Medical Law Review, 2023, **00**, 1–20
https://doi.org/10.1093/medlaw/fwad013
Original article

Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability

Caroline Jones [1,*], James Thornton[2], Jeremy C. Wyatt[3]

[1]Hillary Rodham Clinton School of Law, Swansea University, Swansea, UK
[2]Nottingham Law School, Nottingham Trent University, Nottingham, UK
[3]Wessex Institute, University of Southampton, Southampton, UK

*Corresponding author: caroline.jones@swansea.ac.uk

ABSTRACT

Artificial intelligence (AI) could revolutionise health care, potentially improving clinician decision making and patient safety, and reducing the impact of workforce shortages. However, policymakers and regulators have concerns over whether AI and clinical decision support systems (CDSSs) are trusted by stakeholders, and indeed whether they are worthy of trust. Yet, what is meant by trust and trustworthiness is often implicit, and it may not be clear who or what is being trusted. We address these lacunae, focusing largely on the perspective(s) of clinicians on trust and trustworthiness in AI and CDSSs. Empirical studies suggest that clinicians' concerns about their use include the accuracy of advice given and potential legal liability if harm to a patient occurs. Onora O'Neill's conceptualisation of trust and trustworthiness provides the framework for our analysis, generating a productive understanding of clinicians' reported trust issues. Through unpacking these concepts, we gain greater clarity over the meaning ascribed to them by stakeholders; delimit the extent to which stakeholders are talking at cross purposes; and promote the continued utility of trust and trustworthiness as useful concepts in current debates around the use of AI and CDSSs.

KEYWORDS: Artificial intelligence, Clinical decision support, Clinicians' perspectives, Liability, Trust, Trustworthiness

I. INTRODUCTION

Artificial intelligence (AI) could revolutionise health care, potentially improving patient safety and clinician decision making and reducing the impact of workforce shortages. Press coverage speaks of the hope and promise of 'transforming the NHS', performing 'as well as a

top consultant every time',[1] or (in some cases) claiming to beat them.[2] Projects such as Microsoft's 'InnerEye' (using AI to detect tumours from radiography scans),[3] Google's DeepMind acute kidney failure algorithm Streams, and Babylon's Triage and Diagnostic system claim to be as good as, or better, than human doctors.[4] Such tools could replicate high-level specialist performance in the disease in question (using fixed algorithms, effectively putting the diagnostic expertise of a top consultant into every GP's computer) or, in future, go beyond their initial programming to learn and improve in the field (the so-called 'self-learning' AI systems) becoming better than the state of the art.[5]

Clinical decision support systems (CDSSs) using AI are digital 'active knowledge systems which use two or more items of patient data to generate case-specific advice' for clinicians.[6] CDSSs may provide guidance on diagnosis, treatment, maintenance/follow-up treatment, workflow(s), and patient information, and indeed may use AI or other methods to generate their outputs/advice (our observations can apply to both AI and non-AI-based CDSSs). However, as several commentators have made clear, CDSSs are intended to support rather than to supplant clinical decision making.[7] Systematic reviews of *current* CDSSs show that they make some clinical activities, such as prescribing or preventative care, safer.[8] On the other hand, there have also been criticisms; for example, a response to Babylon's research raised a number of concerns in relation to safety and the methodological limitations of the grand claims made by the firm.[9]

[1] Ian Sample, '"It's Going to Create a Revolution": How AI is Transforming the NHS' *The Guardian* (London, UK, 4 July 2018) <https://www.theguardian.com/technology/2018/jul/04/its-going-create-revolution-how-ai-transforming-nhs> accessed 10 November 2022.

[2] Aliya Ram, 'Google's AI Beats Doctors at Spotting Eye Disease in Scans' *Financial Times* (London, UK, 13 August 2018) < https://www.ft.com/content/3de44984-9ef0-11e8-85da-eeb7a9ce36e4> accessed 10 November 2022.

[3] 'Project InnerEye—Democratizing Medical Imaging AI' (*Microsoft*) <https://www.microsoft.com/en-us/research/project/medical-image-analysis/> accessed 11 November 2022.

[4] Salman Razzaki and others, 'A Comparative Study of Artificial Intelligence and Human Doctors for the Purpose of Triage and Diagnosis' 2018 Babylon Health 1; Jen Copestake, 'Babylon Claims Its Chatbot Beats GPs at Medical Exam' (*BBC*, 27 June 2018) https://www.bbc.co.uk/news/technology-44635134> accessed 10 November 2022.

[5] Such tools have the greatest potential, but also carry the greatest risk. Indeed, some argue the capacity for mistakes or unpredictability (even the developers themselves cannot know how it will learn once released and put into the field) makes them are intrinsically unsuitable for some high stakes/safety contexts, such as automated insulin pumps or nuclear power stations. See Comments from the Governance in AI Research Group (GAIRG) on the proposed EU AI Regulation' <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements_en> accessed 10 May 2023.

[6] Jeremy C Wyatt and David J Spiegelhalter, 'Field Trials of Medical Decision-Aids: Potential Problems and Solutions' (1991) Proceedings of 15th Annual Symposium on Computer Application in Medical Care 3.

[7] Diana Brahams and Jeremy C Wyatt, 'Decision Aids and the Law' (1989) 334 (8663) The Lancet 632; Helen Smith and Kit Fotheringham, 'Artificial Intelligence in Clinical Decision-making: Rethinking Liability' (2020) 20(2) Medical Law International 131; Caroline Jones, James Thornton, and Jeremy C Wyatt, 'Enhancing Trust in Clinical Decision Support Systems: A Framework for Developers' (2021) 28(1) BMJ Health & Care Informatics (Online); on the terminology in this context see the argument that CDSS should instead be referred to as CRSS, ie clinical *reasoning* support systems (emphasis added) to better reflect the 'hybrid intelligence' of clinicians and AI in reaching decisions about specific patients, in Sophie van Baalen, Mieke Boon and Petra Verhoef, 'From Clinical Decision to Clinical Reasoning Support Systems, 2021 27(3) Journal of Evaluation in Clinical Practice 520; Megan Prictor, 'Where Does Responsibility Lie? Analysing Legal and Regulatory Responses to Flawed Clinical Decision Support Systems When Patients Suffer Harm' 2022 Medical Law Review 1.

[8] Amit X Garg and others, 'Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review' (2005) 293(10) JAMA 1223; Pavel S Roshanov, 'Features of Effective Computerised Clinical Decision Support Systems: Meta-regression of 162 Randomised trials' (2013) 346 BMJ (Online); Julian Varghese and others, 'Effects of Computerized Decision Support System Implementations on Patient Outcomes in Inpatient Care: A Systematic Review' (2018) 25(5) Journal of the American Medical Informatics Association 593; Janice L Kwan and others, 'Computerised Clinical Decision Support Systems and Absolute Improvements in Care: Meta-analysis of Controlled Clinical Trials' 2020 370 BMJ (Online).

[9] Hamish Fraser, Enrico Coiera and David Wong, 'Safety of Patient-facing Digital Symptom Checkers' (2018) 392(10161) The Lancet 2263; on the general reliability of studies in this area see also Xiaoxuan Liu and others 'A Comparison of Deep Learning Performance against Health-care Professionals in Detecting Diseases from Medical Imagining: A Systemic Review and Analysis' (2019) 1(6) The Lancet Digital Health e271.

At the forefront of policymakers' concerns are the interrelated concepts of trust[10] and trustworthiness.[11] The House of Lords Select Committee on AI identified a 'need to build public trust and confidence' in AI generally,[12] but in the evidence sessions there was a lack of consensus on this issue. In their evidence to the Committee, 'many AI researchers were concerned that the public were being presented with overly negative or outlandish depictions of AI'; whereas other respondents (from backgrounds in law and ethics, history of science and technology, statistics, and public policy making) 'warned against simplistically attempting to build trust in AI, as at least some applications of AI would not be worthy of trust.'[13] More recently, the European Union's (EU) stated aim (culminating in the Commission's 2021 Proposal for a Regulation on AI)[14] is for 'the development of an ecosystem of trust by proposing a legal framework for trustworthy AI',[15] such that the EU can become 'a global leader in the development of secure, trustworthy and ethical artificial intelligence'.[16]

Yet, as the Department of Digital, Culture Media and Sports' (DCMS) Centre for Data, Ethics and Innovation (CDEI) has concluded, the 'one fundamental barrier' to opportunities with AI was 'low levels of public trust' and, in the health context in particular, that practitioner (clinician) distrust was more prominent than in other sectors.[17] Other empirical studies have also highlighted a lack of trust on the part of medical practitioners, with fears expressed that CDSSs 'may reduce their professional autonomy or may be used against them in the event of medical-legal controversies'.[18] Hence, despite the huge potential of AI in healthcare contexts, uptake of these tools has been described as 'slow',[19] and trust and trustworthiness remain critical concerns for various stakeholders including clinicians.

In this article, we use Onora O'Neill's multi-dimensional understanding of trust and trustworthiness to analyse clinicians' perspectives on the use of AI and CDSSs, set within the current legal and regulatory positions governing their use in the UK.[20] We argue that analysis using O'Neill's framework allows for a productive understanding of the issues raised in this area, and moves beyond the one-dimensional references to trust seen in some earlier policy

[10] Select Committee on Artificial Intelligence, 'AI in the UK: Ready, Willing and Able?' (*House of Lords*, 2018) https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.> accessed 11 November 2022; <Commission, 'Artificial Intelligence for Europe' (Factsheet) COM (2018); Commission, 'Building Trust in Human-Centric Artificial Intelligence' (Communication) COM (2019) 168 final.

[11] Commission High Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (8 April 2019); this report identified three elements for trustworthy AI: lawful, ethical, and robust, but focused only on the latter two. See also the seven key principles for ethical AI: Commission, 'Artificial Intelligence for Europe' COM (2018) 237 final, 2, s 3.3. Commission, 'Liability for Emerging Digital Technologies' (Staff Working Document accompanying Commission, 'Artificial Intelligence for Europe' COM (2018) 237 final). Also, on 'establishing an appropriate governance and regulatory framework', see Commission High Level Expert Group' 'AI, Policy and Investment Recommendations for Trustworthy AI' (2019) 37–42; Commission, 'White Paper on AI: A European Approach to Excellence and Trust' COM (2020) 65 final; European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL)).

[12] Select Committee on Artificial Intelligence (n 10).

[13] ibid 47–48.

[14] Commission, '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 Union Legislative Acts' COM (2021) 206 final.

[15] ibid s 1.1.

[16] ibid para 5.

[17] Roger Taylor, 'AI Barometer Independent Report' (*Centre for Data Ethics and Innovation*, 2020), s 1–s 35.4. <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/894170/CDEI_AI_Barometer.pdf> accessed 11 November 2022.

[18] On the failed introduction of the PRODIGY programme into the NHS see Nikki Rousseau and others 'Practice Based, Longitudinal, Qualitative Interview Study of Computerised Evidence Based Guidelines in Primary Care' 2003 326 BMJ 314; Elisa G Liberati and others, 'What Hinders the Uptake of Computerized Decision Support Systems in Hospitals? A Qualitative Study and Framework for Implementation' 2017 12(1) Implement Science 113; Haroldas Petkus, Jan Hoogewerf, and Jeremy C Wyatt, 'AI in the NHS– are Physicians Ready? A Survey of the Use of AI & Decision Support by Specialist Societies, and Their Concerns' (2020) 20 Clinical Medicine 324; Marie Caroline Lai, M Brian and Marie-France Mamzer, 'Perceptions of Artificial Intelligence in Healthcare: Findings from a Qualitative Survey Study among Actors in France' (2020) 18(1) Journal of Translational Medicine 14.

[19] Baalen, Boon and Verhoef (n 7).

[20] Onora O'Neill, 'Linking Trust to Trustworthiness' (2018) 26(2) International Journal of Philosophical Studies 293.

material.[21] We do not limit our analysis to a particular clinical setting, but instead focus on the key generic features and issues that apply to AI and CDSSs in terms of trust generally and are therefore of the broadest concern and interest. We do not discuss legal aspects of data access for the training of CDSSs, as this is well covered elsewhere[22]; nor patient use of mHealth apps,[23] or health professionals' use of passive 'reference information', such as a simple website or digitised textbook, as opposed to active decision support systems. Equally, we do not cover autonomous systems such as automated diabetes or analgesia controllers, in which there is no human in the control loop. As highlighted above, we focus primarily on clinicians' perspectives (and, to a lesser extent, those of patients)[24]; whilst suppliers/developers play a key role we do not address them in detail because they are not trusting anyone (rather, others are trusting them), and we have considered them in previous work.[25]

II. TRUST AND TRUSTWORTHINESS: THEORETICAL UNDERPINNINGS

Although interrelated 'trust and 'trustworthiness' are separate concepts and should not be conflated.[26] As O'Neill notes, trust is only valuable when directed at agents or things that are worthy of it: that is, those that are 'trustworthy'[27]; as when the untrustworthy are naively trusted, the results can be ruinous, or indeed fatal.[28]

With AI and CDSSs, it is important to be clear about exactly what or who is being 'trusted'. In everyday speech, we might say flippantly that we do or do not 'trust' many things (medicines, vaccines, train timetables, etc), but fundamentally only beings with agency are truly capable of being trusted (or indeed distrusted).[29] As Baier highlights, if we say that we 'trust' a chair not to collapse underneath our weight, then what we really mean is that we trust the people involved, such as the person who designed, built or sold it, rather than the thing itself.[30] The chair does not decide to collapse underneath you (so we would not say it is 'untrustworthy'), but it may collapse if the human who designed/built it did a poor job. This distinction becomes slightly less clear-cut when discussing AI. Baier would still consider trust (or lack of it) to ultimately lie with the human beings who designed (or developed/supplied/tested/used) the AI or CDSS in question,[31] though she does acknowledge that an artificial mind might be capable of assuming a position of trust.[32]

[21] Select Committee on Artificial Intelligence (n 10); Wendy Hall and Jerome Presenti, 'Growing the artificial intelligence industry in the UK' (Independent Report 2017) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf> accessed 20 March 2023.

[22] Eleonora Harwich and Kate Laycock, 'Thinking on Its Own: AI in the NHS' (*Reform*, January 2018) <https://www.wiltonpark.org.uk/wp-content/uploads/Thinking-on-its-own-AI-in-the-NHS.pdf> accessed 11 November 2022.

[23] See Maria K Sheppard, 'mHealth Apps: Disruptive Innovation, Regulation, and Trust—A Need for Balance' (2020) 28(3) Medical Law Review 549.

[24] It is notable and unfortunate that there is very little consideration of patients' perspectives in the current literature. The authors are currently conducting a BA/Leverhulme-funded empirical research project to rectify this gap.

[25] Jones, Thornton, and Wyatt (n 7).

[26] Russell Hardin, *Trust and Trustworthiness* (Russell Sage Foundation 2002) 55; Matthew Fenech, Nika Strukelj and Olly Buston, 'Ethical, Social, and Political Challenges of Artificial Intelligence in Health' (*Future Advocacy*, 2018) <https://cms.wellcome.org/sites/default/files/ai-in-health-ethical-social-political-challenges.pdf accessed> accessed 11 November 2022 41.

[27] O'Neill (n 20); see also Mark A Hall and others, 'Trust in Physicians and Medical Institutions: What Is It, Can It be Measured, and Does It Matter? 2001 79(4) The Millbank Quarterly 613, 616.

[28] Dr Harold Shipman might be considered an extreme example; see the discussion by Paula Case, 'Putting Public Confidence First: Doctors, Precautionary Suspension and the General Medical Council' 2011 19(3) Medical Law Review 339.

[29] Joshua James Hatherly, 'Limits of Trust in Medical AI' (2020) 46(7) Journal of Medical Ethics 478.

[30] Anette Baier, 'What is Trust' in Monique Deveaux and others (eds), *Reading Onora O'Neill* (Taylor & Francis Group 2013).

[31] See also Margit Sutrop, 'Should We Trust Artificial Intelligence?' (2019) 23(4) Trames Journal of the Humanities and Social Science 499, 512.

[32] Baier (n 30).

Irrespective of who (or what) is being trusted, O'Neill argues that there is a much more conceptual problem with understandings of 'trust' generally in most contexts: 'much contemporary work on trust—such as that based on polling evidence—studies generic attitudes of trust in types of agent, institution or activity in complete abstraction from any account of trustworthiness.'[33] Therefore, it is perhaps unsurprising that the literature on AI, CDSSs, and health care tends to take a similar approach. 'Trust' may be presented as a neutral, abstract concept in the literature. However, as the differing perspectives noted in the evidence session to the House of Lords Select Committee on AI suggest, it can be harmful or helpful, depending on whether it is placed in those who are worthy of trust or not.[34] Trusting a fraudster with all one's savings would be harmful, whereas trusting a mainstream bank (instead of storing cash under one's mattress) would be helpful. As O'Neill notes: 'trust is valuable when placed in trust-worthy agents and activities, but damaging or costly when (mis)placed in untrustworthy agents and activities.'[35]

Nevertheless, O'Neill's statement requires further unpacking. For example, it is possible for a doctor to be both honest and dependable, always keeping appointments, never breaching confidentiality, etc, but also to often misdiagnose conditions or misinterpret data: honest, but incorrect. Equally, another doctor might always get diagnosis and treatment right, but often misses appointments, falls asleep on duty, or misuses patient data for their own ends: (medically) correct, but unreliable and dishonest. Whilst we might easily say that trusting either doctor could be harmful (neither individual is worthy of trust), this is in two quite different senses. For O'Neill, this is where the different 'directions of fit' allows for a more nuanced understanding of trust and trustworthiness.

'Trust' can be disaggregated into three interrelated elements, reflecting two key philosophical 'directions of fit':

1) Trust in others' *truth claims* (in the sense that they are likely to be correct),
2) Trust in others' *commitments* to do what they say they will do (in the sense that we might trust our bank to send us a statement every month if they say that they will),
3) Trust in others' *competence* to meet those commitments (in the sense that we trust our dentist's competence to remove a tooth properly).

The first of these addresses an *empirical* 'direction of fit' (does the claim 'fit' the world as it is?), and the other two a *normative* element (does the action 'fit' the relevant norm, for example, commitment, reliability, honesty, competence?). As O'Neill explains, trust in *future behaviour* bridges both empirical and normative directions of fit; that is, some truth claims may be assumed to be honest or accurate, but—to establish trust—greater emphasis is placed on judgments made about the other party's commitment and competence.[36] Hence, in order to be trustworthy, O'Neill argues that one must be 'trustworthy both in word and in deed, both in [empirical] truth claims and in [normative] action.'[37] For example, making a correct diagnosis (a *truth claim*), showing up to the pre-arranged appointment (fulfilling that *commitment*) and carrying out treatment, such as an injection, at that appointment with *competence*.

We have previously outlined the benefits of such an approach for *developers* to utilise the insights provided by O'Neill's tripartite framework to help make clinicians feel more confident about using CDSSs.[38] In this article, we use O'Neill's framework to enhance our

[33] O'Neill (n 20) 293; see also Hardin (n 26).
[34] Select Committee on Artificial Intelligence (n 10) 50.
[35] O'Neill (n 20).
[36] ibid.
[37] ibid.
[38] Jones, Thornton and Wyatt (n 7).

understanding of the legal and regulatory regime (both doctrinally and empirically) in relation to the use of AI and CDSSs, given the trust concerns outlined by clinicians. O'Neill's framework is particularly helpful as it allows us to consider and clarify what is meant by 'trust' and 'trustworthiness' in this context in a multi-dimensional way.

III. CLINICIANS' PERSPECTIVES

Over the last decade, several empirical studies have explored healthcare professionals' perspectives on the use of AI and CDSSs. These studies have included the views of general practitioners (GPs) in Belgium[39] and the UK[40]; in emergency care in the UK[41]; in hospitals in Italy,[42] Belgium,[43] and China[44]; with stakeholders in France,[45] the UK,[46] Australia and New Zealand,[47] and globally.[48] In this section, we focus on studies that identified clinicians' issues around trust, in particular matters of control, medical errors, and legal responsibility/liability.

A. Control

In a qualitative mixed methods study undertaken by Van Cauwenberge and others, case vignettes were used with 24 physicians in Belgium.[49] Although their study focused on the *implementation* of CDSSs rather than trust per se, they identified overarching themes around the 'perceived role' of the AI and physicians respectively. The importance of clinicians having control, expressed as the 'final responsibility' (to make clinical decisions) was emphasised, with concerns raised over the potential automation of key aspects of clinicians' roles regarding diagnosis and treatment. Hence, Van Cauwenberge and others found that administrative tasks would be more readily handed over to AI than medical decisions, as indicated by this quote from a participant: '[Unlike with medical decisions] I do trust the AI when it takes administrative decisions. Those do not look difficult to me.'[50].

In this study, clinicians' reasons for the preservation of control 'differed widely'.[51] In part, they were expressed as being about *empirical* trust. That is, the clinician should have the final say because they know the facts, science, patient context, and their own clinical skill set better than the AI (which is dependent on the quality of the data/information provided to it). We return to the importance of patient context below. However, control concerns were also articulated in terms of norms of professional practice and autonomy, that is trust in others'

[39] Annemie Heselmans and others, 'Family Physicians' Perceptions and Use of Electronic Clinical Decision Support During the First Year of Implementation' 2012 36(6) Journal of Medical Systems 3677.

[40] Charlotte Blease and others, 'Artificial Intelligence and the Future of Primary Care: Exploratory Qualitative Study of UK General Practitioners' Views' 2019 21(3) Journal of Medical Internet Research e12802.

[41] Catherine Pope and others 'Using Computer Decision Support Systems in NHS Emergency and Urgent Care: Ethnographic Study Using Normalisation Process Theory' 2013 13(1) BMC Health Service Research 111.

[42] Liberati and others (n 18).

[43] Daan Van Cauwenberge and others, '"Many Roads Lead to Rome and the Artificial Intelligence Only Shows Me One Road": An Interview Study on Physician Attitudes Regarding the Implementation of Computerised Clinical Decision Support Systems' 2022 23(2) BMC Medical Ethics 50.

[44] Wenjuan Fan and others 'Investigating the Impacting Factors for the Healthcare Professionals to Adopt Artificial Intelligence-based Medical Diagnosis Support System (AIMDSS)' 2020 294(1–2) Annals of Operations Research 567.

[45] Lai, Brian and Mamzer (n 18).

[46] Petkus, Hoogewerf, and Wyatt (n 18).

[47] Jane Scheetz and others 'A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology' 2021 11(1) Scientific Reports 5193.

[48] For example, Sarwar and others cite 487 respondents from 59 countries, whereas Chen et al claim participation from respondents in 39 countries. See Sarwar and others, 'Physician Perspectives on Integration of Artificial Intelligence into Diagnostic Pathology' (2019) 2(1) NPJ Digital Medicine 28; Mingyang Chen and others, 'Acceptance of Clinical Artificial Intelligence among Physicians and Medical Students: A Systematic Review with Cross-Sectional Survey' (2022) 9 Frontiers in Medicine 1.

[49] Van Cauwenberge and others (n 43).

[50] ibid 4.

[51] Van Cauwenberge and others (n 43) 6.

competence: AI was compared with a clinical colleague, whereby the AI could not explain its reasoning for making a suggestion, whereas the colleague could explain their rationale and respond to alternative views from other clinicians. Interestingly, despite the emphasis on 'final responsibility', none of the participants directly referred to legal issues—although this matter has arisen in other studies.

B. Linking control to medical errors and liability

Liberati and others[52] collected data at four Italian hospitals, using semi-structured interviews with a sample of doctors, nurses and others (eg, IT staff), and noted the potential for legal vulnerability arising from the perceived loss of control over decision making. These fears were 'particularly acute' in relation to legal controversies. Participants felt forced to follow CDSS advice (at the expense of contextualised decision making) to avoid becoming 'legally vulnerable' in the event of something going wrong.[53]

The concern expressed here is that the nuances of contextualised medical decision making could be silenced or overridden by external (non-medical) professionals' reliance on the CDSS advice if a medical–legal event arose. For example, one physician was concerned about decisions not to use CDSS being 'used against' them in court, even where doing so would have made no difference to the patient's outcome. IT staff similarly reported being asked by worried physicians about what would happen if they diverged from the CDSSs' recommendation (and struggled to answer, as the legal framework is unclear).[54]

The 'failure' to utilise AI/CDSS tools when available might require justification, once these tools are normalised in medical practice (which is not yet the current state of play).[55] Furthermore, the lack of clarity over the relevant legal frameworks is a concern not only for physicians, but also those health-related colleagues, for example, pharmacy or quality improvement staff, who may be consulted about advice on their use. Similarly, these concerns appear to be primarily about *empirical* trust. Respondents in this study were concerned about AI being given undue primacy in establishing what the empirically 'best' care in the circumstances might be (as it may miss 'other aspects' of the facts),[56] rather than whether the AI can reach a normative standard (of, say, reasonableness, or even 100% accuracy). Equally though, like the above discussion of 'control', there can also be normative trust issues here too. Some AI/CDSS recommendations will be difficult to categorise as empirically 'correct' or 'incorrect' and, rather, reflect an inevitably value-laden judgement on, for example, best interests. For example, suppose the recommendation produced by the AI/CDSS is said to 'reflect the relevant guidance from NICE for this condition'. There is both an empirical element (does it accord with NICE guidance?) and a normative element (does that NICE guidance and the medical expertise it was based upon meet a given standard of competence?) to the question of whether this recommendation is trustworthy.

C. Medical errors and liability

Similar themes have emerged in UK-focused studies. In an early study, focused on the failed introduction of the PRODIGY programme into the NHS, Rousseau and others[57] found a 'strong theme' about the lack of helpfulness of the system, and concerns about trust. For example, one clinician said they 'don't trust' practicing medicine in that way and were concerned about having to defend themselves in court, tribunals, etc. in 'a trial of computer

[52] Liberati and others (n 18).
[53] ibid 6.
[54] ibid.
[55] But see the discussion of how a specific CDSS was normalised in Pope and others (n 41).
[56] Liberati and others (n 18) 6.
[57] Rousseau and others (n 18).

guidelines.'[58] In their qualitative web-based survey of UK GPs' views on AI and primary care, Blease and others[59] highlighted one free text comment on liability: 'the issue is responsibility and liability in legal terms for such tools.'[60]

Interestingly, Blease and others noted that AI researchers were more likely than clinicians to raise a wide range of issues about the design and use of AI tools, including machine learning algorithms (eg biases, reliability, transparency, regulation, privacy, and security).[61] Meanwhile, [anonymised for submission], in their survey of representatives of medical specialty societies in the UK, noted that '[c]oncerns about professional practice, ethics and liability included that the legal liability of doctors following advice is unclear'; this concern ranked highest in importance, 17.5 out of a possible score of 18, suggesting that clarity over the liability regime would be beneficial.[62]

Similarly, in Lai and others[63] qualitative study, in which they interviewed 40 healthcare professionals and other stakeholders in France—the health professionals and health industry representatives noted questions of liability/responsibility as a potential obstacle to the uptake of AI. Thus, physicians 'were not prepared to be held criminally responsible . . . [for errors] . . . made by an AI tool'.[64] Tortious liability was not specifically flagged up regarding clinicians' responsibility, though Lai and others did report the concerns of healthcare industrial partners, citing the potential to inhibit the development of AI tools in France if industry partners were to be held partially responsible for injuries arising from reliance on advice from AI tools.[65] However, the regulators were keenly aware that for uptake of these tools to occur 'it will be necessary for them [healthcare professionals] to be able to trust the assessment process, as in the past.'[66]

Again, in one sense, these trust concerns are empirical in nature. Clinician respondents in these studies speak of their trust concerns in terms of trial by (empirically incorrect, in their patient's context) computer guidelines[67] (instead of being tried against what the empirically or scientifically justified approach was in the circumstances), or capability for 'medical error'[68] which they then get blamed for. For example, the AI/CDSS may suggest a depressed patient be prescribed tricyclic antidepressants, due to that person's previous positive responses in their patient record. That may be sensible advice when based solely upon the information in the system, but may be negligent to follow if the patient presented with suicidal ideation that day. Deliberate overdose would then be a highly relevant risk in that context, which the AI/CDSS advice would not have taken account of. In that example, the AI/CDSS has given a recommendation ('prescribe tricyclic antidepressants'), which is empirically incorrect for that clinician's patient.[69] The concerns of this nature from clinicians in terms of empirical accuracy in the circumstances are to be distinguished from the more normative trust concerns identified by O'Neill and, we later argue, presented by non-clinician stakeholders. On the other hand, it is important to acknowledge that there is still a potential normative element to this sort of recommendation too. Even if the patient had not presented with suicidal

[58] ibid 4.
[59] Blease and others (n 40).
[60] ibid 5.
[61] ibid.
[62] [anonymised] (n 18).
[63] Lai, Brian and Mamzer (n 18).
[64] ibid 6.
[65] ibid.
[66] ibid 7.
[67] Rousseau and others (n 18).
[68] Liberati and others (n 18) 6.
[69] See Garry W Kerr, AC McGuffie and S Wilkie, 'Tricyclic Antidepressant Overdose: A Review' (2001) 18 Emergency Medicine Journal 236; National Institute for Health and Care Excellence, 'Depression: Prescribing Information' (September 2022) <https://cks.nice.org.uk/topics/depression/prescribing-information/> accessed 11 November 2022.

ideation that day, it still requires the clinician to trust the *competence* of the medical expertise that the CDSS recommendation is based upon (here, in relation to the effectiveness of antidepressants more generally).

D. Summary

As outlined in this section, empirical research on clinicians' perspectives on using AI and CDSSs suggests that although the potential benefits for improving patient care and outcomes are well recognised, concerns remain with regard to the legal issues[70]—especially liability—that may arise in connection with their use.[71] What if the clinician trusts and relies upon the AI/CDSS advice and this causes harm to a patient?[72] The clinician must have confidence that the decision support tool will do what it is supposed to and not cause harm for which they may be legally responsible. For clinicians then, trustworthy AI/CDSS should be an accurate tool, whose design, engineering and operation ensures they generate positive outcomes, and mitigates potentially harmful ones. However, the perceived uncertainty about the potential legal consequences does little to facilitate the establishment of trust and confidence in their use.[73] Indeed, although beyond this article's focus on trust of clinicians *in AI/CDSS*, there are further questions to be asked about trust in *the law* and legal processes. Many of the above studies demonstrated a highly cynical view of clinicians' prospects of being judged fairly.

Hence, from these empirical studies of clinicians' views, it appears that there is a strong 'trust' concern in the sense of 'empirical trust' (O'Neill's first limb—ie is the AI/CDSS's recommendation empirically correct?). If AI/CDSS advice is considered trustworthy when it is accurate, then it suggests that this is not a case of decision support tools being untrusted (or untrustworthy) in the normative sense. Clinicians appear to trust that the AI will not be dishonest, etc—these tools can be depended upon to work (provided they are plugged in, installed correctly)—though concerns might arise regarding algorithmic bias. Similarly, the question of competence (of reaching a normative standard, eg of a reasonably qualified peer) does not appear to be the primary concern of these clinicians in relation to AI/CDSS[74]; especially not in relation to matters of 'fact' (such as the size of a tumour), as opposed to judgments about what ought to be done about those facts (treatment options). Questions about competence may, however, arise in relation to the initial medical expertise/value-judgments the AI/CDSS is trained upon, but that is a concern about medical guidelines generally, rather than AI/CDSSs per se. In spite of grand claims by developers that their AI/CDSS is as good as or better than a human clinician, clinicians' own concerns in these empirical studies are often on a different conceptual level, pointing in a different (empirical) 'direction of fit'.

[70] 'Lawfulness' is highlighted in a recent literature review. See Victoria Tucci, Joan Saary and Thomas E Doyle, 'Factors Influencing Trust in Medical Artificial Intelligence for Healthcare Professionals: A Narrative Review' 2022 5 Journal of Medical Artificial Intelligence 4.

[71] Liberati and others (n 18); Derk L and others, 'Acceptance and Barriers Pertaining to A General Practice Decision Support System for Multiple Clinical Conditions: A Mixed Methods Evaluation' 2018 13(4) PLoS One e0193187; Petkus, Hoogewerf, and Wyatt (n 18).

[72] On routes to trust and black-box systems see Robin Feldman, Ehrik Aldana and Kara Stein, 'Artificial Intelligence in the Health Care Space: How We Can Trust What We Cannot Know' 2019 30 Stanford Law and Policy Review 399; on issues regarding opaque AI systems see Helen Smith, 'Clinical AI: Opacity, Accountability, Responsibility and Liability' 2021 36(2) AI and Society 535.

[73] NHS AI LAB & Health Education England, 'Understanding Healthcare Workers' Confidence in AI' (Report 1, May 2022) <https://digital-transformation.hee.nhs.uk/binaries/content/assets/digital-transformation/dart-ed/understandingconfidenceinai-may22.pdf> accessed 14 November 2022 45. This can be contrasted with the consideration of liability for automated vehicles. The Law Commission began its review in 2018, culminating in January 2022 with a joint report published together with the Scottish Law Commission, recommending law reform building on the Automated and Electric Vehicles Act 2018. See Law Commission 'Legal Reforms to Allow Safe Introduction of Automated Vehicles Announced' (26 January 2022) <www.lawcom.gov.uk/legal-reforms-to-allow-safe-introduction-of-automated-vehicles-announced/> accessed 14 November 2022.

[74] Although the inability to interrogate the process of generating the AI/CDSS's recommendation was noted in Van Cauwenberge and others (n 43).

IV. CLINICIANS AND LIABILITY MATTERS

In some ways, clinicians are right to be concerned about being blamed. As a starting point in principle, it has been argued that it is conceptually wrong to apply the law to 'technology'; rather, it is the human(s) who decide to use that technology that should bear the risk of liability.[75] However, such a default position is not necessarily desirable or fair. It would be undesirable if such an approach led to reluctance to use useful technology for fear of liability (as the previous section suggests it currently does). It seems unfairly absolutist to *always* lay the blame with the *user*, as opposed to for example the developer, or the hospital/trust that procured, and thus potentially imposed that system on the user (though, as we note below, there are potential liability routes for these parties too). Equally, it is naïve to treat self-learning 'intelligent' tools in the same way as a scalpel or X-ray machine.

The Academy of Medical Royal Colleges has queried the 'medico-legal position for a clinician who disagrees with the AI'[76]; and it has also suggested that 'the nature of negligence claims may change as patients adapt to the availability of AI-generated decisions and recommendations', which in turn may have implications for medical defence organisations.[77] Unsurprisingly, proposals to address liability in this field have been made.[78] However, at the time of writing, we could not find any reported legal decisions that turned on the use of AI or CDSS advice. Searches were conducted on Lexis Library, Westlaw, BAILII, and PubMed for phrases around AI and CDSS. The following search terms were used: 'adviser', 'algorithm', 'automated tool', 'decision support', 'digital technology', 'expert system', 'flowchart risk score'—the results were narrowed by 'health', 'liability', 'medicine', 'medical', and 'tort', but did not generate any relevant reported legal decisions; nor have other researchers been able to locate published or reported decisions in the UK, Europe, or USA.[79]

This raises the question of what if the clinician trusts and relies upon the AI/CDSS advice and this causes harm to a patient? Liability may arise under a variety of heads against a range of possible defendants. For example, clinicians may be liable in negligence; the NHS may be liable directly—through a breach of statutory duty, or safe systems of care, or if a non-delegable duty was established[80]—or it may be held vicariously liable for the actions or omissions of negligent clinicians; and developers might be held liable under product liability legislation or in negligence, or face enforcement powers under consumer protection legislation.[81]

A. Clinicians' competence

It is well established that clinicians will remain legally responsible for the medical advice and treatment given to their patients, irrespective of the use of AI/CDSS.[82] The same principles

[75] Chris Reed cited in Select Committee on Artificial Intelligence, 'Corrected Oral Evidence: Artificial Intelligence' (*House of Lords*, 17 October 2017) <https://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-comittee/artificial-intelligence/oral/71898.html> accessed 14 November 2022 Q31; Commission Expert Group on Liability and New Technologies—New Technologies Formation, 'Liability For Artificial Intelligence and Other Emerging Digital Technologies' (2019) <https://op.europa.eu/en/publication-detail/-/publication/1c5e30be-1197-11ea-8c1f-01aa75ed71a1/language-en> accessed 14 November 2022, 6, para [8].

[76] Academy of Medical Royal Colleges, 'Artificial Intelligence in Healthcare' (January 2019) <www.aomrc.org.uk/wp-content/uploads/2019/01/Artificial_intelligence_in_healthcare_0119.pdf> accessed 14 November 2022, 32.

[77] ibid 28.

[78] Smith and Fotheringham (n 7); Helen Smith and Kit Fotheringham, 'Exploring Remedies for Defective Artificial Intelligence Aids in Clinical Decision-making in Post-Brexit England and Wales' (2022) 22(1) Medical Law International 33; for the Australian legal context see Prictor (n 7).

[79] John Fox and Richard Thomson, 'Clinical Decision Support Systems: A Discussion of Quality, Safety and Legal Liability Issues' 2002 Proceedings of the AMIA Symposium 265; Smith (n 72).

[80] Paula Giliker, 'Non-delegable Duties and Institutional Liability for the Negligence of Hospital Staff: Fair, Just and Reasonable?' (2017) 33(2) Tottel's Journal of Professional Negligence 109. Since publication there have been cases on PCT liability: *JMH v Akramy; Badger Group and NHS Commissioning Board* [2020] EWHC 3445 (QB), also reported as *Hopkins (A Child) v Akramy* [2020] EWHC 3445 (QB); and *Hughes v Rattan* [2022] EWCA Civ 107.

[81] See Prictor (n 7).

[82] [anonymised for submission] (n 7); Smith and Fotheringham (n 7).

apply to doctors, 'nurses, midwives, dentists and opticians', that is, '[a]ny person who professes expertise in any aspect of medical treatment is required to exercise reasonable skills and care, judged by the standards of his [or her] own particular profession.'[83] An action in negligence could potentially arise, therefore, if a patient was harmed following a clinician's reliance on faulty AI/CDSS advice, or where an inappropriate CDSS was chosen, without reflection or application of their own expert knowledge to the situation. Although clinicians (as with other professionals) are expected to remain up to date, they are not expected to be aware of every single development in their field, but rather to keep abreast of 'common practice'.[84] Thus, the extent to which a clinician should be aware of a specific issue will turn on the facts and established practice at the time. This approach is subject to the requirement to disclose to the patient 'any material risks in any recommended treatment, and of any reasonable alternative or variant treatments'.[85] Materiality of risk is assessed by reference to what 'a reasonable person in the patient's position' would likely attach significance to, or by a determination of what the doctor did or should reasonably have known regarding significance of risk for their 'particular patient'.[86]

The Medicines and Healthcare Products Regulatory Agency (MHRA) has noted that clinicians may increasingly rely on the outputs of AI/CDSS without accessing or reviewing the raw data.[87] Hence, some caution over the (possibly blithe) acceptance of AI/CDSS advice in medical practice would be beneficial to minimise the risks associated with 'automation bias' (the human tendency to follow advice from a computer system, even when the computer is incorrect and the human would have made the correct decision), especially regarding a failure to disclose material risks regarding treatment, and/or options.[88] The alternative—where the clinician either refuses or fails to use the AI/CDSS at all (especially where this *is* accepted as common practice), or refutes or overrides the AI/CDSS advice with no good reason—will raise questions as to the defensibility of their approach and standpoint.[89]

Interestingly, in theory, the legal test applied to clinicians therefore appears to focus much more on a normative basis of trust (versus the trust clinicians themselves assess AI/CDSSs by). Applying first principles, whether the clinician is factually/empirically correct or not (and therefore trustworthy in the empirical sense), is ultimately not the issue in negligence liability. The focus in finding a breach of duty of care is ultimately about whether they have fallen below the standard of the reasonable clinician in the circumstances. Assessment against a *normative* standard therefore suggests that, in terms of trust in the CDSS-using clinician themselves, the issue is of a fundamentally different type of trust than that which the clinician might themselves judge the AI/CDSSs by. This is not a purely semantic issue. Indeed, as we shall see in relation to analogous cases, the difference in standards matters very much in certain contexts and may explain why clinicians have such fears over legal liability.

[83] John Powell and others, 'Medical Practitioners' in Powell and others (eds), *Jackson and Powell on Professional Liability* (8th edn, Sweet and Maxwell 2017) 13-044; drawing on *Bolam v Friern Hospital Management Committee* [1957] 1 WLR 58 and *Bolitho v City of Hackney Health Authority* [1998] AC 232.

[84] David I Bainbridge, 'Computer-aided Diagnosis and Negligence' (1991) 31(2) Medicine, Science and the Law 127. See eg, *Bayley v George Eliot Hospital NHS Trust* [2017] EWHC 3398 (QB) where the court was 'not satisfied that a reasonably competent vascular surgeon would or ought to have known about' an alternative treatment, and therefore dismissed a claim that the physicians were negligent in not informing the claimant about it [99].

[85] *Montgomery v Lanarkshire Health Board* [2015] UKSC 11, [2015] AC 1430.

[86] ibid.

[87] Medicines and Healthcare Products Regulatory Agency, 'Guidance: Medical Device Stand-alone Software' (2017) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/648465/Software_flow_chart_Ed_1-04.pdf> accessed 28 May 2018.

[88] Kate Goddard K, Andul Roudsari, and Jeremy C Wyatt, 'Automation Bias: A systematic Review of Frequency, Effect Mediators, and Mitigators' (2012) 19(1) Journal of the American Medical Informatics Association 121.

[89] Academy of Medical Royal Colleges (n 76).

B. Analogous reported cases

Reported examples where clinicians and healthcare professionals have fallen short of the reasonably competent standard in relation to the use (or failure to use) of other medical technologies include: a doctor whose interpretation of a cardiotocogram was found wanting[90]; failures by midwives regarding the use of ultrasound and cardiotocogram in the monitoring of foetal well-being during childbirth[91]; paramedics' failure to advise a patient to attend hospital, when two ECGs appeared normal to them despite a computer-generated report on the face of one printed ECG indicating an abnormality[92]; and a trainee paediatrician convicted of gross negligence manslaughter following the death of a child patient.[93]

In each of these cases, the professional conduct of the clinician in question was assessed according to the reasonable competent standard of a person specialised in their field,[94] having scrutinised the professionals' evidence to ensure it withstood logical analysis.[95] Although it is rare for evidence to be rejected on the basis of the common practice being illogical, it is not unheard of.[96] In such circumstances, 'logic' may be understood as a different normative standard that is being compared against, and the judiciary has shifted its comparator of trustworthiness (because the profession's 'common practice' is no longer deemed worthy of trust) towards a normative standard of 'logic'. Accordingly, the 'norm' shifts from *the profession's* 'common practice', to *the court's* view of what is logically defensible practice.

More radically, sometimes the direction of fit for understanding trust and assessing trustworthiness (between normative and empirical trust) changes. In *Muller*, in obiter remarks, a distinction was drawn between 'pure diagnosis' and 'pure treatment' cases (negligence was ultimately found here on the *Bolitho* basis of the expert evidence of 'common practice' not withstanding logical analysis).[97] Regarding 'pure diagnosis', Jackson has suggested that:

> [T]there cannot be two right answers to the question of how a patient should be diagnosed, as might be the case in a "negligent treatment" case. Rather the diagnosis is simply wrong, and an expert witness who claims that a pathologist would have acted competently by missing obvious signs of melanoma was not expressing a defensible opinion.[98]

This reflects Kerr J's view:

> In a case involving advice, treatment or both, opposed expert opinions may in a sense both be 'right', in that each represents a respectable body of professional opinion. The same is not true of a pure diagnosis case such as the present, where there is no weighing of risks

[90] *Robertson v Nottingham HA* [1997] 8 Med LR 1.

[91] *Popple v Birmingham Women's Hospital NHS Foundation Trust* [2011] EWHC 2320 QB, upheld on appeal [2013] EWCA Civ 1628.

[92] *Taaffe v East of England Ambulance Service NHS Trust* [2012] EWHC 1335 QB.

[93] *R v Bawa-Garba* [2016] EWCA Crim 1841, *Bawa-Garba v GMC* [2018] EWCA Civ 1879. In the Court of Appeal, reference was made to the systemic failings at the hospital at the time of the patient's death, including the communication of blood test results. The NHS trust did not face any legal action for these failings, and although reference was made to its investigation report on the failings, no details were provided in any of the reported decisions. However, one of the contributing factors to the 'failure to act on abnormal test results' was the fact that just three minutes after the patient's full screen of routine blood and other tests had been sent to the laboratory for processing, the iLab system used to report the results failed. The benefit of the iLab system is that it would flag any abnormal results. When the laboratory staff provided a verbal report, the medical personnel were 'not informed that any of the results were abnormal'; and it was established that the personnel in question would have relied upon the iLab system to highlight potential issues.

[94] *Bolam v Friern Hospital* [1957] 1 WLR 582.

[95] *Bolitho v City of Hackney Health Authority* [1998] AC 232.

[96] *Marriott v West Midlands RHA* [1999] Lloyd's Rep Med 23; *C v North Cumbria University Hospitals NHS Trust* [2014] EWHC 61 (QB), [2014] Med. L.R. 189; *Lane v Worcestershire Acute Hospitals NHS Trust* [2017] EWHC 1900 (QB); *Muller v King's College Hospital NHS Foundation Trust* [2017] EWHC 128 (QB); *Bradfield-Kay v Cope* [2020] EWHC 1351 (QB).

[97] *Muller v King's College Hospital NHS Foundation Trust* [2017] EWHC 128 (QB).

[98] Emily Jackson, *Medical Law Text Cases and Materials* (5th edn, OUP 2019), 135.

> and benefits, only misreporting which may or may not be negligent. The experts expressing opposing views on that issue cannot both be right. And the issue is, par excellence a matter for the decision of the court, which should not, as a matter of constitutional propriety, be delegated to the experts.[99]

Hence, in 'pure diagnosis' cases, Kerr J appears to reject a purely normative assessment of trustworthiness, instead also considering the question of which view is ultimately empirically right (and accepting that there is only one that can be). This incorporates a question of *empirical* correctness (O'Neill's first kind of trust): the issue is not whether conduct matches a given normative standard of *competence*, rather it is either (*empirically*) correct or incorrect. Cases turning on 'advice, treatment or both', however, return to the normative direction of fit: the court is simply asking whether the conduct 'fits' the standard of 'a respectable body of professional opinion'.

Thus, the *context* of the use of the AI/CDSS in question may determine the appropriate standard of care applicable. If the clinician uses, or elects not to use, AI/CDSS in *diagnosis*, then the determination of the court in *Muller* would suggest that the diagnosis issue is binary—it is either correct or not. Accordingly, the competence of the clinician is either made out, or not, and the *Bolam* test will not 'save' the clinician from liability. However, where the AI/CDSS is used in judging *treatment* options then it is feasible, per *Bolam*, *Bolitho*, and *Muller*, that a reasonable competent clinician might reach a different conclusion as to the appropriate treatment pathway.

C. Professional matters

Three further interrelated issues arise. The first concerns the matter of providing effective training for doctors to use digital technologies; noted in 2018 by Andrew Goddard, (then)President of the Royal College of Physicians,[100] and echoed more recently by the WHO[101] and NHS AI Lab and Health Education England.[102] There is an important question of fairness if clinicians are held responsible for consequences arising from misuse of AI/CDSS when they have never been given the opportunity to learn how to use it properly: they would be set up to fail.

Secondly, there is the (arguably undue) significance of professional guidelines (let alone AI/CDSS which incorporates and purports to apply such guidelines) assisting courts in establishing the appropriate standard of reasonable care. Whilst, as Jackson notes, '[g]uidelines are, by definition, not mandatory',[103] Samanta and others[104] empirical study on the use of clinical guidelines in medical litigation shows they can strongly influence decisions about settling (or abandoning) a claim at an early stage in the litigation process. Further, '[a]lthough not dispositive, CGs [clinical guidelines] can be persuasive or influential upon judicial decision-making'[105]; although the difficulties caused by 'forensic dissection',[106] or the 'wisdom of extracting words from such guidance out of context as if they were legal

[99] *Muller* (n 97) [75] per Kerr J.
[100] 'Safety and Regulation of Digital Technologies' (*Royal College of Physicians*, 25 July 2018) <https://www.rcplondon.ac.uk/news/safety-and-regulation-digital-technologies> accessed 14 November 2022.
[101] World Health Organization, 'Ethics and Governance of Artificial Intelligence for Health: WHO Guidance' (2021)/ <https://www.who.int/publications/i/item/9789240029200> accessed 14 November 2022.
[102] NHS England, 'Developing Healthcare Workers' Confidence in AI' (October 2022) <https://digital-transformation.hee.nhs.uk/binaries/content/assets/digital-transformation/dart-ed/developingconfidenceinai-oct2022.pdf> accessed 14 November 2022, 7.
[103] Jackson (n 98)136.
[104] Ash Samanta, Jo Samanta and Joanne Beswick, 'Responsible Practice or Restricted Practice? An Empirical Study of the Use of Clinical Guidelines in Medical Negligence Litigation' (2021) 29(2) Medical Law Review 205, 229.
[105] ibid.
[106] ibid 222.

instruments' has been vehemently critiqued.[107] Thus, it remains important to be cognisant of Montgomery and Montgomery's concerns about the 'unpredictability' of the judicial interpretation and application of clinical guidelines in determining liability in cases which do proceed to court.[108] There is a danger that similar deference may be applied to an AI/CDSS's recommendations when based on such guidelines. Again, emphasis on guidelines is very much considering the issues from a normative trustworthiness angle. The clinician is being judged against a normative standard of competence rather than an empirical discussion as to whether what they did was correct or not.

Thirdly, there must be robust procurement processes in place to ensure appropriate AI/CDSS tools are sourced for clinical use. NHS Trusts and primary care practices directly procure computer systems such as AI and CDSS for use by their staff. They also procure external services that may use AI or CDSS, such as a radiology or lab test reporting service, or a remote monitoring service that automatically detects risky events in patients at home from measured data, or an online primary care triage or symptom-checking service to advise patients on their best action given their symptoms. If any of these systems fail this could cause patients harm, so clinicians need to be able to trust that their employer has used an appropriate procurement method to purchase a high-quality system or service. IBM Watson Health (since rebranded Merative) and Google's DeepMind are two high profile brands that have emerged in the context of AI and health applications, although neither are without criticism[109] or controversy.[110] Thus, reliance on the brand alone to facilitate trust will likely be insufficient.[111] Further, where clinicians have concerns about the quality of AI or CDSS provision their professional code of conduct requires them to remove patients from risks (from such tools or equipment) and to report incidents that risk patient safety.[112] Again, from clinicians' perspective, the concern is not much that these companies have nefarious motives or are going to provide software that is full of bugs (which raises normative trust questions of commitments and competence), but rather whether they can have confidence/trust the programme to provide an empirically/medically 'correct' answer.

Squaring the circle, and returning to Samanta and others work, they reported that participants in their study expressed concern that shortages of resources should not be a 'blanket defence for decisions not to follow relevant guidance.'[113] Further, liability in negligence should lie with organisations (not individuals) where non-compliance with guidance was due to resource limits; citing the dissent of Sir Nicolas Browne-Wilkinson (then)V-C: 'A health authority which so conducts its hospital that it fails to provide doctors of sufficient skill and experience to give the treatment offered at the hospital may be directly liable in negligence to the patient.'[114]

[107] Jonathan Montgomery and Elsa Montgomery, 'Montgomery on Informed Consent: An Inexpert Decision?' (2016) 42(2) Journal of Medical Ethics 89, 89.
[108] ibid 93.
[109] On IBM Watson Health see Smith and Fotheringham (n 7) citing Casey Ross and Ike Swetlitz, 'IBM Pitched Watson as a Revolution in Cancer Care. It's Nowhere Close' (*STAT News*, 2017), <https://www.statnews.com/2017/09/05/watson-ibm cancer/> accessed 17 September 2022.
[110] Following an investigation by the Information Commissioner into DeepMind's Streams app, see The Royal Free London NHS Foundation Trust, 'Audit of the Acute Kidney Injury Detection System Known as Streams' (17 May 2018) <http://s3-eu-west-1.amazonaws.com/files.royalfree.nhs.uk/Reporting/Streams_Report.pdf> accessed 14 November 2022. In May 2022, Mishcon de Reya announced it was acting on behalf of individuals in a representative action in the High Court against Google and DeepMind for the unlawful use of patients' confidential medical records with the Streams app.
[111] Jones, Thornton and Wyatt (n 7).
[112] General Medical Council, 'Good Medical Practice' (2019); also see Smith and Fotheringham (n 78).
[113] Samanta, Samanta and Beswick (n 104) 228.
[114] ibid 228 citing *Wilsher v Essex AHA*, [1987] Q.B. 730 (1986) per Browne-Wilkinson 778 A-C.

D. Summary

Hence, whilst clinicians in published empirical studies (Section III) tended to mention numerous issues about AI/CDSS, many of these concern questions of empirical trustworthiness: will the AI provide me with the correct answer about this patient? Similarly, their concerns around legal liability are often framed in terms of the AI/CDSS getting something empirically wrong (and the clinician themselves getting the blame). In contrast (with some exceptions discussed earlier), the courts themselves will often apply O'Neill's third kind of trust: normative competence. The implications for there sometimes being such dissonance in standards (or perceptions thereof) will be further explored in the next section, on trust by patients in their clinicians.

V. PATIENTS' PERSPECTIVES

It is fair to say that coverage of patient/service-user views on their clinicians' use of AI and/or CDSS is often inadequate. Their perspectives tend to be either missing entirely, implicit, or assumed to be homogenous. For example, the House of Lords Report[115] mentioned patients in many areas, but did not include or refer to any commentary from patients or representative groups. Notable exceptions (principally in the mental health context) include Davies and others (2017) discussion of service-users' perspectives on a mental health app[116]; Hill and others (2017) on the importance of collaborative design and development involving service-users and clinicians[117]; and Hollis and others emphasis on the importance of service-users' needs and priorities to be driving development in digital technology mental health support (not least given concerns over the safety and efficacy of digital interventions versus face-to-face engagement and care).[118]

This research gap is clearly unfortunate. As Laï and others point out, AI could jeopardise the physician–patient relationship, for example 'the "black box" phenomenon could prevent the doctor from providing clear information' to patients.[119] Hence, even though this article is primarily concerned with *clinicians'* perspectives on trust; nonetheless, we think it is important and helpful to consider patient/service-user and more widely members of the publics' perspectives regarding their clinicians' use (or indeed non-use) of AI/CDSS.[120] Contrasting these with clinicians' own concerns can be illuminating in this regard. Indeed, we argue that the (admittedly limited) available evidence suggests service-users and clinicians may be talking past one another in this context. When one uses O'Neill's trust lens, patient concerns in relation to trust and trustworthiness tend to be of a different kind (or 'direction of fit') to those of clinicians.

Several empirical studies indicate how the legal principles outlined in the previous sections might operate in cases involving clinician use of CDSS. These studies examine attitudes of members of the public, acting out the role of patients and juries, towards the use of AI/CDSS by doctors. Whilst such experiments are obviously limited in that they do not consider *real* patients involved in real cases/treatment (or AI or CDSSs which use AI), they still reveal an interesting trust dynamic. Such studies suggest that patients, like clinicians, are concerned

[115] Select Committee on Artificial Intelligence (n 10).
[116] E Bethan Davies and others, 'Proportionate Methods for Evaluating a Simple Digital Mental Health Tool' (2017) 20(4) Evidence-Based Mental Health 112.
[117] Claire Hill and others, 'Navigating the Challenges of Digital Health Innovation: Considerations and Solutions in Developing Online and Smartphone-Application-Based Interventions for Mental Health Disorders' (2017) 211(2) British Journal of Psychiatry 65.
[118] Chris Hollis and others, 'Identifying Research Priorities for Digital Technology in Mental Health Care: Results of the James Lind Alliance Priority Setting Partnership' (2018) 5(10) Lancet Psychiatry 845.
[119] Lai, Brian and Mamzer (n 18) 4.
[120] Maria Sheppard has considered trust in apps used by patients directly, see Sheppard (n 23).

with whether CDSSs get things right ('empirical' trust—correctness of decision). However, they also show that, unlike clinicians, even when a CDSS gets it wrong (and therefore any 'empirical' trust has broken down), study participants still rate the clinician using the AI/CDSS *significantly less* harshly than when the clinician gets it wrong without using AI/CDSS. How can this be? We argue that O'Neill's trust lens helps to make some sense of this puzzle, by identifying some enduring trust in the AI/CDSS by hypothetical patients that is not about accuracy ('empirical' trust) but concerns commitment and reliability ('normative' trust). Notably, in contrast to the earlier discussion on clinicians' own views, this suggests that people acting out the role of patients are more willing to trust AI/CDSSs than their clinicians are. Hence, clinician fear of litigation due to their use of AI/CDSSs may be misplaced. In fact, these studies show quite the opposite: appropriate use of a CDSS may protect a clinician from potential liability in the eyes of potential litigants.

Pezzo and Pezzo conducted two experiments on participants' (asked to role play patients) perceptions of clinician use of CDSS.[121] In the first, 59 students read a hypothetical medical scenario where they were the patient. Some participants were told their doctor had successfully detected and treated a severe condition. Others were told their doctor had missed it, severely affecting their quality of life. Half were also told their doctor had used a computer CDSS to help analyse test results, and that this had a better diagnostic accuracy than clinicians. They were then asked to rate the doctor's decision quality out of five and whether they would recommend this doctor to a friend. Those who were told their doctor had missed the diagnosis were also asked how negligent they felt the doctor was, and how likely they would be to bring a negligence case against them. Generally, ratings on the decision quality scale were higher for those where the doctor diagnosed correctly: an average score of 3.47 for correct diagnosis versus 2.69 for incorrect where the CDSS was used; and 3.97 for correct diagnosis versus 2.03 for incorrect where the CDSS was not used. However, across both the (in-)correct diagnosis scenarios, CDSS use took 'some of the credit for a positive outcome and some of the blame for a negative outcome.'[122]

Hence, Pezzo and Pezzo's first experiment suggests that, far from subjecting clinicians to an increased risk of liability, reliance on CDSS may act as a shield from adverse judgment. This conclusion does, however, require some caution. Unlike the studies in Section III above (in relation to clinicians' views), participants here were explicitly told to assume that the CDSS had better diagnostic accuracy than a human. Against that background, the positive views of 'patients' are not all that surprising. However, in their second experiment, where 'patients' also reported very positively about the use of CDSSs, they had not been given any indication about its diagnostic accuracy. This study involved 154 students reading a malpractice case where a radiologist was alleged to have negligently caused a cancer patient's death. Cases were edited to reflect three variations: the doctor agreed with a CDSS ('agree'), disagreed but followed its advice anyway ('heed'), or disagreed and did not follow it ('defy'). There was a control group where no CDSS was used. Participants indicated (on a 7-point scale) to what extent the radiologist was at fault for the death. It was found that 'greater fault was perceived' in 'defy' than 'heed' or 'agree' (with the CDSS) cases and, further, the doctor was considered significantly more at fault in the control group (ie no CDSS used at all) than in the 'agree' group. Hence, like their first experiment, the results suggest that where a clinician relies on CDSS, 'use of a computer aid in the context of medical error is protective'.[123]

[121] Mark V Pezzo and Stephanie P Pezzo, 'Physician Evaluation After Medical Errors: Does Having a Computer Decision Aid Help or Hurt in Hindsight?' (2006) 26(1) Medical Decision Making 48.
[122] ibid.
[123] ibid.

Again, participants did not appear to share the reluctance to trust CDSS that affects some clinicians.

Likewise, in another study, Arkes and others conducted a series of mock jury experiments in which 657 members of the public of all ages[124] participated. (recruited by an external company and therefore drawn from the entire US population with access to a telephone.)[125] Each participant was provided with a DVD showing a mock malpractice trial. Changes were made to the cases that different participants viewed, to examine the effect on their judgement of several variables, one of which was whether the doctor used a CDSS. After viewing the DVD, participants were asked whether the physician met the standard of care and therefore whether they were guilty of negligence. The questions addressed both of these elements (standard of care and guilt). Those who answered 'guilty' were then asked how deserving of punishment the physician was (from 1 to 9 in order of punitiveness). For our purposes, the most interesting finding was that decisions about meeting (or not) the standard of care, and thus guilty or not guilty verdicts, were not influenced by the use of the CDSS. Of those who found the defendant guilty (ie concluding that they fell below the standard of care), use of a CDSS significantly reduced the punitiveness score (4.74/9 (53%) versus 5.68/9 (63%): $P \leq 0.05$).[126] In other words, the use of CDSSs could, at worst, make no difference and, at best, operate as a shield from adverse judgment.

Hence, these empirical studies suggest that, where mistakes are made, patients and other members of the public could view clinicians who rely on CDSSs more favourably than either those who do not consult a CDSS or do but ignore it. Patients appear readier to trust/CDSSs than their clinicians and, consequently, are more forgiving when those clinicians who *do* use CDSSs make mistakes. We suggest that this may be down to the different types of trust being applied by the hypothetical patients and jurors in this context, compared with clinicians. As the previous sections considered, a number of clinician concerns about AI/CDSSs appear to be grounded in concerns about accuracy: will the AI/CDSS get it right? Can I rely on the conclusions it has drawn? In other words, a lack of trust of empirical 'truth claims'.[127] In contrast, these studies show that accuracy of CDSSs is not the only thing hypothetical patients and jurors are concerned with. Of course, accurate truth claims are also important to patients and the studies do reflect that: general ratings in these experiments were higher when the clinician using a CDSS got the diagnosis right.[128] However, this does not explain the observed difference in views when the correct decision was not reached (when the clinician relying on a CDSS was still judged more favourably than those who did not). Such a judgment suggests that a different kind of trust in CDSSs is present here, even when there is a breakdown in 'empirical' trust (because the doctor has arrived at the wrong answer; its 'truth claim' was false). That trust is O'Neill's *normative* trust in commitments and competence. As outlined earlier, one can trust someone's commitment and competence if it measures up to a normative standard, irrespective of whether its claims about the world are true. Trust need not be based on the belief that truth claims are accurate, but as O'Neill suggests 'on judgements about their commitment or reliability, and about the competence or expertise they will bring to action'.[129] Hence, this kind of trust is about reaching a particular norm or standard,[130] or about the process by which its conclusions are reached, rather than what the

[124] Hal R Arkes, Victoria A Shaffer and Mitchell A Medow, 'The Influence of a Physician's Use of a Diagnostic Decision Aid on the Malpractice Verdicts of Mock Jurors' 2008 28(2) Medical Decision Making 201, 203–205.
[125] ibid.
[126] ibid, 204. Indeed, this was the only variable that made a statistically significant difference. Demographics, such as age, made no significant difference.
[127] O'Neill (n 20).
[128] And, equally, when the clinician did *not* use CDSS and got it right.
[129] O'Neill (n 20) 294.
[130] ibid 295.

ultimate conclusion is.[131] Certainly, this is true of CDSS. We may wonder whether a human can be depended upon to turn up for appointments and keep their promises, but a machine (provided it is plugged in and is functioning correctly) appears more dependable. Likewise, the fact that AI/CDSS can have a substantial amount of data, knowledge, and expertise programmed into them (which is not all marketing hype)[132] could provide a basis for a high level of expertise and therefore a basis for trusting the process, if not the outcome.

As with the preceding sections, this analysis illustrates how O'Neill's trust lens can provide new insights and relevance to the legal field. Rather than just concluding that patients trust AI/CDSSs and clinicians do not, we can look deeper into exactly how and why that might be the case. It shows that patients are sometimes using a type of trust (or direction of fit) that clinicians are not. That is not a reassuring insight though, and indeed, it could be problematic. One might ask what normative standards are AI/CDSSs and the process by which they reach conclusions being held to? The fact that even failure is accepted by participants in empirical studies (or at least significantly more accepted than it is for human clinicians acting alone), suggests the comparison may be too lenient: trusted, but not trust*worthy*.[133] It is also interesting that the 'black box' concerns of clinicians discussed above (about the process the CDSS uses to arrive at conclusions) do not feature in studies on hypothetical patients at all. Perhaps, such trust is the result of participants comparing AI/CDSS to the tools and computers that they use in other areas of their lives, such as satnavs and weather apps; which sets the bar rather low. Further, if that is the baseline comparator, the vast difference in stakes ought to be acknowledged: it is not unheard of for satnavs to go wrong, but when they do the outcome is unlikely to be life-threatening. Whereas, if a CDSS is wrong, then the consequences can be fatal.

It could be argued that the difference here is partly down to the different agents being judged in these studies: In this section, the studies involved hypothetical patients, etc judging *clinicians*, whereas the studies in Section III involve clinicians judging *AI/CDSS*. However, even when it comes to healthcare apps used by patients directly, there appears to be a higher default level of trust than with humans. Fritsch and others found less than 5% of surveyed patients in hospital waiting rooms viewed AI in medicine negatively (when asked in principle/abstract).[134] In line with the above studies, they also found 'predominantly positive reactions' when asked in principle about physicians using AI.[135] Going even further, Gratch and others found that patients were *more* trusting when told that they were messaging a chatbot about their mental health versus being told that they were messaging a human.[136] Concerningly though, research looking at public ratings of smoking cessation apps has found unduly high regard for such apps, irrespective of their medical validity.[137] If the normative

[131] O'Neill has written similarly on 'judgements'. We can distinguish between judging how something measures up against certain standards versus how it measures up to what we see in the world: 'fit[ting] the world rather than to make the world fit or live up to' a principle. Onora O'Neill, 'Experts, Practitioners, and Practical Judgement' (2007) 4(2) Journal of Moral Philosophy 154.

[132] Amit X Garg and others, 'Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review' (2005) 293(10) JAMA 1223; Pavel S Roshanov and others, 'Features of Effective Computerised Clinical Decision Support Systems: Meta-regression of 162 Randomised Trials' (2013) 346 British Medical Journal f657.

[133] O'Neill (n 20).

[134] Sebastian J Fritsch and others, 'Attitudes and Perception of Artificial Intelligence in Healthcare: A Cross-Sectional Survey Among Patients' (2022) 8 Digital Health 1, 5.

[135] ibid.

[136] Jonathan Gratch and others, 'It's Only a Computer: The Impact of Human-Agent Interaction in Clinical Interviews' (2014) Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems 85.

[137] Lorien C Abroms and others found no correlation between smoking cessation apps' ranking/rating and their quality (measured by how closely apps adhered to US Public Health Service's Good Practice Guidelines), see Lorien C Abroms and others, 'A Content Analysis of Popular Smartphone Apps for Smoking Cessation' (2013) 45(6) American Journal of Preventative Medicine 732. See also, Pouyan Esmaeilzadeh, Tala Mirzaei, and Spurthy J Dharanikota who found Americans considered AI 'less trustworthy compared to traditional diagnostic and treatment processes when they interact directly with the

standards against which AI/CDSSs are judged by the public do indeed fall far below the legal standard to which human clinicians are held, then lawyers and regulators must seriously consider whether such trust is misplaced.

VI. CONCLUSION

This article began by noting that 'trust' and 'trustworthiness' are stated frequently as the key concerns with clinician use of AI in clinical decision support for a variety of stakeholders, such as policymakers, courts, clinicians, patients, and the public. However, the meanings of those terms are often assumed or implied. Different groups, whilst seemingly agreeing in principle that 'trust' and 'trustworthiness' are important, can in fact be referring to very different concepts and talking past one another. Without drilling down into what these terms mean (or what various stakeholders mean when they use them), it is impossible to move important debates about AI, liability, and the future of healthcare forward. As Sutrop puts it, 'although there is much talk about trust, surprisingly little is said about what constitutes trust and what it depends upon'.[138] Sutrop preferred to use the term 'reliance' in the context of much AI.[139] Similarly, when the NHS AI Lab conceptualised trust purely in terms of 'reliability', this conceptualisation was found to be inadequate. Based on this finding, the NHS AI Lab preferred to talk in terms of ((in)appropriate) 'confidence' rather than 'trust'.[140] However, we have shown in this article that these conceptual problems can be resolved. We do not have to abandon 'trust' and 'trustworthiness', nor suffer from using one-dimensional definitions. Instead, by building on O'Neill's tripartite trust framework (empirical accuracy, normative commitment, and normative competence), this article has shown a way of drawing these distinctions out into the open, facilitating analytical comparison, and allowing this important debate about trustworthy medical AI to move forward.

In the context of AI and clinical decision support, using O'Neill's lens has allowed us to uncover how clinicians' reported trust concerns tend to focus primarily on one kind of trust: *empirical* accuracy (can we trust that the answer from the AI is empirically correct?), whereas the liability regime, courts' interpretation of it, and the (albeit limited) empirical studies of public/patients' views on this in hypothetical legal scenarios take a different view. Their view incorporates more normative concerns of *commitment* and *competence*. This allows us to make sense of the otherwise perplexing examples in some of the studies discussed where non-clinician stakeholders still appear to 'trust' AI even when they know it was *empirically* wrong and therefore untrustworthy in fact (if not deed).[141] In doing so, we move beyond previous studies which, having identified trust as a problematic term, abandon it (to a greater or lesser extent) in favour of other terms, or rely on only one definition of trust.[142] Further, by examining clinicians' perspectives on their AI clinical decision support tools, this article has addressed a gap in the literature, which has thus far focused largely on apps used by patients directly, rather than their clinicians.[143] In analysing the Anglo-Welsh liability regime in terms of this trust framework, it builds upon Smith and Fotheringham's doctrinal critique

physicians.' Although unlike Abroms and others, their study was asking hypothetical questions rather than providing real advice, see Pouyan Esmaeilzadeh, Tala Mirzaei and Spurthy J Dharanikota, 'Patients' Perceptions Toward Human-Artificial Intelligence Interaction in Health Care: Experimental Study' 2021 23(11) Journal of Medical Internet Research 1.

[138] Sutrop (n 31) 500.

[139] ibid 512.

[140] Mike Nix, George Onisiforou and Annabelle Painter, 'Understanding Healthcare Workers' Confidence in AI' (*NHS AI Lab & Health Education England*, May 2022) 6.

[141] O'Neill (n 20).

[142] For eg, see Nix, Onifisiforou and Painter (n 140); Sutrop (n 31).

[143] Sheppard (n 23).

of duty of care in England and Wales[144] and Prictor's analysis of the Australian liability regime.[145] It has also provided analysis of patients' perspectives of their clinicians' use of AI and CDSSs (and allowed for comparison with clinicians' perspectives)—a perspective that has thus far been lacking.

The challenge for policymakers and researchers in future work on AI in digital health care is to ensure that the philosophically loaded and complex terms 'trust' and 'trustworthiness' are engaged with, defined overtly and multi-dimensionally (eg using O'Neill's tripartite framework). This allows for nuances between the different possible meanings of the words to be unpicked and contrasted. In doing so, we can move the debate forward without talking past one another, and hope for greater trust placed appropriately by all stakeholders in more trustworthy AI and clinical decision support tools.

ACKNOWLEDGEMENTS

We owe thanks to the anonymous reviewers for their valuable feedback, and similarly to our colleagues/peers in the Medical Law stream at the SLS annual conference, and at the Healthcare Disparities: Disruptive healthcare technologies and the patient conference, Manchester University. We gratefully acknowledge Nisan Alici for ensuring our compliance with the style guide.

Conflict of interest statement. JCW has received consultancy payments from the NHSX AI Lab/Transformation Directorate at NHS England for advising on the validation and evaluation of AI systems for the NHS, but the views shared in this article are his own. The authors declare no other competing interest that might be relevant to the views expressed in this article.

[144] Smith and Fotheringham (n 7).
[145] Prictor (n 7).