be used to represent information flow between different components of the sociotechnical system. The developed ontology is also capable of representing relations between different actors and components of the system. We have four main entities represented in the developed ontology: (1) Organisation, (2) AI/Autonomous System, (3) Stakeholders in the system (e.g., Action subject, Developers), (4) Entities to represent various dialogue components (e.g., Reason, Remedy).

The mediator agent is equipped with an ontology instance to explain the reasoning process behind the specific actions, where actions are associated with a set of possible reasons. If there is a question that the mediator agent is not able to respond or it has no authority to respond automatically (i.e., it requires approval from a human operator), the mediator agent will identify responsible agents within the organisation using its ontology (e.g., the lead developer working on the implementation of an algorithm) to help building a collective response. Or the mediator agent can also connect people with a human operator, if required.

3.3 Chatbot-based Prototype

Our initial prototype is based on a chatbot technology that facilitates dialogue between different actors. In this prototype, the mediator agent connects the dialogue components with its ontology to enable answerability. The mediator agent iterates on the process of giving answers (automated or semi-automated) to the patients of responsibility based on their queries. To evaluate the usability aspects of such a prototype, we are planning to conduct user studies with different scenarios from various sectors.

ACKNOWLEDGMENTS

This research was supported in part by the UKRI Engineering and Physical Sciences Research Council (grants EP/V026607/1 and EP/W011654/1); UKRI Arts and Humanities Research Council (grant AH/X007146/1); The John Templeton Foundation and Fetzer Foundation (a Consciousness and Free Will Joint Grant); and The Wellcome Trust (grant 209519/Z/17/Z).

REFERENCES

[1] Andreas Matthias. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6, 175–183. DOI: https://doi.org/10.1007/s10676-004-3422-1
[2] Mark Coeckelbergh. 2020. Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics 26, 2051–2068. DOI: https://doi.org/10.1007/s11948-019-00146-8
[3] Daniel W. Tigard. 2021. There is no techno-responsibility gap. Philos. Tech. 34, 589–607. DOI: https://doi.org/10.1007/s13347-020-00414-7
[4] Deborah G. Johnson. 2011. Software agents, anticipatory ethics, and accountability. In: Marchant, G., Allenby, B., Herkert, J. (eds) The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight. Springer, Dordrecht. 61-76.
[5] Johannes Himmelreich. 2019. Responsibility for killer robots. Ethical Theory and Moral Practice 22:3, 731-747.
[6] Peter Königs. 2022. Artificial intelligence and responsibility gaps: what is the problem? Ethics Inf Technol 24, DOI: https://doi.org/10.1007/s10676-022-09643-0
[7] Köhler, S., Sauer, H., & Roughley, N. (2017). Technologically blurred accountability? Technology, responsibility gaps and the robustness of our everyday conceptual scheme. In C. Ulbert, P. Finkenbusch, E. Sondermann, & T. Debiel (Eds.), Moral Agency and the Politics of Responsibility (pp. 51–67). Routledge.
[8] Daniel W. Tigard. 2021. Technological answerability and the severance problem: Staying connected by demanding answers. Sci Eng Ethics 27(59), 1-20. DOI: https://doi.org/10.1007/s11948-021-00334-5
[9] Maximilian Kiener. 2022. Can we bridge AI's responsibility gap at will? Ethic Theory Moral Prac 25, 575–593. DOI: https://doi.org/10.1007/s10677-022-10313-9
[10] R. A. Duff. 2009. Legal and moral responsibility. Philos Compass 4(6): 978–986. DOI: https://doi.org/10.1111/j.1747-9991.2009.00257.x
[11] Shannon Vallor and Bhargavi Ganesh. 2023. AI and the imperative of responsibility: Reconceiving AI governance as social care. In M. Kiener (Ed.), The Routledge Handbook of Philosophy of Responsibility (Chapter 31). Routledge.
[12] Lyria Bennett Moses. 2007. Recurring dilemmas: The law's race to keep up with technological change. Illinois Journal of Law, Technology and Policy, Vol 2007:2. 239-285.
[13] David Nersessian and Ruben Mancha. 2020. From automation to autonomy: legal and ethical responsibility gaps in artificial intelligence innovation. Michigan Technology Law Review 27: 55. DOI: https://doi.org/10.36645/mtlr.27.1.from
[14] Johan Ordish. 2023. Large language models and software as a medical device. MedRegs, MHRA. Retrieved from https://medregs.blog.gov.uk/2023/03/03/large-language-models-and-software-as-a-medical-device/
[15] Martin Sand, Juan Manuel Duran and Karin Rolanda Jongsma. 2021. Responsibility beyond design: Physicians requirements for ethical medical AI. Bioethics 36: 162 – 169. DOI: https://doi.org/10.1111/bioe.12887
[16] Umang Bhatt, McKane Andrus, Adrian Weller and Alice Xiang. 2020. Machine learning explainability for external stakeholders. arXiv: 2007.05408. Retrieved from https://arxiv.org/abs/2007.05408
[17] Pouyan Esmaeilzadeh. 2020. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. BMC Med Inform Decis Mak 20: 170. DOI: https://doi.org/10.1186/s12911-020-01191-1
[18] NHS AI Lab & Health Education England. 2022. Understanding healthcare workers' confidence in AI. Report 1 of 2. Retrieved from https://digital-transformation.hee.nhs.uk/binaries/content/assets/digital-transformation/dart-ed/understandingconfidenceinai-may22.pdf. 90 pages.
[19] Jordan P. Richardson, Cambray Smith, Susan Curtis, Sara Watson and Richard R. Sharp. 2021. Patient apprehensions about the use of artificial intelligence in healthcare. npj Digit. Med. 4, 140. DOI: https://doi.org/10.1038/s41746-021-00509-1
[20] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. Qualitative Research in Psychology 3(2): 77 – 101, DOI: 10.1191/1478088706qp063oa